Computational techniques for cell signaling

by

Yasir Suhail

Submitted to the Department of Biomedical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

August 2017

Acknowledgements

I am grateful for the opportunity to pursue a PhD at Johns Hopkins and the support I was afforded. I will attempt to thank some of the many people whose support has been invaluable to me.

My research advisor, Dr. Joel Bader has been an incredible source of support, advice, and innovative ideas for solutions to research problems. His background in statistical physics and computational chemistry has shown me succinct, novel, and meaningful insights and relations between different problems.

Many collaborators within the University and beyond have made it possible for me to undertake this research and work on many problems. In no particular order, I would like to acknowledg some among these here. Dr. Jef Boeke and his group were responsible for the generation of highthroughput genetic interaction screen data that introduced me to the practical applications of network analysis in bioinformatics. Drs. Kshitiz Gupta, David Ellison, and Andre Levchenko constructed the devices, conducted experiments, and provided valuable insights for the work on modeling molecule transfer through tunneling nanotubes and the development of the flow ELISA method (not presented in this dissertation). Drs. Florent Villiers and June Kwak conducted the abscisic acid treatment experiments on *Brassica napus* protoplasts and provided me with the RNAseq short read data.

I also want to thank Drs. Larry Schramm, Donald Gelman, and Sarah Wheelan for all their time and patience revieweing my work and giving supportive suggestions and criticism.

I also want to thank my family and friends for being supportive and helping me throughout this time.

Abstract

Cells can be viewed as sophisticated machines that organize their constituent components and molecules to receive, process, and respond to signals. The goal of the scientist is to uncover both the individual operations underlying these processes and the mechanism of the emergent properties of interest that give rise to the various phenomena such as disease, development, recovery or aging. Cell signaling plays a crucial role in all of these areas. The complexity of biological processes coupled with the physical limitations of experiments to observe individual molecular components across small to large scales limits the knowlege that can be gleaned from direct observations. Mathematical modeling can be used to estimate parameters that are hidden or too difficult to observe in experiments, and it can make qualitative predictions that can distinguish between hypotheses of interest. Statistical analysis can be employed to explore the large amounts of data generated by modern experimental techniques such as sequencing and high-throughput screening, and it can integrate the observations from many individual experiments or even separate studies to generate new hypotheses.

This dissertation employs mathematical and statistical analyses for three prominent aspects of cell signaling: the physical transfer of signaling molecules between cells, the intracellular protein machinery that organizes into pathways to process these signals, and changes in gene expression in response to cell signaling. Computational biology can be described as an applied discipline in that it aims to further the knowledge of a discipline that is distinct from itself. However, the richness of the problems encountered in biology requires continuous development of better methods equipped to handle the complexity, size, or uncertainty of the data, and to build in constraints motivated by the reality of the underlying biological system. In addition, better computational and mathematical methods are also needed to model the emergent behavior that arises from many components. The work presented in this dissertation fulfills both of these roles. We apply known and existing techniques to analyse experimental data and provide biological meaning, and we also develop new statistical and mathematical models that add to the knowledge and practice of computational biology.

Much of cell signaling is initiated by signal transduction from the exterior, either by sensing the environmental conditions or the recpetion of specific signals from other cells. The phenomena of most immediate concern to our species, that of human health and disease, are usually also generated from, and manifest in, our tissues and organs due to the interaction and signaling between cells. A modality of inter-cellular communication that was regarded earlier as an obscure phenomenon but

has more recently come to the attention of the scientific community is that of tunneling nanotubes (TNs). TNs have been observed as thin (of the order of 100 nanometers) extensions from a cell to another closely located one. The formation of such structures along with the intercellular exchange of molecules through them, and their interaction with the cytoskeleton, could be involved in many important processes, such as tissue formation and cancer growth. We describe a simple model of passive transport of molecules between cells due to TNs. Building on a few basic assumptions, we derive parametrized, closed-form expressions to describe the concentration of transported molecules as a function of distance from a population of TN-forming cells. Our model predicts how the perfusion of molecules through the TNs is affected by the size of the transferred molecules, the length and stability of nanotube formation, and the differences between membrane-bound and cytosolic proteins. To our knowledge, this is the first published mathematical model of intercellular transfer through tunneling nanotubes. We envision that experimental observations will be able to confirm or improve the assumptions made in our model. Furthermore, quantifying the form of inter-cellular communication in the basic scenario envisioned in our model can help suggest ways to measure and investigate cases of possible regulation of either formation of tunneling nanotubes or transport through them.

The next problem we focus on is uncovering how the interactions between the genes and proteins in a cell organize into pathways to process call signals or perform other tasks. The ability to accurately model and deeply understand gene and protein interaction networks of various kinds can be very powerful for prioritizing candidate genes and predicting their role in various signaling pathways and processes. A popular technique for gene prioritization and function prediction is the graph diffusion kernel. We show how the graph diffusion kernel is mathematically similar to the Ising spin graph, a model popular in statistical physics but not usually employed on biological interaction networks. We develop a new method for calculating gene association based on the Ising spin model which is different from the methods common in either bioinformatics or statistical physics. We show that our method performs better than both the graph diffusion kernel and its commonly used equivalent in the Ising model. We present a theoretical argument for understanding its performance based on ideas of phase transitions on networks. We measure its performance by applying our method to link prediction on protein interaction networks. Unlike candidate gene prioritization or function prediction, link prediction does not depend on the existing annotation or characterization of genes for ground truth. It helps us to avoid the confounding noise and uncertainty in the network and annotation data. As a purely network analysis problem, it is well suited for comparing network

analysis methods. Once we know that we are accurately modeling the interaction network, we can employ our model to solve other problems like gene prioritization using interaction data.

We also apply statistical analysis for a specific instance of a cell signaling process: the drought response in *Brassica napus*, a plant of scientific and economic importance. Important changes in the cell physiology of guard cells are initiated by abscisic acid, an important phytohormone that signals water deficit stress. We analyse RNA-seq reads resulting from the sequencing of mRNA extracted from protoplasts treated with abscisic acid. We employ sequence analysis, statisitical modeling, and the integration of cross-species network data to uncover genes, pathways, and interactions important in this process. We confirm what is known from other species and generate new gene and interaction candidates. By associating functional and sequence modification, we are also able to uncover evidence of evolution of gene specialization, a process that is likely widespread in polyploid genomes.

This work has developed new computational methods and applied existing tools for understanding cellular signaling and pathways. We have applied statistical analysis to integrate expression, interactome, pathway, regulatory elements, and homology data to infer *Brassica napus* genes and their roles involved in drought response. Previous literature suggesting support for our findings from other species based on independent experiments is found for many of of these findings. By relating the changes in regulatory elements, our RNA-seq results and common gene ancestry, we present evidence of its evolution in the context of polyploidy. Our work can provide a scientific basis for the pursuit of certain genes as targets of breeding and genetic engineering efforts for the development of drought tolerant oil crops. Building on ideas from statistical physics, we developed a new model of gene associations in networks. Using link prediction as a metric for the accuracy of modeling the underlying structure of a real network, we show that our model shows improved performance on real protein interaction networks. Our model of gene associations can be use to prioritize candidate genes for a disease or phenotype of interest. We also develop a mathematical model for a novel inter-cellular mode of biomolecule transfer. We relate hypotheses about the dynamics of TN formation, stability, and nature of molecular transport to quantitative predictions that may be tested by suitable experiments. In summary, this work demostrates the application and development of computational analysis of cell signaling at the level of the transcriptome, the interactome, and physical transport.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prelude

1.1.1 Biological systems: order and disorder

Life is characterized by the ability to organize matter into cells during growth and reproduction, while maintaining homeostasis and responding to changing environmental conditions. Following the definition of Schrodinger (1992), life is the ability to maintain (or accumulate) negative entropy. In general, a closed system should always be moving to a state of higher entropy; on the other hand, life seems to be moving in the opposite direction of the second law of thermodynamics. Microscopic deviations from the second law, however, can occur by chance. Theoretical advances in non-equilibrium statistical physics have made it possible to quantify the probabilities of irreversible processes happening away from equilibrium. For a system in contact with a heat bath, short term increases in entropy in violation of the second law of thermodynamics occur with a non-zero probability, but according to the Cook's fluctuation theorem (Crooks, 1999; Gallavotti and Cohen, 1995; Kurchan, 1998), the chances of this accumulation of negative entropy decrease exponentially with time and the rate of entropy decrease. Specifically, it says that the probability $\Pr(-S)$ a system changing its entropy at the rate $-S$ per unit time, for a total time τ is given by

$$\frac{\Pr(-S)}{\Pr(S)} = \exp(-S\tau)$$

Therefore, it is exponentially more likely for a system to increase its entropy than to decrease it.

Of course, this is not at all how cells work – as a kind of Maxwell's demon. They do not maintain homeostasis or divide simply by using microscopic fluctuations away from equilibrium. Instead, they employ a complex machinery that actually uses an energy source (such as light or a chemical diet) and a heat sink (the surrounding environment) to dump the excess entropy generated. The entropy of the living matter decreases or is maintained at the cost of increasing the entropy of the environment.

There is another way of looking at this. Instead of just thinking about the increase in entropy as the *cost* of maintaining life, a somewhat speculative argument suggests that life serves as the catalyst for what nature demands, i.e., the fastest and largest possible increase in entropy.

Michaelian (2009) and others such as Annila and Annila (2008) argue that life and the biosphere are optimized for the maximum total production of entropy on earth. The earth as a whole absorbs radiation from the sun, and dissipates this energy back into space. The maximal increase in entropy occurs when incoming high frequency radiation (such as in the UV range) is converted to lower frequency thermal radiation which is exported into space. Photosynthesis is of course one of the ways of directly capturing some of the energy from the sun, and this is one way in which life on earth captures energy. However, the largest engine of entropy production is the water cycle which radiates (at much larger wavelengths than the incoming radiation) when it condenses in the atmosphere. However, Michaelian (2009) argues that the biosphere as a whole accelerates this working of the water cycle engine. On a molecular level, photosynthesis captures energy which may be later radiated downstream by animals, fungi, or bacteria. Wang et al. (2007) calculated that leaves and the stomata actually operate as if optimizing for maximal transpiration from the leaves, rather than simply energy capture from photosynthesis. Michaelian (2009) also presents arguments that parts of the biosphere other than plants also operate as a system to maximize entropy production at a scale vastly bigger than would happen on a lifeless planet; and that early RNA and DNA molecules by themselves may have been the optimal molecules to capture photons from UV light and transmit this energy to surrounding water, stimulating evaporation and feeding the water cycle entropy engine. Recently, the non-equilibrium thermodynamics of systems driven by an external energy source to undergo irreversible processes was applied to understand self-replication rates of simple molecules such as RNA (England, 2013, 2015). Similar arguments have been applied for available reservoirs of chemical energy, arguing that biological matter was able to generate entropy by hydrogenating

the carbon dioxide that saturated the early atmosphere (Yung and Russell, 2010). The overriding theme here is that the complexity and organization of the biology systems, including the cell walls, intracellular compartments, the information content of the genome, and protein structure, is related to its property of entropy production. We can put circular teleological conjectures aside; whatever the odds of its emergence, the supposed "purpose" of life as an entropy maximization agent of the universe is more accurately the emergent property of self-replication dynamics, given the constraints of the conditions, physical laws, and its dynamic environment. Regardless of these arguments, what is clear is that the complex machinery of biological matter enables it to perform a very complex task: to overcome the energy barriers that separate the state of its available resources from the energy valleys they could potentially occupy.

1.1.2 Signaling and dynamics

While the complexity of cells and organisms is remarkable in terms of its static organization, the cell is a dynamic machine that needs to control various non-equilibrium processes. The steps of transcription, translation, chaperone-assisted folding, and post-translational modifications have to occur sequentially. The different phases of the cell cycle need to be sensed, started, and stopped. Vesicles have to be formed, transported, and absorbed. The control of all of these processes means that various signals have to be generated, stored, and processed. All of this would still be necessary if the cell existed in a static environment with a constant and uniform source of energy. However, a cell also has to take into account a dynamic macro-environment; it has to maintain its organization not just in the face of the microscopic disorder of the second law, but also the macroscopic dynamic disorder of the availability of resources and changing physical conditions.

Cell signaling is required for sensing and observing the environment. Chemical concentrations, temperature, and mechanical signals all convey information about resource availability and probability of damage. Secondly, the dynamic optimization of metabolic processes according to resource availability, avoidance of environmental insults, repair of the proceeding injury, and optimal timing of division or reproduction are complex decisions that require a complex computational machinery to solve. This problem of computing the best response for a changing environment with limited resources can also be framed in terms of trying to predict the future of the environment (Still et al., 2012). This computational problem, whether framed as an optimization problem, a prediction task, or a nonlinear control system is one that has a tremendous effect on biological fitness but that must

be solved with limited time, memory, and processing units. It is obvious that one way to help this would be some sort of co-operative means employing a number of cells like the various ensemble learning, message passing, and parallel processing algorithms employed on computers.

Multicellular organisms have to ensure that the whole organism responds appropriately to the present or predicted environment. The environmental sensing signal might have to be conveyed to a cell spatially removed from the site of optimal sensing, or to a cell that is specialized to perform the adaptive action. Both of these scenarios occur in a problem that we study in this work: the abscisic acid signaling of drought response, where the signal is communicated from a spatially removed organ (the root to the leaves), or from cells of one phenotype to the other (mesophyll cells to guard cells). Cell-to-cell communication also occurs in unicellular organisms. Quorum sensing in bacteria is a well known example of unicellular organisms employing cell-to-cell communication to regulate their gene expression as a function of population density. Signaling to and from cells can occur in different modalities. Chemical signaling by various small and large molecules is pervasive in biology. Neurons and myocytes respond to, and neurons generate, electrical signals. Cells can also sense mechanical stresses via strain on cytoskeletal proteins (Ulbricht et al., 2013) and "communicate" mechanical signals through the extracellular matrix (Reinhart-King et al., 2008).

1.2 Motivation

All processes of sensing, computation, and decision-making performed by an organism are necessarily operations on signals. A cell sensing its environment implies receiving external signals; their processing involves intracellular signaling through mechanisms like protein binding, conformation changes, protein modifications or gene expression; and part of its response could be the communication of signals to other cells. At a much higher level of organization, even information processing in the brain is built on chemical and electrical cell signaling between neurons. Understanding the signaling involved in a certain process, whether in disease, development or homeostasis, is key to understanding the process. Knowledge of the signaling starting from the extracellular receptors to the intracellular pathway to the secreted signals means that we can predict the response to new conditions, find targets for pharmaceutical intervention, or propose candidates for genetic engineering.

In this work, we apply and develop computational and statistical techniques that help in understanding various aspects of cell signaling. From the perspective of systems biology, we can divide the phenomena involved in cell signaling into various levels of abstraction. At the lowest level, we have individual molecules whose physical and chemical properties are responsible for the whole system. These individual molecules, whether nucleotide polymers, proteins or metabolites, interact with each other in some fashion. Sets of these interactions happening in concert with each other give rise to whole cell processes and responses. Going further up the hierarchy, these cellular responses may then affect other cells. At each level, there are individual processes occurring that give rise to emergent properties.

This dissertation tackles three problems at different levels of abstraction. We are interested in both uncovering the biological phenomena and developing tools to represent and model the phenomena. In practice, however, a specific effort may be more of an exploration of biological observations or of conceptual advance. The unifying theme in our studies is the application of quantitative tools in the service of understanding cell signaling at different levels of abstraction and organization.

1.2.1 Cell-cell communication

Cell-cell communication is of primary importance for understanding human disease and development since they are a product of the behavior of the whole system. In terms of the number of nucleotides or functional genes, genomes can increase to very large sizes. Whether this alone leads to a more

"complex" organism, for some behavioral measure of complexity, is debatable. Many plants in fact, have genomes many times the size of the human genome. Most empirical scaling laws suggest that simply increasing the size of a certain system increases some measure of its performance up to some saturating limit. For example, West et al. (1999) shows that the scaling of the rates of biological processes scale slower than body size in animals, and this is due to the vascular structure needed to transport nutrients and communication. Bettencourt et al. (2007) shows that a city's productivity, in terms of wealth generation, rises faster than its population, according to a scaling law leading to bigger and denser cities, until this regime collapses leading to a breakdown.

Whatever the scaling of biological functional complexity may be with respect to its gene set size, there is a stage beyond which one requires a qualitatively different kind of structure to continue to increase efficiency. Larger organisms are multi-cellular rather than simply giant cells. All of these multicellular organisms need cell-cell communication to exist and function as a whole. Major modes of cell-cell communication include endocrine signaling (where the signaling molecule circulates in the blood to its target), paracrine signaling (where the molecule diffuses locally to other cells in the source cell's environment), and Delta-Notch signaling between the membrane bound receptor and signaling molecules of cells in contact with each other.

We turn our attention in this work to a kind of cell-cell communication about which comparatively little is known, tunneling nanotubes (TNs). Cells send out protruding nanotubes from their cell bodies. The nanotubes are able to fuse with the cell wall of a nearby cell to form a nanotube connection, facilitating the transfer of signals. We present a basic model of the dynamics of molecule transfer with these TNs. TNs have been observed in many different situations including neuronal (Costanzo et al., 2013), corneal(Chinnery et al., 2008) and lung cells(Lou et al., 2012) and they have been implicated in many important phenomena such as tumour formation (Thayanithy et al., 2014), chemoresistance (Dickson et al., 2014), and stem cell differentiation (Vallabhaneni et al., 2012). It has been hypothesised that TNs may represent a new kind of supercellular organisation in animal tissues (Rustom, 2016). By modeling the dynamics of biomolecule transfer by TNs, we open the door to experiments that can characterize the level of regulation in a signal of interest.

1.2.2 Networks and signaling

Biological processes within a cell occur due to many molecular components working together. Our knowledge of the physics of protein structure and binding is too rudimentary to be able to calculate

their behaviour and role in the system from first principles. Interaction networks provide additional information about the relative roles of proteins in the cellular system. These networks can be composed of many different kinds of interactions. For example, regulatory interactions denote the direct or indirect effect of one protein on the expression or activity of another, protein-protein interactions usually denote the physical binding of proteins to each other, genetic interactions arise from some co-operative behaviour that leads to non-additive phenotypes in double mutants, and co-expression networks, as the name suggests, record correlated gene expression of certain genes in many different conditions.

The cellular response to a stimulus is, in almost all cases, composed of a number of genes (or more accurately their protein or RNA products) working together. If we have even partial knowledge of the response or signaling pathway in terms of certain genes, a great deal can potentially be predicted using the interaction networks. For example, interaction networks can be used to search for additional genes that are highly associated with a core gene set involved in a pathway or disease phenotype. They can be used to characterize the unknown function of genes, predict dynamic behavior, arrange the genes into functional modules, or prioritize genes for experimental characterization or validation.

In this study, we have used interaction data for exploratory analysis of the guard cell abscisic acid response. We also attempted to define the causal genes involved in quantitative trait linkage (QTL) of regions of the genome to certain drought related phenotypes. However, the sparsity of available network data and limited resolution of the genetic markers mapping the QTL regions was insufficient to give us statistically significant or even informative results. The model we developed for gene associations over networks could solve these kind of problems as experimental data sets improve, however. We had earlier demonstrated how the graph diffusion kernel, a common model for network analysis, can be adapted to genetic interactions (Qi et al., 2008). We present, as the second topic here, what we show to be an improvement over the graph diffusion kernel, and show its close relationship to models developed from the statistical physics of spin lattices.

1.2.3 Cell signaling and gene expression

An extracellular chemical signal is usually sensed either by a cell membrane receptor or in the case of diffusible small molecules, by an intracellular target. In either case, any signal will usually trigger a cascade of downstream events through various processes. These downstream events can

include many different processes, especially in eukaryotes, such as phosphorylation, ubiquitination, transport, binding, or changes in the conformational structure of existing proteins. For certain classes of cells and signals, such as an electrical spike train sensed in the dendrite of some neurons, the response may primarily be the transmission of a modulated signal. Other stimuli may induce slower and longer lasting changes such as in metabolic fluxes or cytoskeleton structure. Due to energetic constraints, a cell is unlikely to maintain a large surplus of protein machinery that is not in regular use. Therefore, an important signal that requires a significant change in the functioning of the cell, especially one that does not have a relatively uniform repetitive frequency, will require additional protein machinery to be assembled. This implies a large and systematic change in gene expression in response to the stimulus. In guard cells, abscisic acid is one such important signal. Its response includes a large amount of ion transfer, with the resulting changes in turgidity proceeding to close or narrow the stomatal opening. The sudden and large change in ion concentrations and turgidity leads to osmotic stress, which requires the expression of osmoprotectants. In addition, water deficit also implies changes in metabolic resources, and might be predictive of falling levels or photosynthesis. This might necessitate transport of photoassimilate to safe storage or other changes in the metabolic program. All of these targets require stepping up the rates of some processes and production of its machinery.

While *Brassica napus* is a commercially important food crop, especially in certain countries like Canada and China, and a target for selective breeding, it has not been a model organism for scientific research. Its first draft genome was published in 2014. Therefore, its genes have not been annotated in detail, and no interactome information is available. We were able to use RNA-seq to characterize the gene expression changes in response to abscisic acid in the guard cells. We also use homology to *Arabidopsis* and *Arabidopsis* interaction and transcription factor binding site to infer the processes and genes involved.

1.3 Research aims

1.3.1 To propose a model for inter-cellular molecule transfer through tunneling nanotubes

Cell-to-cell communication proceeds through a number of modalities. Tunneling nanotubes are one medium of communication of emerging scientific interest. Nanotube structures have been observed to form between many different kinds of cells and can transport, proteins, cell components, and electrical signals. We develop a mathematical model of the transport of intracellular and membrane-bound molecules using model in which the nanotube permits mixing of the two cells' membranes, and we relate the rates of transfer to nanotube development dynamics. The model makes simple assumptions for dynamics and diffusion, and it can be augmented once experiments generate different parameter values that confirm or invalidate model predictions.

1.3.2 To unify and develop a model for understanding network structures

Understanding the network structure of an interaction graph is an important problem in systems biology, and it can be used for biological discovery and hypothesis generation in a number of ways. In the context of this dissertation, we wished to prioritize candidate genes from drought responsive QTL regions using protein interactions. For this and similar problem classes, we developed new algorithms for network analysis. We review the graph diffusion kernel, a popular model for predicting gene associations in networks, and relate it to the Ising model for the statistical physics of spin lattices. We develop a novel approximation of spin-spin correlations in the Ising model. We evaluate the performance of our model for predicting missing links in protein-protein interactions and show that our formulation shows better performance than similar methods. We show how our model may be applicable for other applications and how it provides a more unified view of related network analysis techniques.

1.3.3 To study the response of abscisic acid signaling in *Brassica napus* guard cells

We study the response of the *Brassica napus* guard cell to an externally applied abscisic acid signal. Abscisic acid is a major phytohormone and signals water deficit stress. We use short read sequencing

of mRNA in response to the application of abscisic acid to isolated protoplasts. We apply statistical tests to understand the changes in gene expression, and we then integrate cross-species and regulatory interaction data in a systematic fashion to understand the genes and interactions involved in drought response. In addition, we show that abscisic acid response in *Brassica* has evolved since it and the related *Arabidopsis* lineage split from a common ancestor.

Chapter 2

Modeling intercellular transfer of biomolecules through tunneling nanotubes

2.1 Abstract

Tunneling nanotubes (TNs) have previously been observed as long and thin transient structures forming between cells and intercellular protein transfer through them has been experimentally verified. It is hypothesized that this may be a physiologically important means of cell-cell communication. This paper attempts to give a simple model for the rates of transfer of molecules across these TNs at different distances. We describe the transfer of both cytosolic and membrane bound molecules between neighboring populations of cells and argue how the lifetime of the TN, the diffusion rate, distance between cells, and the size of the molecules may affect their transfer. The model described makes certain predictions and opens a number of questions to be explored experimentally.

2.2 Introduction

Cell-cell communication plays an important role in coordinating collective cell decisions. It is also critical to maintain both structural and functional homeostasis in a tissue. Since coordination between cells is an essential requirement for the successful functioning of a multicellular organism, many mechanisms have evolved to allow cells to communicate with each other bearing important outcomes in both normal functioning of the tissues and pathology. For example, high twitch muscle cells are in close proximity with blood vessels and nerve endings, and their close interactions are essential for the correct muscle functioning (Behnke et al., 2011; Vikne et al., 2012). Cell-cell interactions between cancer cells and endothelial cells occur within solid tumors, and metastatizing cancer cells extravasating the endothelium (Weis and Cheresh, 2011; Qin et al., 2012; Stine et al., 2011). Extensive research has explored the mechanisms of these cell-cell interactions, resulting in extensive information on the chemical cell-cell signaling pathways occurring in autocrine (Lichtenberger et al., 2010), paracrine (Abou-Khalil et al., 2009), and juxtracrine manner (Bosenburg and Massague, 1993; Singh and Harris, 2005). However, recent evidence has also suggested another form of cell-cell interaction that occurs via direct transfer of cellular components from one cell to another, thus transferring information without involvement of traditionally implicated chemical mechanisms (Niu et al., 2009; Ahmed and Xiang, 2011; Li et al., 2010; Pap et al., 2009; Prochiantz, 2011; Mack et al., 2000). Although the degree of such intercellular transfer of cellular components and its role in defining cell and tissue behavior *in vivo* remain less understood, the evidence for existence of this novel communication mechanism is overwhelming, suggesting that it could potentially have a significant effect in influencing the recipient cell phenotypes in such diverse processes as cancer progression (Ambudkar et al., 2005), immunity (Baba et al., 2001; Carlin et al., 2001; Quah et al., 2008), HIV infection (Mack et al., 2000), transfer of drug resistance (Levchenko et al., 2005), and ribosomal recruitment in neuronal axons (Twiss and Fainzilber, 2009). Direct protein-protein transfer is therefore important to understand in greater detail, both experimentally and computationally.

Previous studies have reported multiple examples of transfer of membrane proteins between cells (Levchenko et al., 2005; Guescini et al., 2012; Agnati et al., 2011; Al-Nedawi et al., 2008; Davis, 2007). In addition, small cytoplasmic biochemical components have also been shown to be transferred between cells in a size-dependent manner (Niu et al., 2009). However, intercellular transfer of large cytoplasmic proteins has not been yet examined with conclusive results. Various mechanisms have been suggested for intercellular transfer of cellular components, including forma-

tion of tunneling nanotubes (TNs) between cells (Guescini et al., 2012; Rustom, 2004), spontaneous secretion and integration of microvesicles (Valadi et al., 2007; Denzer et al., 2000), and transient cell-cell fusion (Dreisen et al., 2005). As is frequently the case with poorly understood biological phenomena, it is not easy to discriminate between putative mechanistic details and generate most plausible models of this cell communication phenomenon. It is also possible that the mechanisms may be cell-type specific and multiple mechanisms might coexist in diverse physiological and pathophysiological contexts. However, certain findings have been suggestive of the constraints that can be placed on the mechanistic models of this process. For instance, the reports of membrane protein transfer are much more frequent and better supported than the reports of transfer of large cytosolic components, including of proteins (Agnati et al., 2011; Camussi et al., 2010). We questioned whether a reason for this discrepancy might lie in the properties of the transfer process itself. Another, potentially more revealing constraint comes from the observation that transfer of cytosolic, but not membrane components is strongly dependent on the molecular weight of the transferred molecules (Niu et al., 2009). Thus, a plausible model of the transfer process has to be able to explain these particular well-established features of the intracellular transfer of different cellular components.

Here, we propose a mathematical model to explain passive protein transfer between cells via formation of tunneling nanotubes (TNs), which have been observed in various studies to be responsible for intercellular protein transfer (Guescini et al., 2012; Rustom, 2004; Bukoreshtliev et al., 2009). Our steady state model explains that while membrane protein transfer may be unrelated to the mass of the protein, cytoplasmic proteins may follow an inverse correlation with size. Though no existing report conclusively shows the transfer of cytoplasmic proteins between cells, smaller cytoplasmic components have been shown to be transferred between cells in a size-dependent manner (Niu et al., 2009), as predicted by the model. The model explains that while transfer of cytoplasmic proteins may occur between cells, it would be in relatively smaller amounts in comparison to smaller biochemical components present in the cytosol, or membrane proteins. Further, we predict that protein transfer may depend on the stabilization of TNs for longer duration.

2.3 Methods

2.3.1 Basic assumptions

Previous studies have revealed that proteins and other cellular components can transfer between cocultured cells (Niu et al., 2009; Li et al., 2010; Prochiantz, 2011; Davis, 2007). Typically donor and recipient cells are defined according to the criterion of observation for the transfer. Commonly, these observations are specific to the transferred component, e.g., by using an antibody or fluorescent tag to observe the dynamics of transfer of a biochemical molecule from one cell to another. The schematic in Figure 2-1 details a typical experimental setup used to detect transfer of cellular components between cells. For simplicity, in the schematic and in the model, we assume that both membrane and cytosolic components are transferred from a donor population to a second recipient population.

Variables and constants used in the model are described in Table 2.1. We assume that cells are cultured as adherent cells in a dish. Consider a cell located at the origin of a system of coordinates superimposed onto the cell adhesion substratum. The cell can exchange proteins or other molecules with cells around it by sampling the space in some manner, by means of tunneling nanotubes (TNs) protruding into the extracellular space (Figure 2-2).

The maximal length of TNs will be limited by the physical and energetic constraints of the cell. The growth of the exploring TN can be expected to be driven by some sort of polymeric growth, like the filopodia or actin growth. In any given direction, this growing nanotube can only expect to make a connection with the closest cell. If the cells are uniformly distributed points, and the placement of one cell is independent of another, the distance of any cell to its nearest cell will follow an exponential distribution. The physical limit of the TN growth, however, limits the exponential at its tail and most of the distribution thus lies in the linear regime of the exponential. As a first-order approximation, we can thus approximate the abundance of the TN lengths to fall linearly with length. Denoting the maximum length as l, we assume that the length r of such TNs follows a distribution:

$$p(r) \propto \begin{cases} l - r & \text{if } r < l, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Hence, the probability of forming a connection to transfer protein with a part of another cell located within the infinitesimal region $(r\,dr,\, d\theta)$ at the polar co-ordinates (r, θ) within some time unit is

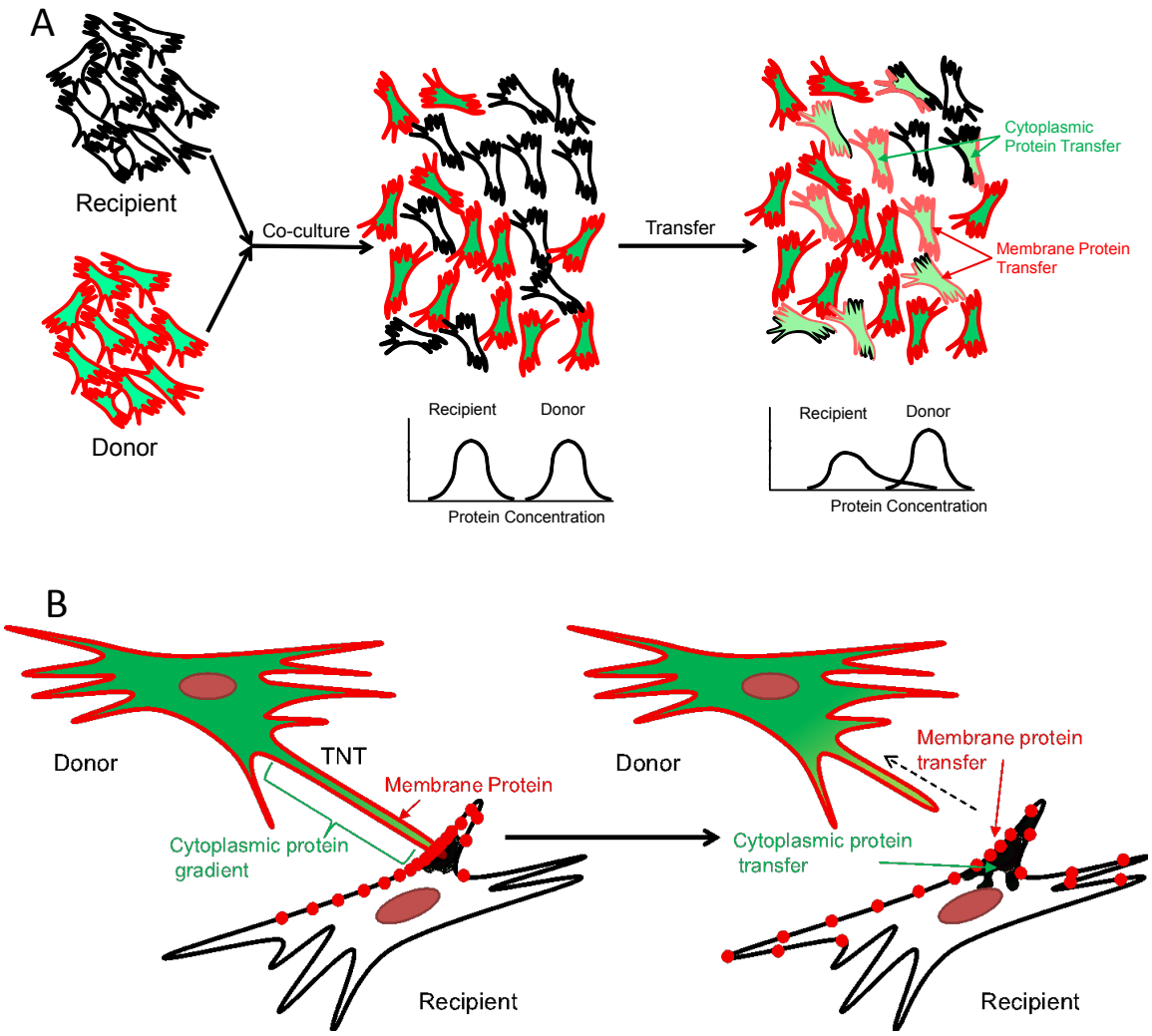

Figure 2-1: Schematics showing the experimental observations of molecule transfer and the TNs. (A) Transfer of membrane and cytosolic protein transfer between cells in coculture. The donor and recipient cells are defined according to the observation criterion. After coculture, both membrane and cytosolic proteins are transferred to the recipient cells from the donor cells at a relatively slow rate in comparison to the rate of production of proteins in the donor cells. Observation post coculture depicts a small population of recipient cells that received transferred protein that can now be detected. (B) Schematic showing transfer of membrane and cytosolic components from acceptor to donor cells via tunneling nanotubes (TNs). In this model, donor cells contains higher amount of cytosolic component (shaded), and membrane bound component (dots) than the recipient cell. Coculture results in formation of TNs from the donor cell that can transiently connect with the recipient cell, resulting in transfer of both cytosolic component, and membrane-bound component. In the model, the membrane composition of TN remains similar to the rest of the donor cell membrane, the cytosol within the TN shaft contains a gradient of cytosolic components till steady state is reached. Since most TNs are transient (i.e., their lifetime is smaller than that required for the concentration of cytosolic components within the TN shaft to attain steady state), the transfer of cytosolic components to the recipient cell is determined by the concentration of the component at the site of connection between the TN and the recipient cell. The cytosolic component is green and the membrane proteins are red dots.

Figure 2

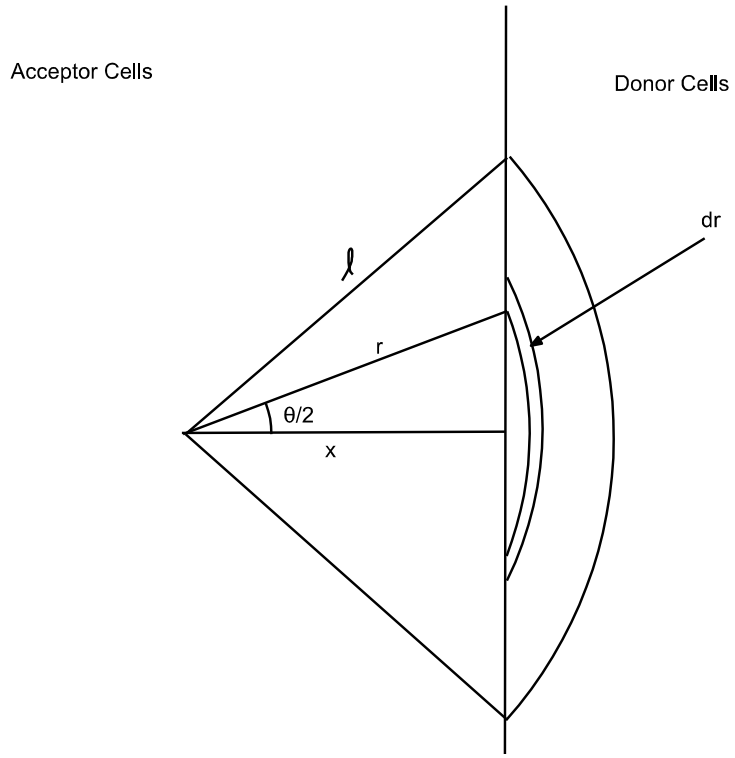

Figure 2-2: Schematic showing the calculation of the probability of a tunneling nanotube (TN) connection between an acceptor and donor cell. As explained in Equation (2.2), consider r as the length of the TN, and l the maximum length. For a cell located at a distance x from the boundary, there is an arc of angle $2\arccos(x/r)$ with cells located at a distance r. This corresponds to the illustrated infinitesimal area $2\arccos(x/r)r\,dr$, which can be integrated from $r = x$ to l for the total applicable area

$$dp(\text{Connection}|r,\theta) \propto \begin{cases} l-r & \text{if } r < l, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider another cell of radius b located distance r away from the donor cell of interest sending out TNs. Assuming that the region of sampling by the TNs is much larger than the dimensions of the cells ($l \gg b$ and $r \gg b$), the probability of forming such a connection will be $p(\text{Connection}|r) \propto (l-r)\pi b^2$.

Once such a connection is formed, we consider the cases of cytoplasmic and membrane proteins transferred from a region of donor cells to a region of acceptor cells.

Table 2.1: Description of all the constants and variables used in the mathematical model of the inntercellular transfer of molecules through nanotubes.

Description	Symbol used	Value used in the analysis	Source of the parameter value	Comments	Dimensions
Radius of a cell	b	No value used, analysis done algebraically			Length
Maximum length of TN	l		Inexact estimate in the range of observations reported in (Rustom et al. 2004)		Length

Description	Symbol used	Value used in the analysis	Source of the parameter value	Comments	Dimensions
Diffusion coefficient	D	See Table 2.2			Length^2/time
Stoke's radius	r	See Table 2.2			Length
Constant related to membrane bound molecule transfer	A	No value used, analysis done algebraically		Related to the area of membrane transferred on,each connection, the frequency with which a cell sends out TNs, and the,active transport of the molecule to donor cell membrane	Molecules/Length^3
Constant related to cytosolic molecule transfer	B	No value used, analysis done algebraically		Related to the frequency with which a cell sends out TNs and the concentration of the molecule in the donor cell	Molecules/Length^3
Cell density	ρ	No value used, analysis done algebraically			Cells/Length^2

2.3.2 Transfer of cellular components by TN

We assume that when a TN from a donor cell reaches the recipient cell, and connects with the membrane of a recipient cell, it can "donate" a small portion of the membrane to the donor cell. This process may actually occur as "exchange" of membrane portions, but here we describe only the transfer of observable membrane and cytoplasmic components present exclusively in the donor cells. Furthermore, we assume that TNs are open to diffusive transport of donor cell components and

the concentrations of the transferred components are not necessarily at the steady state in the TNs. Due to more extensive reservoirs of the potentially transferable cytosolic vs. membrane components (e.g., cytosolic proteins) and the potential for lower diffusivity through the cytosolic vs. membrane parts of TNs, the diffusion of the membrane components may lead to a more effective exchange vs. that of the cytosolic ones during the transient, TN-mediated cell-cell fusion. Thus, the transfer of membrane components may be limited by the rate of their access to an individual TN on the donor cell side, with the membrane density otherwise reaching a steady state within the TN. On the other hand, the transfer of cytosolic components may be limited by the rate of reaching the steady state in the TN, with transport mostly resulting in and dependent on the spatial gradient of the component within the TN.

2.3.3 Transfer of membrane proteins

Consider an acceptor cell located at a distance x from the region of donor cells, each of radius b. There is an arc of radius r subtending an angle of θ radians from the acceptor cell that falls on the region of the donor cells where $\theta = 2\arccos(x/r)$. Assuming a cell density of ρ cells per unit area, the probability of our cell making a connection with any cell located in the donor region at the distance x away will thus be

$$p(\text{Connection}|x) \propto \rho\pi b^2 \int_x^l (l-r) 2\arccos\left(\frac{x}{r}\right) r\, dr \tag{2.2}$$

Membrane bound molecules are actively transferred to the membrane by the cellular machinery. Every time a TN connection is formed there is a merging of the membranes of the two cells at one end of the TN. We assume that a small amount of membrane protein is transferred to the acceptor cell due to the TN-cell membrane contact. The total amount of membrane bound molecules (ϕ molecules per cell) transferred into an acceptor cell at a distance x from the region of the donor cells will be the frequency or abundance TN connections (represented by quantity of Equation (2.2)) multiplied by a constant related to the amount of molecules of interest transferred in each TN connection:

$$\left[\frac{d\,\phi}{dt}\right]_{\text{Transfer}} = A\rho\pi b^2 \begin{cases} \int_x^l (l-r) 2\arccos\left(\frac{x}{r}\right) r\,dr & \text{if } x < l, \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \frac{A\rho\pi b^2}{3}\left(-2lx\sqrt{l^2-x^2} + l^3\arccos\left(\frac{x}{l}\right) - x^3\log\left(\frac{x}{l+\sqrt{l^2-x^2}}\right)\right) & \text{if } x < l, \\ 0 & \text{otherwise.} \end{cases}$$

where A is a constant with dimensions of molecules/$m3$ collapsing all the unknowns such as density of the protein on the donor cell membrane, efficiency of transfer across cells, etc. With a protein degradation rate of β with units s^{-1}, we have the dynamics

$$\frac{d\,\phi}{dt} = \left[\frac{d\,\phi}{dt}\right]_{\text{Transfer}} - \beta\phi$$

This leads to the steady state condition of

$$\begin{aligned} \phi(x) &= \tfrac{1}{\beta}\left[\tfrac{d\phi}{dt}\right]_{\text{Transfer}} \\ &= \begin{cases} \frac{A\rho\pi b^2}{3\beta}\left(-2lx\sqrt{l^2-x^2} + l^3\arccos\left(\frac{x}{l}\right) - x^3\log\left(\frac{x}{l+\sqrt{l^2-x^2}}\right)\right) & \text{if } x < l \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{Al^3\rho\pi b^2}{3\beta}\left(-2\sqrt{1-(x/l)^2} + \arccos(\frac{x}{l}) - (x/l)^3\log\left(\frac{x/l}{1+\sqrt{1-(x/l)^2}}\right)\right) & \text{if } x < l \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2.3}$$

The protein is transferred up to a distance equal to the maximum length of the TNs (l) and the concentration of the transferred molecule at the boundary of the donor-acceptor cells is

$$\phi(x)_{x\to 0} = \frac{Al^3\rho\pi b^2}{3\beta},$$

while the slope of the concentration of transferred molecules at the boundary is

$$\left[\frac{d\phi}{dt}\right]_{x\to 0} = -\frac{Al^2\rho\pi b^2}{\beta}.$$

2.3.4 Transfer of cytoplasmic proteins

We consider an identical arrangement of donor and acceptor cells in this case. The chances of TN connections formed between the donor and acceptor cells is

$$p(\text{Connection}|x) \propto \rho\pi b^2 \int_x^l (l-r) 2 \arccos\left(\frac{x}{r}\right) r\, dr.$$

Here, we assume that the diffusive transport through a TN is the rate limiting step. According to Fick's law, in one-dimensional diffusion from a source of density η, the density at a distance r at time t is $\eta\text{Erfc}(\frac{r}{2\sqrt{Dt}})$ where D is the coefficient of diffusion and Erfc is the complimentary error function. We assume that once a connection is made, the transfer of a cytosolic component occurs due to diffusion for a certain amount of time (i.e., the effective connection time).

The amount of protein transfer per connection is thus proportional to $\text{Erfc}(\frac{r}{2\sqrt{DT}})$. Now the rate of protein transferred will be proportional to the number of connections made per unit time and the amount of protein transferred per connection:

$$\left[\frac{d\,\phi}{dt}\right]_{\text{Transfer}} = B\rho\pi b^2 \begin{cases} \int_x^l (l-r) 2 \arccos(x/r) \text{Erfc}(r/C) r\, dr & \text{if } x < l, \\ 0 & \text{otherwise} \end{cases}$$

where B is a constant of dimension molecules/m^3 (corresponding to the parameter A for the membrane bound case) incorporating the chemical unknowns and C is the mean diffusion length

$$C = 2\sqrt{Dt} \tag{2.4}$$

collapsing the diffusion coefficient and effective mean connection time of the TN connections. The integral with the error function can be computed numerically but is analytically cumbersome.

Since we have already used the one-dimensional approximation and assumed no diffusion within the TN before the formation of the connection, we can make one more simplifying approximation and use a linearized approximation to the error function,

$$\text{Erfc}(x/C) = \begin{cases} 1 - \frac{2x}{C\sqrt{\pi}} & \text{if } x < \frac{C\sqrt{\pi}}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Like the case of the membrane bound proteins with degradation rate β, we can solve for the steady state with the aforementioned approximation of the error function to arrive at

$$\phi(x) = \frac{B\rho b^2\sqrt{\pi}}{C\beta}\begin{cases} \frac{1}{48}\Big[(C^2\pi - 8Cl\sqrt{\pi} - 8x^2)2x\sqrt{C^2\pi - 4x^2} \\ +(4l - C\sqrt{\pi})C^3\pi^{3/2}\arccos(\frac{2x}{C\sqrt{\pi}}) \\ +16x^2(C\sqrt{\pi} + 2l)\log\left(\frac{C\sqrt{\pi}+\sqrt{C^2\pi-4x^2}}{2x}\right)\Big] & \text{if } x < \sqrt{\pi}C/2 \leq l, \\ \frac{1}{3}\Big[(l^2 - 2Cl\sqrt{\pi} - 2x^2)x\sqrt{l^2 - x^2} \\ +(C\sqrt{\pi} - l)l^3\arccos(\frac{x}{l}) \\ -(2l + C\sqrt{\pi})x^3\log\left(\frac{x}{l+\sqrt{l^2-x^2}}\right)\Big] & \text{if } x < l < \sqrt{\pi}C/2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Therefore, the region where the acceptor cells receive the protein is limited by both the maximum length of the TNs and also by the mean diffusion length. The maximum protein levels seen at the boundary can be evaluated from Equation (2.5) as

$$\phi(x)_{x\to 0} = \left(\frac{B\rho b^2\sqrt{\pi}}{C\beta}\right)\begin{cases} \frac{l^3(\sqrt{\pi}C - l)}{3} & \text{if } C \geq \frac{2l}{\sqrt{\pi}} \\ \frac{(4l - \sqrt{\pi}C)C^3\pi^{3/2}}{48} & \text{Otherwise} \end{cases} \tag{2.6}$$

The rate of decline of protein levels with distance in the acceptor cells at the boundary is also dependent on both the maximum length of the TNs and the mean diffusion length.

$$\left[\frac{d\phi(x)}{dx}\right]_{x\to 0} = \left(\frac{B\rho b^2\sqrt{\pi}}{C\beta}\right)\begin{cases} 2l^3(1 - \frac{C\sqrt{\pi}}{2l}) & \text{if } C \geq \frac{2l}{\sqrt{\pi}} \\ \frac{C^2\pi}{24}(C\sqrt{\pi} - 6l) & \text{Otherwise} \end{cases} \tag{2.7}$$

Interestingly, we see that while greater mean diffusion length increases the observed levels of the transferred molecules transferred adjacent to the donor cells(Equation (2.6)), it also sharpens the fall in the concentration of the molecule as we move farther from the donor cells(Equation (2.7)).

To calculate the mean diffusion lengths, we assume that the diffusion coefficient follows the Einstein-Stokes equation $D = \frac{k_B T}{6\eta\beta r}$, where r is the radius (or the effective Stokes radius for nonspherical particles) of the diffusing molecule, k_B is the Boltzmann's constant, T the absolute temperature,

and η the viscosity of the medium. We considered a few representative molecules with varying sizes, the diffusion coefficients of which are tabulated in Table 2.2.

Table 2.2: Table detailing the diffusion coefficients of various molecules simulated by the model in Figure 2-4.

Molecule	Stoke's radius	Diffusion coefficient D
Glucose		500 μm (Groebe et al., 1994)
Dextran (3 kDa)	13 Å (Nicholson and Tao, 1993)	37 μm
Dextran (10 kDa)	23 Å (Nicholson and Tao, 1993)	20 μm
Dextran (40 kDa)	73 Å (Nicholson and Tao, 1993)	6.5 μm
GFP	23 Å (Nicholson and Tao, 1993)	20 μm
Cytochrome C	$23\text{Å}\sqrt[3]{\frac{12\text{ kDa}}{30\text{ kDa}}} = 17\text{Å}$, estimated from GFP	28 μm
Legumainpreprotein (AEP)	$23\text{Å}\sqrt[3]{\frac{433\text{ amino-acids}}{238\text{ amin-acids}}} = 28\text{Å}$, estimated from GFP	17 μm

According to Gregor et al. (2005), and Luby-Phelps (1986), the viscosity of the cytosol is approximately 4 times of that of water, therefore, we simulated our model with the value of value of $\eta = 4.2 \times 10^{-3}\text{kg m}^{-1}\text{s}^{-1}$.

2.4 Results

2.4.1 Range of profusion of the transferred molecule

Simulating our model for protein transfer by TN from donor to acceptor cells for membrane proteins, we observe that the membrane proteins can be transferred into the acceptor cells within the distance up to the maximum length of the TN, l. Also, the decline of the protein levels is approximately linear with the distance from the boundary of the donor cells.

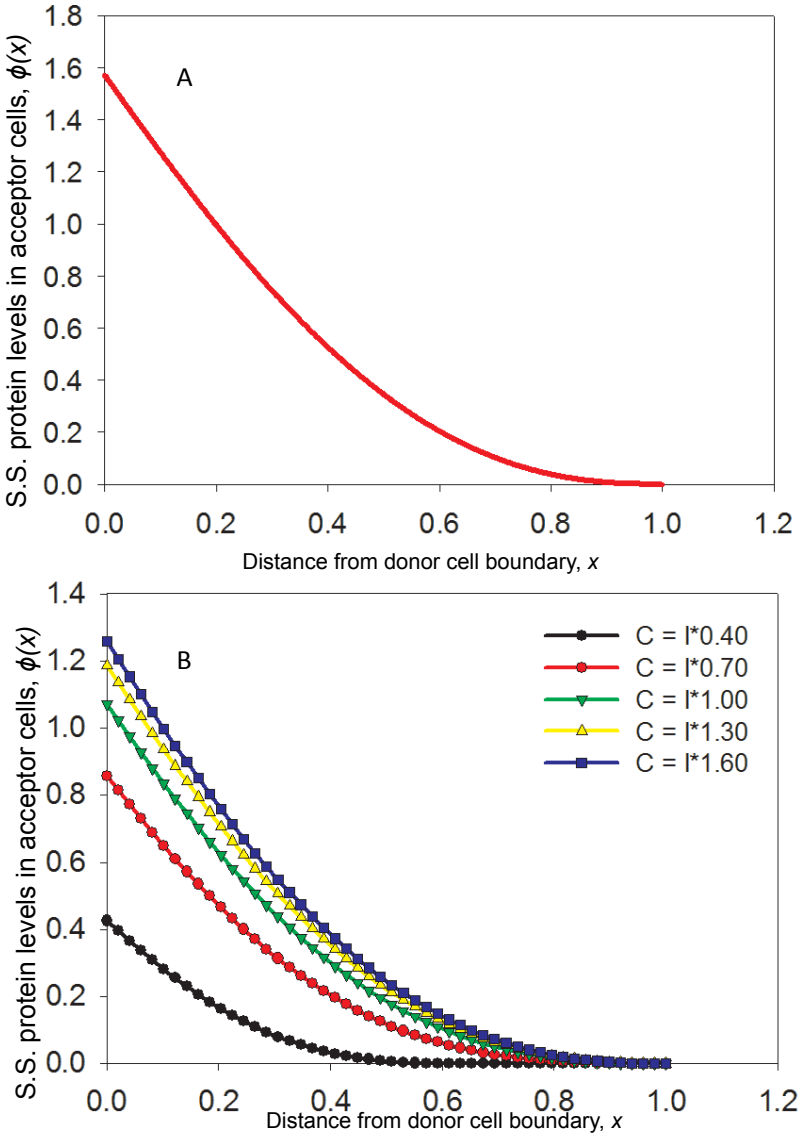

Figure 2-3: Simulated transfer of molecules from donor to accepted cells via TN. Caption continued on the next page.

When simulated for cytoplasmic proteins, the model predicts a similar profile for the levels of trans-

Figure 2-3: (Continued caption) Simulated transfer of molecules from donor to accepted cells via TN. (**A**) Transferred membrane-bound molecule levels in acceptor cells at a distance x from the boundary of the donor cells region, calculated from Equation 2.3. The distances used in the plot are in units of the maximum TN length l. The levels of transferred membrane molecules are given in units of $\frac{Al^3\rho\pi b^2}{3\beta}$ where A is a constant related to the physics of the membrane contact and level in the donor cells, ρ the density of the donor cells, and b the radius of the cells. The level is maximal at the boundary and gradually decreases to zero at a distance equal to the maximum TN length l, after which no TN connection can be made between the donor and acceptor cells. (**B**) Transferred cytoplasmic molecule levels in acceptor cells at a distance x from the boundary of the donor cells region, calculated from Equation 2.5 for various values of the mean diffusion length (C). The distances (x) used in the plot are in units of the maximum TN length l. The level of transferred molecule levels are given in units of $\frac{Bl^3\rho\pi b^2}{3\beta}$ where B is a constant related to the level in the donor cells, ρ the density of the donor cells, and b the radius of the cells. The values of the mean diffusion length C considered are specified as fractional multiples of the maximum TN length l. More of the cytosolic molecule is transferred for larger mean diffusion lengths. Both the amount of molecules transferred to a particular distance and the maximal distance to which it is transferred is limited by the mean diffusion length. The mean diffusion length itself may depend on both the diffusion constant and mean time of stable TN formations (Equation 2.4), which is explored further in Figure 2-4.

ferred cytoplasmic molecules into the acceptor cells (Figure 2-4). However, the amount of transferred molecules, as well as the distance over which they are effectively transferred also depends on the mean diffusion length. This can be attributed to the fact that cytoplasmic constituents, during a transient TN formation, exist in the form of concentration gradient with the highest concentration in the location of cytoplasm before TN was formed. The concentration of cytoplasmic constituents is lowest at the tip of the TN in connection with the recipient cells (Figure 2-3). Similar to the case of membrane-bound molecules, the fall in the levels of transferred cytoplasmic molecule is approximately linear with distance from the donor cell region boundary (Figure 2-4). We then compared model predictions for a number of different biomolecules detailing the efficiency of transfer into the acceptor cells after a steady state of the transfer process is reached. For cytoplasmic molecules, both the size and the duration of stable TN connection were found to determine the levels of transferred molecules.

2.4.2 Possible mechanisms for the regulation of TN molecular transfer

A number of regulatory mechanisms for the transfer of molecules across TNs are consistent with our model. TN length and its stability can be modulated experimentally by stabilizing the actin cytoskeletal assembly forming the TNs. The model predicts that the stability of TN, and thereby connection of donor and recipient cells, will have a significant effect on the level of cytoplasmic molecules in donor cells. The length of time TN connections are made will also influence the transfer of membrane bound molecules, with longer stable connections leading to higher membrane molecule

Figure 6

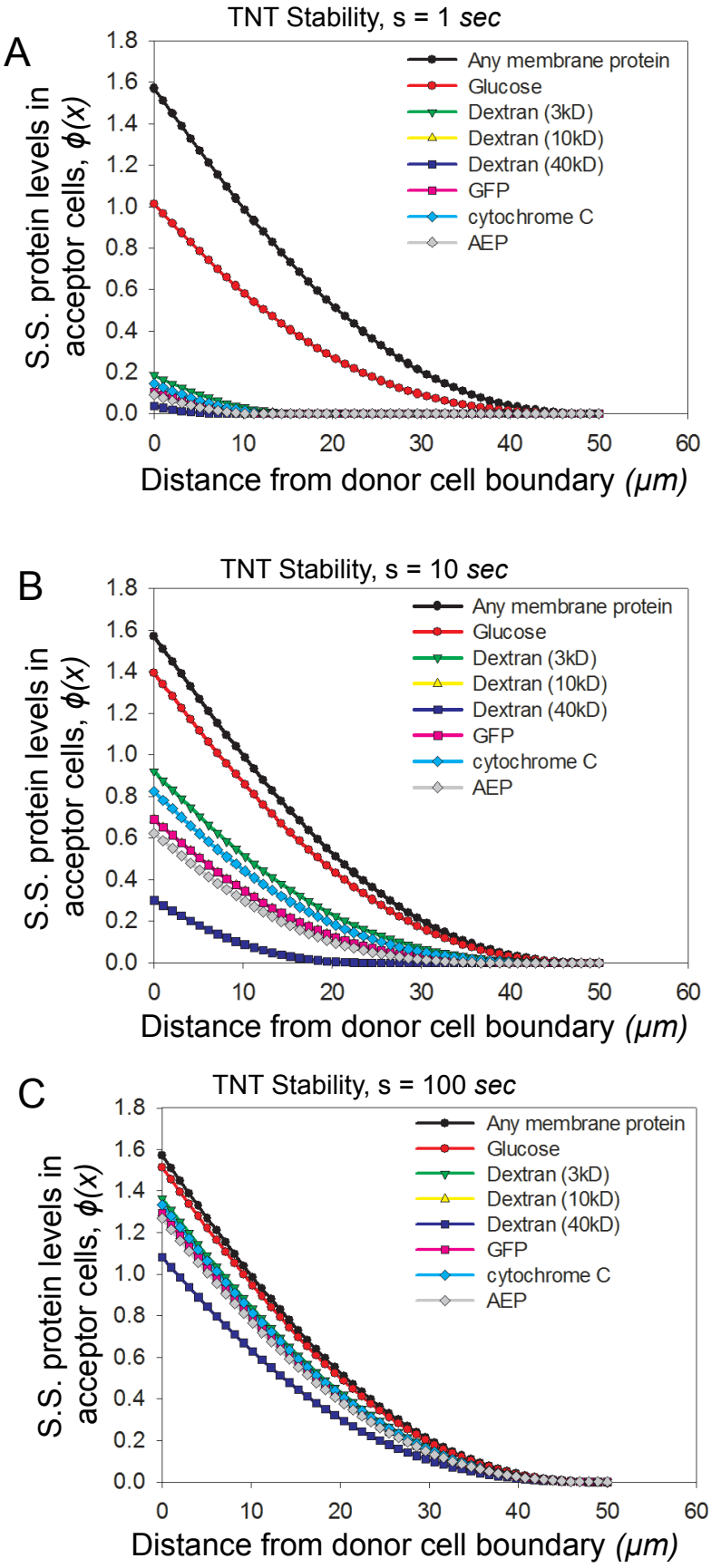

Figure 2-4: Effect of the stability of TNs. Caption continued on the next page.

Figure 2-4: (Continued caption) Stability of TN connection determines extent of cytoplasmic protein transfer in a size dependent manner. Level of transferred molecules transferred from donor to acceptor cells via TN formation, simulated to be stabilized for mean durations of (A) 1 second, (B) 10 seconds, and (C) 100 seconds. In all cases, levels of transferred molecules in the recipient cells are plotted against the distance x from the region of donor cells, calculated from Equations refeq:memss and refeq:meandiffl. The levels of transferred molecules are given in units of $\frac{Al^3\rho\pi b^2}{3\beta}$ for membrane bound molecules, and $\frac{Bl^3\rho\pi b^2}{3\beta}$ for cytosolic molecules similar to Figure reffig:fig3. In all cases, size of TN is considered to be 50ţm

transfer. Changes in the transport of the molecule through the Golgi apparatus, micro-vesicles, and endosomes to and from the membrane and the effect of post translational control of its binding with the membrane in the donor cells will also determine the amount of membrane-bound molecule available for transfer on the TNs. These changes in membrane recruitment and binding could be modulated in the cells by various pathways.

2.4.3 Effect of the size of the transferred molecules

The model predicts that small molecules are quite robust in their transfer across the TNs while larger proteins require favorable conditions, for example, stable TN that retract after longer durations. Our model predicts that, in general, in a typical TN that exists for a few seconds to tens of seconds, transfer of membrane proteins will be appreciably higher than cytoplasmic molecules. Among cytoplasmic molecules, large molecules will have an extremely low transfer efficiency, with transfer occuring only within cells that are extremely close to each other (Figure 2-4A). The difference of extent of transfer between small and large molecules is quite pronounced, suggesting that the size of molecules plays a significant role in cytoplasmic protein transfer between cells.

Small cytosolic molecules, for example, glucose (Groebe et al., 1994) or other metabolites, transfer at a much faster rate. Thus the model explains previous observations that dextran molecules of different sizes showed a size dependent intercellular transfer amounts between dextran containing Chinese Hamster Ovarian (CHO) cells to those without dextran (Figures 2-4A, B). Since typical cytosolic proteins are much larger than smaller metabolites (e.g., green fluorescent protein used as a reporter probe in cell biology experiments, which has a Stoke's radius of 23 Å, similar to a 10 kDa Dextran (Phillips, 1997)), the extent of their transfer is expected to be much lower, suggesting a possible reason why the detection of cytosolic protein transfer has been rare, or has remained unreported.

However, as the TN becomes more stabilized, the diffusion of cytoplasmic molecules within the TN

shaft shifts more towards a steady state, and becomes shallower. This results in higher transfer of cytoplasmic molecules (Figure 2-4B). As TNs are stabilize even further, the extent of transfer of cytoplasmic molecules approaches the transfer of membrane proteins for all distances between donor and recipient cells (Figure 2-4C). The difference of transfer within cytoplasmic molecules of different sizes becomes less pronounced, suggesting that size remains a smaller factor in cytoplasmic protein transfer as TNs stabilize (Figure 2-4C).

2.5 Discussion

Intercellular transfer of cellular components is a fascinating phenomenon largely because it is understood so little, and does not seem to obey any known classical cell-cell communication mechanisms. This transfer also seems surprising since it suggests a phenomenon over which the recipient cells can only have partial or no active control. Even then, it has been implicated as important in various physiological and pathological contexts. Physiological heterotypic and homotypic cell-cell interactions occur frequently in various tissues, including blood vessels, muscles, and nerve fibers, etc., potentially allowing transfer of molecules between cells. Pathological contexts such as cancer present new avenues for intercellular transfer of biomolecules to play significant roles. For example, drug-resistant cancer cells could transfer small molecules conferring drug-resistance to neighboring drug-susceptible cells causing lateral transfer of resistance. However, in spite of the importance of this passive form of intercellular communication, our present understanding about it is limited.

In spite of the relatively few reports of intercellular transfer of biomolecules, largely due to the small amounts of transfer that can frequently go undetected, a few trends stand out in the reported studies. It has been observed that the membrane bound molecules could transfer more readily from one cell to another, but cytoplasmic molecules transfer has been reported less frequently. Interestingly, while there are no reports that the efficiency of transfer of membrane-integrated molecules is dependent on the molecular weight of the transferred components, cytoplasmic molecules have been reported to transfer in a size dependent manner. In addition, there has not been any conclusive demonstration of cytoplasmic protein transfer between cells. Here, we propose a mathematical model describing intercellular transfer of biomolecules via TNs that explains these observations and makes useful predictions.

The model makes a critical assumption about distinct characteristics of transport of membrane vs. cytosolic components through TNs. In particular, it is assumed that large cytosolic components, such intracellular proteins, can diffuse over the length of TN much slower that the membrane components. Thus the rate limiting step in the transfer of the membrane components is the rate of their access to TNs on the donor cell side, whereas the rate limiting step of the transport of cytoslic components is their diffusion over TN. As a consequence and due to the transient nature of TNs, membrane but not cytosolic components would reach a steady state distribution over the length of a TN, with cytoslic components forming a diffusion based gradient. Since diffusion is dependent on the size of the molecule, and consequently results in a size/mass dependent transfer of cytosolic

molecules.

For both the membrane bound and cytosolic molecules, the transfer is limited by the spatial separation between an acceptor cell and the nearest donor cells. Due to the size dependence of the cytosolic molecular transfer, both the amount of the molecule transferred and the maximum separation between the cells that allows for any observable transfer becomes negligible for large proteins. A consequence of the model then is that in most physiological contexts, any signaling happening across cells in this fashion is limited to either membrane bound molecules or small cytosolic molecules. Thus, our model provides a physical basis for the observation of signaling by membrane proteins and small cytosolic molecules.

The model predictions regarding the importance of the length of the TNs the time scale of TN lifetimes open new avenues for of the analysis of intercellular communication through individual TN formations. In addition, it raises the question of whether there could be specific pathways regulating the formation and behavior of such TNs. For example, it has been reported that HIV induces the formation of TNs in macrophages (Eugenin et al., 2009). This hypothesis can be tested by modulating the frequency of TN formation by cells, achievable by chemical and environmental means (Lou et al., 2012). Another hypothesis generated by this model is that increased stability of TNs could reduce the differential transfer of molecules of different sizes, and this can be tested by modulating TN stability (Marzo et al., 2012). Recent reports of transfer of endocytotic organelles due to TNs, which can be controlled by a number of molecular signals (Gurke et al., 2008), suggest a scope for more detailed theoretical models than the one presented here, taking into account the active and modulated TN formation frequency and dynamics and how they affect the transfer of components in response to specific biological signals. It is plausible that TN formation may be regulated by cells as response to various stresses or other stimuli, resulting in a controlled selection of the nature, size, and amount of the transferred components and the corresponding phenotypes.

2.6 Directions for future experimental studies

The model presented here, in the absence of precise values for the multitude of physical parameters involved in the process, makes a number of assumptions in order to provide some qualitative predictions. Careful experimental studies may validate or correct certain aspects of this model. These predictions and assumptions should help to tease out the role the transfer of molecules across TNs plays physiologically.

We have a simplistic linear relation between the length of TNs and their abundance in a uniform density of cells (Equation (2.1)). The elongation of the TNs proceeds through actin (Wittig et al., 2012) and/or microtubule (Wang et al., 2012) polymerization; it is therefore intuitive that TN forming cells will be able to survey their immediate neighborhood more exhaustively than the relatively distal regions. However, imaging a large number of cells forming TNs could help to provide us with a better understanding of the dynamic of TN formation and their static distribution. Our model can thus be updated with a more accurate length distribution of TNs during growth. We have also ignored the effect of the narrow TN channel on the diffusion coefficient. However, this should be an additional factor that is absorbed into the mean diffusion length, a parameter we have handled algebraically while deriving the expressions for concentrations of the transferred molecule.

Our model derives an analytical expression for the concentration of molecules reaching into a population of cells from another population of TN forming plated next to them. We derive simple expressions for the slope of the concentration with distance from the boundary of the two cell populations for membrane bound and cytosolic molecules. Further, we show how the size of the cytosolic molecules should determine the extent of profusion; while for membrane bound molecules, it is their concentration on the TN membrane rather than size that should determine the extent of profusion. These are qualitative predictions that can be tested independent of the exact measurement of various parameters involved in the model. Experiments measuring the profusion of cytosolic and membrane bound molecules of differing sizes can shed light on the validity of our models of diffusive transfer of cytosolic molecules and membrane transfer at the tip of the TNs. We hope that with further experimental evidence, our model can be refined to better reflect the physiological mechanics of TN formation and molecule transfer.

As mentioned earlier, experiments perturbing the frequency and stability of the TNs provide another avenue for testing the model and at least one possible regulatory mechanism.

Once certain quantitative characteristics of the transfer of these molecules have been verified for some control molecules, any deviation from these transfer rates for physiologically important proteins opens the way for investigating the signaling pathways the cells employ for regulating this intracellular traffic.

Chapter 3

Link prediction in protein interaction networks: graph diffusion and the Ising model

3.1 Introduction

3.1.1 Problem definition

Characterizing all the genes that are discovered in a genome or expressed as proteins is an important goal of systems biology. The physiological influence of genes and proteins, however, is often the result of physical or other functional interactions between genes, proteins, and other biomolecules, rather than individual contributions of isolated components. Interaction networks are often defined as graphs with genes or proteins as vertices, which often have labels describing membership in defined groups, and interactions as pairwise edges between vertices, which may be of different types, directed or undirected, and weighted or unweighted. These networks, whether built from a compilation of small studies or the result of high throughput studies, are often noisy, with both false positive spurious interactions and false negatives absent from the network.

The problem that we address in this study is predicting missing interactions using the rest of the tested interactions in the network. The benefit of using this assessment, rather than prediction of

vertex labels, is that interactions can be experimentally measured whereas gene annotations are less rigorously defined. Link prediction generally assumes a network where genes with similar functions or activities are more likely to connect to similar genes than to the rest of the network. We focus on two approaches to this problem: the random walk or graph diffusion kernel, and the Ising model. We show how these models are related and compare a number of approximate solutions of these models in the context of link prediction.

3.1.2 Motivation

The human genome has about $20,000$ protein-coding genes while the much simpler common yeast (*Saccharomyces cerevisiae*), an awesomely powerful model organism for understanding eukaryotic biology, has about $6,000$ protein-coding genes. Detailed experimental studies to elucidate the biology of each gene using conventional molecular and cellular biology techniques such as fluorescent microscopy for protein localization, co-immunoprecipitation, blotting, and phenotyping knock-outs, are labor and resource intensive and have only been performed for a subset of genes. Systems biology is an attempt to understand emergent phenomena in physiology, development, and disease that arise from interactions between various genetic and environmental variables.

Central to the study of systems biology in creating predictive models is the ability to map the interactions between genes. The types of interactions commonly studied include physical protein binding or protein-protein interactions (PPI), direct regulation of gene expression by transcription factors (regulatory interactions), and genetic interactions defined by non-additive phenotypes such as lethality or slow growth of double mutants when the individual mutants have high fitness. This study relates primarily to undirected, unweighted networks, a frequently used model for PPI networks. These can be considered as graphs with nodes or vertices defined by proteins and edges defined by existence of a protein-protein interaction. However, the developed techniques can be directly applied to link prediction in other undirected networks, and with some modifications, to weighted networks, directed networks, and networks with multiple edge types.

High-throughput protein interaction studies using techniques such as yeast two-hybrid screens (Fields and Song, 1989) and tandem affinity purification (Puig et al., 2001) have been used for systematic studies to generate protein-protein interaction (PPI) networks with large coverage of proteins in species such as yeast (Fields et al., 2000; Ito et al., 2001), *Escherichia coli* (Rajagopala et al., 2014), fruit fly (Giot et al., 2003) and human (Rual et al., 2005; Rolland et al., 2014). In addition, bioin-

formatics databases such as BioMart (Haider et al., 2009) or MIPS (Pagel et al., 2005), and species specific portals such as for yeast (Cherry et al., 1998), bacteria (Su et al., 2008), and human (Stelzl et al., 2005) collect and curate interactions both from publications reporting individual interactions and high throughput systematic studies.

These PPI interactomes, along with other interaction networks, have been used to understand the systems biology of various biological phenomena, and to make specific predictions and generate testable hypotheses about gene function. Interaction networks have been used to predict protein functions (Sharan et al., 2007; Chua et al., 2006) and associate genes with disease (Köhler et al., 2008; Oti et al., 2006), search for causal genes in eQTL regions (Suthram et al., 2008), refine regulatory networks (Nariai et al., 2003), assemble the temporal order of pathways (Farach-Colton et al., 2004), and expand and understand gene lists in context (Lachmann and Ma'ayan, 2010).

Even the high-throughout studies, however, usually only capture a subset of all the interactions in an organism. Stumpf et al. (2008) estimated the true sizes of the interactomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Drosophila melanogaster*, and *Homo Sapiens*, assuming random sampling of proteins included in various systematic studies. While the true networks are probably very sparse, we have not been able to sample all the interactions. Deane et al. (2002) used gene expression and interactions among paralogous genes to estimate that only about half of the interactions found in high throughput studies are physiologically relevant. The known interactome thus compiled from these studies may have under-sampled some proteins while oversampling others, and there are unequal sampling biases in distinct high throughput studies (von Mering et al., 2002; Bader et al., 2002). This sampling bias affects our estimation of the global properties of the interaction networks (Han et al., 2005). Huang et al. (2007) modeled protein specific promiscuities in the prey-bait counts of yeast two-hybrid experiments to arrive at false-positive rates of 25% to 45% and false-negative rates of 75% to 90% in the yeast, worm, and fly datasets.

3.1.3 Related methods for predicting links and improving interaction network quality

Computational approaches to improve the quality of experimentally measured interactions and predict additional interactions have followed two general approaches, depending on the information used: biological approaches based on experimental and computed properties of individual proteins, and statistical approaches based on the observed interaction network. In this section we briefly

review some studies that have used one or both of these approaches to improve interaction networks, and then focus on some of the common themes in network analysis techniques for link prediction.

Some studies use additional biological information like gene expression, co-expression of transcripts or co-existence of proteins in the same cell type or cellular localization, and additional statistical characterization of the results of different high throughput studies to classify pairs as false negatives and false positives, thus improving the network. Kemmeren et al. (2002) use differential co-expression and Matthews et al. (2001) use cross-species homologs to predict new interactions.

Other studies have used the topological structure of the interaction networks to predict likely missing links. These methods assume that interacting pairs of nodes are more well connected to each other than chance through paths of length two or greater. Biological networks have been shown to be scale-free, modular, and hierarchical (Barabási and Oltvai, 2004; Ravasz and Barabási, 2003; Ravasz et al., 2002). Cellular physiology is thought to arise from a modular organization of pathways and interaction networks (Hartwell et al., 1999). It is known that interaction networks are more clustered than predicted by chance. Kashtan and Alon (2005) argued that the evolutionary model of gene duplication followed by divergence naturally results in an interaction network which looks modular and where interacting pairs share many common neighbors. Indeed, the presence of community structure is a common property of various real-world networks (Girvan and Newman, 2002) and is a common framework for the study on complex networks. Przulj et al. (2004) showed that PPI networks are better modeled by embedding the nodes in a metric space and drawing connections between nearby nodes than the scale-free or Erdos-Renyi random graphs.

One way to conceptualize link prediction in the context of network topologies is to assign to each pair of nodes in an interaction network a distance or similarity score. This distance measure depends on our model of the graph topology, but nodes deemed more similar or closer to each other are more likely to interact. Multiple measures of common neighbors between nodes have been used as measures of similarity, with different normalizations based on the node degree (Kossinets, 2006; Ravasz et al., 2002; Newman, 2001; Saito et al., 2003). Goldberg and Roth (2003) modeled interactions assuming a small-world network rather than explicit modularity assign confidence scores to individual interactions from a high-throughput interaction study. By fitting the hypergeometric distribution to the number of common neighbors of a vertex pair, they introduced the idea of the mutual clustering coefficient as a local distance measure. Zhou et al. (2009) compared a number of these local measures of node similarity for predicting missing links in PPI and other networks.

Chen et al. (2005) assessed the reliability of PPI interactions by considering the shortest weighted alternative path connecting the two proteins. Other methods assign reliabilities to interactions by building a model of the whole interactome rather than a simple combination of terms involving the closest neighbors. Guimerà and Sales-Pardo (2009) tested the reliability of links and networks using a fully Bayesian stochastic block model of the network. Pan et al. (2016) predicted links by learning the likelihoods of cycles of various lengths. Exclusive community structure (Zheleva et al., 2008) and hierarchical structure (Clauset et al., 2008) have also been used to predict missing links.

Others have combined biological information with network statistics to improve network quality. Bader et al. (2004) use a number of distance measures in networks derived from different yeast two-hybrid and co-immunoprecipitation studies as predictors to arrive at a confidence score for each possible interaction. Other studies have combined the topological structure and knowledge of specific proteins to arrive at reliable links. For example, Ahmed and Glasgow (2014) use both protein domain knowledge and common neighbors in a particle swarm optimization framework to predict high confidence interactions.

Barzel and Barabási (2013) predicted direct interactions between genes and proteins using the correlation matrix of gene expression patterns in *E. coli*. Unlike the other works discussed here, their study did not predict missing links in the same network used as an observation to the algorithm. However, it uses the notion of direct interactions versus correlations that we also employ in the Ising model method here.

In this study, we are especially interested in the class of similarity functions over nodes called graph diffusion kernels (Kondor and Lafferty, 2002), which are often motivated by probability distributions describing random walks on graphs. Graph diffusion kernels became very popular after their commercially successful use in web search engines (Brin and Page, 1998). Qi et al. (2008) used graph diffusion kernels to predict links in protein and genetic interaction networks, which included signed edges akin to ferromagnetic and anti-ferromagnetic interactions in a spin lattice. The graph diffusion kernel can be described as the probability of a random walker starting at one vertex and end at another vertex, while following certain rules for stochastic transitions across network edges. Variant forms of the graph diffusion kernel are motivated by slightly different physical analogies and mathematical construction. They may be constructed as the equivalent electrical resistance in a graph composed of resistors (Klein and Randić, 1993), the diffusion of heat (Chung, 2007), or as the probability that random walkers starting from two nodes meet at some third node (Jeh and

Widom, 2002), among others. These functions can be mathematically described as the weighted sum of paths of various lengths through the network, or the probability distribution for random walkers that decays over time. Related functions on graphs have appeared in other work (Katz, 1953; Leicht et al., 2006), although these functions were not presented as graph diffusion kernels.

Here we draw connections between the stochastic model of graph diffusion and the equilibrium statistical physics of the Ising model, introduced to model regular lattices of magnetic dipoles in solids (Ising, 1925). Ising models have a Hamiltonian that is defined by signed and weighted edges connecting pairs of vertices, each having a spin variable with states $+1$ and -1, and possibly also coupled to an external field. The Ising model has been used to model communities (Son et al., 2006) and propagation dynamics (Grabowski and Kosiński, 2006; Fraiman et al., 2009) across networks. We explore isomorphisms between the spin-spin correlation function and the graph diffusion kernel to connect existing approaches and develop new approximations to identify strongly coupled vertices in the Ising sense, and likely interactions in the PPI domain. We apply the Ising model and our new approximations to general graphs arising from interaction studies and other real-world networks rather than regular or semi-regular lattices observed in materials science or statistical physics.

3.1.4 Link prediction in other contexts

Link prediction in complex networks is also useful for domains other than protein interaction networks and biological networks. Link prediction has been used to predict links in social network of friendships or relationships between people (Fire et al., 2013). Liu and Ning (2011) used link prediction on a bipartite network to rank candidates for employment positions. Collaborative filtering, a commercially important problem that seeks to recommend a product that is likely to appeal to an individual, can be cast as a link prediction problem in a bipartite network of consumers and products (Huang et al., 2005; Li and Chen, 2013). Link prediction can be used to predict likely links or followers in online social networks (Valverde-Rebaza and de Andrade Lopes, 2013). Kastrin et al. (2016) used link prediction in a network of scientific literature to generate hypotheses in the form of related biomedical concepts.

3.1.5 The approach used in this study

In this study, we relate graph diffusion kernels to the Ising model. We show how the Ising model can be used for link prediction and why the two approaches are intimately related. We derive new analytical approximations for the spin-spin correlation function in an Ising model, and we demonstrate that our approximations are overall better for link prediction than either graph diffusion kernels or existing Ising model approximations. We also present a theoretical argument supporting the superior performance of our network analysis method. Our new method could be a powerful new approach for related problems spanning many fields of network science: gene candidate prioritization, graph clustering, and community detection.

3.2 Methods

3.2.1 The Ising model

The Ising model is defined for a graph with spins on vertices (s_u) and pairwise interactions as edges (J_{uv}) with the the Hamiltonian (or energy function):

$$H(\mathbf{s}) = -\sum_{u\sim v} J_{uv} s_u s_v - \sum_u h_u s_u,$$

where each spin variable s_u can take the values $+1$ and -1.

The equilibrium probability of a state, defined by the configuration of spins, is given by the Gibbs-Boltzmann distribution,

$$\begin{aligned}
\Pr(\mathbf{s}) &= \frac{\exp[-\beta H(\mathbf{s})]}{\sum_{\mathbf{s}'} \exp(-\beta H(\mathbf{s}'))} \\
&= \frac{\exp\left[\beta(\sum_{u\sim v} J_{uv} s_u s_v + \sum_u h_u s_u)\right]}{\mathrm{Tr}_{\mathbf{s}'} \exp\left[\beta(\sum_{u\sim v} J_{uv} s'_u s'_v + \sum_u h_u s'_u)\right]} \\
&= \frac{\exp\left[\beta(\sum_{u\sim v} J_{uv} s_u s_v + \sum_u h_u s_u)\right]}{Z}.
\end{aligned} \tag{3.1}$$

where $Z = \mathrm{Tr}_{\mathbf{s}} \exp\left[\beta(\sum_{u\sim v} J_{uv} s_u s_v + \sum_u h_u s_u)\right]$ is the partition function.

By symmetry, the expected value of any spin is 0 when the external field h_u is 0. For a non-zero external field, the expectation is

$$\langle s_u \rangle = \frac{\sum_{\mathbf{s}} \exp[-\beta H(\mathbf{s})] s_u}{\sum_{\mathbf{s}} \exp[-\beta H(\mathbf{s})]}.$$

An exact calculation requires a full enumeration of each possible state, 2^N for a lattice with N spins. For most lattices, approximations are required. A very useful approach is the mean field approximation, in which the average is effectively moved from the outside to the inside of the

exponential:

$$\begin{aligned}\langle s_u \rangle &= \sum_{s_u} s_u \mathrm{P}(s_u) \\ &= \Pr(s_u = 1) - \Pr(s_u = -1) \\ &\approx \frac{\exp(\beta(\sum_{v \neq u} J_{uv}\langle s_v \rangle + h_u)) - \exp(-\beta(\sum_{v \neq u} J_{uv}\langle s_v \rangle + h_u))}{\exp(\beta(\sum_{v \neq u} J_{uv}\langle s_v \rangle + h_u)) + \exp(-\beta(\sum_{v \neq u} J_{uv}\langle s_v \rangle + h_u))} \\ &= \tanh(\beta(\sum_{v \neq u} J_{uv}\langle s_v \rangle + h_u)).\end{aligned}$$

Under zero external field ($h_u = 0, \forall u$), one of the solutions to this equation is $s_u = 0, \forall u$. This naïve mean field method was extended to include the effect of pair interaction correlations by the Bethe mean field method (Bethe, 1935). Yedidia (2001) showed a general method to include larger correlations using Plefka's formulation of the Gibbs free energy (Plefka, 1982).

3.2.1.1 Spin correlations

One measure of distance between nodes in a graph is the correlation of spins, $\langle (s_u - \langle s_v \rangle)(s_v - \langle s_v \rangle) \rangle = \langle s_u s_v \rangle - \langle s_u \rangle \langle s_v \rangle$. Under zero external field and magnetization, this reduces to $\langle s_u s_v \rangle$.

3.2.1.1.1 A non-linear mean field solution using partial relaxation

Now let us consider the correlations $\chi_{uv} = \langle s_u s_v \rangle$ between two nodes u and v under 0 external field, in which each spin has expectation 0. For convenience, we denote the rest of the nodes as $\mathbf{R} = \{r : r \notin (u, v)\}$. We then consider separately the terms of the Hamiltonian for the spins of interest

$$H_0 \equiv -h_u s_u - h_v s_v - s_u J_{uv} s_v,$$

the terms of the Hamiltonian for the rest of the system,

$$H_R \equiv -\sum_{r \in \mathbf{R}} h_r s_r - \sum_{r \sim r' \in \mathbf{R}} s_r J_{rr'} s_{r'},$$

and the terms connecting the system to the rest of the network,

$$V \equiv -s_u \sum_{r \in \mathbf{R}} J_{ur} s_r - s_v \sum_{r \in \mathbf{R}} J_{vr} s_r,$$

with $H = H_0 + H_R + V$.

The vector of spins $\mathbf{s}$ is similarly considered separately as two blocks of the spins of interest $\mathbf{s}_{uv} = [s_u\,, s_v]^T$ and the rest $\mathbf{s}_\mathbf{R} = [..., s_r, ...]$ where $r \in (\mathbf{R}$ and $r \notin (u, v)$.

The partition function can then be written as

$$\begin{aligned} Z &= \mathrm{Tr}_\mathbf{s} \exp\left[-\beta H_0(\mathbf{s}_{uv})\right] \exp\left[-\beta H_R(\mathbf{s}_R)\right] \exp[-\beta V(\mathbf{s}_{uv}, \mathbf{s}_R)] \\ &= \mathrm{Tr}_{s_u s_v} \exp\left[-\beta H_0(\mathbf{s}_{uv})\right] \mathrm{Tr}_{\mathbf{s}_R} \exp\left[-\beta H_R(\mathbf{s}_R)\right] \exp[-\beta V(\mathbf{s}_{uv}, \mathbf{s}_R)] \end{aligned}$$

Consider the Boltzmann weight obtained by summing over the rest of the system,

$$\begin{aligned} &\mathrm{Tr}_{\mathbf{s_R}} \exp[-\beta H_R(\mathbf{s_R})] \exp[-\beta V(\mathbf{s}_{uv}, \mathbf{s_R})] \\ =&\mathrm{Tr}_{\mathbf{s_R}} \exp[-\beta H_R(\mathbf{s_R})] \frac{\mathrm{Tr}_{\mathbf{s_R}} \exp[-\beta H_R(\mathbf{s_R})] \exp[-\beta V(\mathbf{s}_{uv}, \mathbf{s_R})]}{\mathrm{Tr}_{\mathbf{s_R}} \exp[-\beta H_R(\mathbf{s_R})]} \\ =&Z_\mathbf{R} \langle \exp[-\beta V(\mathbf{s}_{uv}, \mathbf{s_R})] \rangle_\mathbf{R} \\ =&Z_\mathbf{R} \left\langle \exp\left[\sum_{r \in \mathbf{R}} \beta(s_u J_{ur} s_r + s_v J_{vr} s_r)\right] \right\rangle_\mathbf{R} \end{aligned} \tag{3.2}$$

We wish to arrive at a self-consistent equation for χ without summing over all the discrete states ($\mathrm{Tr}_\mathbf{s}$). Note that what makes the problem hard is the discrete nature of the spins; a quadratic form could be calculated exactly. We therefore approximate the discrete $\mathbf{s_R}$ with a continous approximation of a multivariate Gaussian distribution with means $m_r = \langle s_r \rangle = 0$ and covariance matrix given by $\langle s_r s_{r'} \rangle - \langle s_r \rangle \langle s_{r'} \rangle = \chi_{r,r'}$ where $r, r' \in \mathbf{R}$. Using this approximation, Equation (3.2) can be approximated as,

$$\begin{aligned} &Z_\mathbf{R} \left\langle \exp\left[\sum_{r \in \mathbf{R}} \beta(s_u J_{ur} s_r + s_v J_{vr} s_r)\right] \right\rangle_\mathbf{R} \\ \approx&Z_\mathbf{R} \int_\mathbf{R} \frac{1}{\sqrt{(2\pi)^{|R|} |\chi_\mathbf{R}|}} \exp\left[-\frac{1}{2} \mathbf{s_R}^T \chi_\mathbf{R}^{-1} \mathbf{s_R}\right] \exp\left[\sum_{r \in \mathbf{R}} \beta(s_u J_{ur} s_r + s_v J_{vr} s_r)\right] d\mathbf{s_R} \\ =&Z_\mathbf{R} \int_\mathbf{R} \frac{1}{\sqrt{(2\pi)^{|R|} |\chi_\mathbf{R}|}} \exp\left[-\frac{1}{2} \mathbf{s_R}^T \chi_\mathbf{R}^{-1} \mathbf{s_R}\right] \exp\left[\beta \mathbf{s}_{uv} \mathbf{J}_{uv \times \mathbf{R}} \mathbf{s_R}\right] d\mathbf{s_R}. \end{aligned} \tag{3.3}$$

Now, since the moment generating function of a zero-mean multivariate normal distribution is

$$\mathbb{E}\left[\mathbf{t}^T \mathbf{X}\right] = \exp(\frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t}),$$

where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We can substitute this result in Equation (3.3) with $\mathbf{t}^T = \beta \mathbf{s}_{uv} \mathbf{J}_{uv}$ to get

$$\begin{aligned}
&\mathrm{Tr}_{\mathbf{s_R}} \exp[\beta H_R(\mathbf{s_R})] \exp[\beta V(\mathbf{s}_{uv}, \mathbf{s_R})] \\
&= Z_{\mathbf{R}} \left\langle \exp\left[\sum_{r \in \mathbf{R}} \beta(s_u J_{ur} s_r + s_v J_{vr} s_r)\right] \right\rangle_{\mathbf{R}} \\
&\simeq Z_{\mathbf{R}} \exp\left[\frac{1}{2}\beta^2 \mathbf{s}_{uv} \mathbf{J}_{uv\times\mathbf{R}} \chi_{\mathbf{R}} \mathbf{J}^T_{uv\times\mathbf{R}} \mathbf{s}^T_{uv}\right].
\end{aligned} \tag{3.4}$$

By definition of the Boltzmann distribution. the correlation between s_u and s_v is,

$$\begin{aligned}
\chi_{uv} &= \frac{\mathrm{Tr}_{\mathbf{s}} s_u s_v \exp[-H(\mathbf{s})]}{Z} \\
&= \frac{\mathrm{Tr}_{\mathbf{s}} s_u s_v \exp[-H_0] \exp[-H_R] \exp[-V]}{Z} \\
&= \frac{\mathrm{Tr}_{s_u, s_v} s_u s_v \exp[-H_0] \mathrm{Tr}_{\mathbf{s_R}} \exp[-H_R] \exp[-V]}{\mathrm{Tr}_{s_u, s_v} \exp[-H_0] \mathrm{Tr}_{\mathbf{s_R}} \exp[-H_R] \exp[-V]}
\end{aligned}$$

Using the approximation of Equation (3.4) in the numerator and denominator, we get

$$\begin{aligned}
\chi_{uv} &= \frac{\mathrm{Tr}_{\mathbf{s}} s_u s_v \exp[H_0] \exp[H_R] \exp[V]}{Z} \\
&= \frac{\mathrm{Tr}_{s_u, s_v} s_u s_v \exp[H_0] Z_{\mathbf{R}} \exp\left[\frac{1}{2}\beta^2 \mathbf{s}_{uv} \mathbf{J}_{uv\times\mathbf{R}} \chi_{\mathbf{R}} \mathbf{J}^T_{uv\times\mathbf{R}} \mathbf{s}^T_{uv}\right]}{\mathrm{Tr}_{s_u, s_v} \exp[H_0] Z_{\mathbf{R}} \exp\left[\frac{1}{2}\beta^2 \mathbf{s}_{uv} \mathbf{J}_{uv\times\mathbf{R}} \chi_{\mathbf{R}} \mathbf{J}^T_{uv\times\mathbf{R}} \mathbf{s}^T_{uv}\right]} \\
&= \frac{\mathrm{Tr}_{s_u, s_v} s_u s_v \exp\left[\beta h_u s_u + \beta h_v s_v + \beta s_u J_{uv} s_v + \sum_{r\sim r' \in \mathbf{R}} \frac{1}{2}\beta^2 s_u J_{ur} \chi_{rr'} J_{r'v} s_v + \sum_{r\sim r' \in \mathbf{R}} \frac{1}{2}\beta^2 s_v J_{vr} \chi_{rr'} J_{r'u} s_u\right]}{\mathrm{Tr}_{s_u, s_v} \exp\left[\beta h_u s_u + \beta h_v s_v + \beta s_u J_{uv} s_v + \sum_{r\sim r' \in \mathbf{R}} \frac{1}{2}\beta^2 s_u J_{ur} \chi_{rr'} J_{r'v} s_v + \sum_{r\sim r' \in \mathbf{R}} \frac{1}{2}\beta^2 s_v J_{vr} \chi_{rr'} J_{r'u} s_u\right]}.
\end{aligned}$$

In the absence of external fields ($h_u = 0$ anhd $h_v = 0$), we get

$$\begin{aligned}
\chi_{uv} &= \frac{\mathrm{Tr}_{s_u, s_v} s_u s_v \exp\left[\beta s_u J_{uv} s_v + \sum_{r\sim r' \in \mathbf{R}} \beta^2 s_u J_{ur} \chi_{rr'} J_{r'v} s_v\right]}{\mathrm{Tr}_{s_u, s_v} \exp\left[\beta s_u J_{uv} s_v + \sum_{r\sim r' \in \mathbf{R}} \beta^2 s_u J_{ur} \chi_{rr'} J_{r'v} s_v\right]} \\
&= \tanh\left[\beta J_{uv} + \sum_{r\sim r' \in \mathbf{R}} \beta^2 J_{ur} \chi_{rr'} J_{r'v}\right].
\end{aligned}$$

If we define $\overline{\tanh}$ is the element-wise tanh operation, then in matrix form this can be written as

$$\chi = \overline{\tanh}\left(\beta \mathbf{J} + \beta^2 \mathbf{J} \chi \mathbf{J}\right),$$

where the only difference is the sum over r_1 and r_2 was not supposed to include the points u and v. This is for the case where $u \neq v$. If $u = v$, we know that $\chi_{uu} = \langle s_u s_u \rangle - \langle s_u \rangle^2 = 1$. Finally,

$$\chi = \mathbf{I} + (\mathbf{1} - \mathbf{I}) \diamond \overline{\tanh} \left(\beta \mathbf{J} + \beta^2 \mathbf{J} \chi \mathbf{J} \right),$$

where $\mathbf{1}$ is a square matrix of all 1's, and $\diamond$ is the element-wise product (Hadamard product).

3.2.1.1.2 A mean field derivation

A mean field solution for the average magnetization is given by

$$m_i = \tanh\left(\sum_k J_{ik} m_k + h_i\right).$$

When there is no external field, and we define the matrix $\mathbf{J}$ to be symmetric, this can also be written as

$$m_i = \tanh\left(\sum_k J_{ik} m_k\right)$$

Now, we can find the average magnetization of spin at i when the spin $s_j = 1$ at node j as

$$m_{i|s_j=1} = \tanh\left(\sum_{k\neq i,j} J_{ik} m_{k|s_j=1} + J_{ij}\right)$$

Let us denote these dependent magnetizations $m_{i|s_j=1}$ by a marix $m_{i|s_j=1} = Q_{ij}$. This leads to

$$Q_{ij} = \tanh\left(\sum_{k\neq i,j} J_{ik} Q_{kj} + J_{ij}\right)$$

which leads to the following recursive formula

$$\mathbf{Q} = \mathbf{I} + (\mathbf{1} - \mathbf{I}) \diamond \overline{\tanh}(\mathbf{JQ})$$

Now, in the absence of an external magnetic field, we expect symmetry, which gives us

$$m_{i|s_j=1} = -m_{i|s_j=-1}$$

The correlation can thus be found as

$$\begin{aligned}
\chi_{ij} = \langle s_i s_j \rangle &= \langle s_i | s_j = 1\rangle \mathrm{P}(s_j = 1) + \langle s_i | s_j = -1\rangle (-1) \mathrm{P}(s_j = 1) \\
&= Q_{ij} \mathrm{P}(s_j = 1) + (-Q_{ij})(-1)\mathrm{P}(s_j = -1) \\
&= Q_{ij} \left[\mathrm{P}(s_j = 1) + \mathrm{P}(s_j = -1)\right] \\
&= Q_{ij}
\end{aligned}$$

But we know that the correlation function is symmetric ($\chi_{ij} = \chi_{ji}$); therefore, we must have $Q_{ij} = Q_{ji}$. We can thus combine the requirement of symmetry and the previous recursive equation to arrive at

$$\chi = \mathbf{I} + (\mathbf{1} - \mathbf{I}) \diamond \overline{\tanh}\left(\frac{1}{2}\mathbf{J}\left(\chi + \chi^T\right)\right)$$

where $\diamond$ denotes the element-wise (Hadamard) product and $\overline{\tanh}$ denotes the element-wise tanh function.

3.2.1.1.3 Using the linear response theorem

The following formulation of the spin correlation has been derived and has proven useful to the neural networks community (Tanaka, 1998; Kappen and Rodriguez, 1998). Consider the Helmholtz free energy function

$$F(J,h) = -\ln Z.$$

The average magnetization at each node is

$$m_u = -\frac{\partial F}{\beta \partial h_u}.$$

The variables m_u and h_u can be replaced by the Legendre transformation to the Gibbs free energy

$$G(J,m) = F(J,h) + \beta \sum_u h_u m_u$$

According to the method of Plefka (1982), one can derive a series of mean field approximations by parametrizing the Hamiltonian as

$$H(\mathbf{s},\gamma) = -\gamma \sum_{u \sim v} J_{uv} s_u s_v - \sum_u h_u s_u$$

where for the true Hamiltonian $\gamma = 1$, and for the zeroth order (independent) approximation $\gamma = 0$.

The Gibbs free energy for this Hamiltonian is

$$\begin{aligned} G(J,m,\gamma) &= F(J,h,\gamma) + \beta \sum_u h_u m_u \\ &= -\ln Z(J,h,\lambda) + \beta \sum_u h_u m_u \\ &= -\ln \left(\mathrm{Tr}_s \exp(\sum_{uv} \beta\gamma s_u J_{uv} s_v + \sum_u \beta h_u s_u) \right) + \beta \sum_u h_u m_u. \end{aligned}$$

The Taylor's series approximation of $G(J,m,\gamma)$ is

$$G(J,m,\gamma) = G(J,m,0) + \gamma \left[\frac{\partial}{\partial \gamma} G(J,m,\gamma) \right]_{\gamma=0} + \gamma^2 \left[\frac{\partial^2}{2! \partial \gamma^2} G(J,m,\gamma) \right]_{\gamma=0} + \gamma^3 \left[\frac{\partial^3}{3! \partial \gamma^3} G(J,m,\gamma) \right]_{\gamma=0} + \ldots \tag{3.5}$$

Since we have introduced $m_u = \langle s_u \rangle$, we can remove h_u by noting that at $\gamma = 0$,

$$m_u = \langle s_u \rangle = \frac{\exp(\beta h_u) - \exp(-\beta h_u)}{\exp(\beta h_u) + \exp(-\beta h_u)}$$

which reduces to

$$h_u = \frac{1}{2\beta} \ln \left(\frac{1 + m_u}{1 - m_u} \right).$$

The successive terms of the Taylor's series, starting with the zeroth order term, are as follows:

$$\begin{aligned}
G(J, m, 0) &= -\ln(\mathrm{Tr}_{\mathbf{s}} \exp(\sum_u -\beta h_u s_u)) + \beta \sum_u h_u m_u \\
&= -\sum_u \ln(\exp(\beta h_u) + \exp(-\beta h_u)) + \beta \sum_u h_u m_u \\
&= -\sum_u \left[\ln \left(\left(\frac{1 - m_u}{1 + m_u} \right)^{\frac{1}{2}} + \left(\frac{1 + m_u}{1 - m_u} \right)^{\frac{1}{2}} \right) + \frac{1}{2} m_u \ln \left(\frac{1 - m_u}{1 + m_u} \right) \right] \\
&= \sum_u \frac{1}{2} (1 + m_u) \ln \left(\frac{1 + m_u}{2} \right) + \frac{1}{2} (1 - m_u) \ln \left(\frac{1 - m_u}{2} \right).
\end{aligned}$$

The first derivative is readily evaluated,

$$\left[\frac{\partial}{\partial \gamma} G(J, m, \gamma) \right]_{\gamma = 0} = -\langle \beta \sum_{uv} s_u J_{uv} s_v \rangle_{\gamma = 0} = -\beta \sum_{uv} m_u J_{uv} m_v.$$

Therefore, up to the first order, the Gibbs free energy is

$$\begin{aligned}
G(J, m)^{\{1\}} &= \left[G(J, m, \gamma)^{\{1\}} \right]_{\gamma = 1} \\
&\simeq \frac{1}{2} (1 + m_u) \ln \left(\frac{1 + m_u}{2} \right) + \frac{1}{2} (1 - m_u) \ln \left(\frac{1 - m_u}{2} \right) - \beta \sum_{uv} m_u J_{uv} m_v.
\end{aligned}$$

To solve for the average magnetization m_u, we re-introduce the field variables $h_u = (\partial / \partial m_u) G(J, m)$, to find

$$h_u = \frac{1}{2\beta} \ln \left(\frac{1 + m_u}{1 - m_u} \right) + \sum_v J_{uv} m_v,$$

which reduces to

$$m_u^{\{1\}} = \tanh \left(\beta \sum_v J_{uv} m_v^{\{1\}} + \beta h_u \right)$$

And this is identical to the naïve mean field equation.

For the second order terms, we find the recursive equation (which is the same as the Bethe mean field approach)

$$m_u^{\{2\}} = \tanh\left(\sum_v \left(\beta J_{uv} m_v^{\{2\}} + \beta^2 J_{uv}^2 m_u^{\{2\}}(1-(m_v^{\{2\}})^2\right) + \beta h_u\right).$$

Plefka (1982) derived the terms up to the second order. Georges and Yedidia (1991) used a similar series, and Nakanishi and Takayama (1997) independently used a computer algebra system to derive terms up to the fourth order.

To calculate the correlations χ_{uv}, consider that $\chi_{uv} = (\partial^2 \ln Z)/(\beta^2 \partial h_u\, \partial h_v) = (\partial m_u)/(\beta \partial h_v)$.

Using the field variables h as the independent variable and the average spins m as the dependent variable, we can calculate $(\partial m_u)/(\partial h_v)$. For the na"ive mean field (the first order approximation, which is the first non-zero term), we find that

$$\begin{aligned}\chi_{uv}^{\{1\}} = \frac{\partial m_u}{\beta \partial h_v} &= \frac{1-m_u^2}{\beta}\left(\sum_w \beta J_{uw} \frac{\partial m_w}{\partial h_v} + \beta \delta_{uv}\right) \\ &= (1-m_u^2)\left(\sum_w J_{uw}(\beta \chi_{wv}^{\{1\}}) + \delta_{uv}\right)\end{aligned}$$

For $m_u = 0$, the correlation function is

$$\chi_{uv}^{\{1\}} = \delta_{uv} + \beta \sum_w J_{uw} \chi_{wv}^{\{1\}}.$$

In matrix notation,

$$\chi^{\{1\}} = \mathbf{I} + \beta \mathbf{J} \chi^{\{1\}},$$

or, solving for the correlation function,

$$\chi^{\{1\}} = (\mathbf{I} - \beta \mathbf{J})^{-1}$$

Similarly, for the second order approximation of the free energy, we have

$$\chi_{uv}^{\{2\}} = \frac{\partial m_u}{\beta \partial h_v} = \frac{1-m_u^2}{\beta}\left(\sum_w \left(\beta J_{uw}(\beta \chi_{wv}^{\{2\}}) + 2\beta^2 J_{uw}^2 m_u^2 (1-m_w)(\beta \chi_{wv}^{\{2\}})\right) + \beta \delta_{uw}\right).$$

Higher order terms can be derived using the free energy approximations (as given in Nakanishi and Takayama (1997))

$$\begin{aligned} G(J,m) =& \frac{1}{2}(1+m_u)\ln\left(\frac{1+m_u}{2}\right) + \frac{1}{2}(1-m_u)\ln\left(\frac{1-m_u}{2}\right) \\ & - \sum_{uv} \beta m_u J_{uv} m_v - \sum_{uv} \beta^2 J_{uv}^2 (1-m_u^2)(1-m_v^2) \\ & - 4\sum_{uv} \beta^3 J_{uv}^3 m_u m_v (1-m_u^2)(1-m_v^2) \\ & - 6\sum_{uvw} \beta^3 J_{uv} J_{vw} J_{uw} (1-m_u^2)(1-m_v^2)(1-m_w)^2. \end{aligned}$$

3.2.1.1.3.1 Valid regime for the linear response correlation

The linear response correlation function is $\chi = (\mathbf{I} - \beta\mathbf{J})^{-1}$. At very high temperatures, β is small; $\mathbf{I} - \beta\mathbf{J}$ is close to the identity matrix and hence invertible and positive definite. As the temperature decreases, β increases, until $\mathbf{I} - \beta\mathbf{J}$ is not invertible. This is the phase transition temperature β_c. Therefore,

$$\begin{aligned} \beta_c &= \inf_{\beta>0} [\det(\mathbf{I} - \beta\mathbf{J}) = 0] \\ &= \inf_{\beta>0} \left[\beta \det(\frac{1}{\beta}\mathbf{I} - \mathbf{J}) = 0\right] \\ &= \inf_{\beta>0} \left[\det(\mathbf{J} - \frac{1}{\beta}\mathbf{I}) = 0\right] \end{aligned}$$

We know that $\det(\mathbf{J} - \lambda\mathbf{I}) = 0$ is the characteristic polynomial, the zeros of which are the eigenvalues of $\mathbf{J}$. Therefore, the largest such zero $\lambda = 1/\beta$ (the largest eignevalue) gives the largest β that is still in the invertible regime. Therefore, $\beta_c = \frac{1}{\lambda_1}$, the inverse of the largest eigenvalue of $\mathbf{J}$.

For β just below the transition temperature β_c, we will have at least one negative eigenvalue for $(\mathbf{I} - \beta\mathbf{J})$ and hence for $\chi = (\mathbf{I} - \beta\mathbf{J})^{-1}$. We know that the correlation matrix must be positive semi-definite; therefore, the linear response approximation of the correlation matrix is only valid for β below the transition point (i.e., at high temperatures).

3.2.2 Graph diffusion

Following the definition in Qi et al. (2008), consider a graph with the adjacency matrix $\mathbf{A}$. We view the links as pipes with flow conductance proportional to the edge weights, and the nodes as reservoirs capable of holding a certain fluid. We also imagine a constant loss of fraction γ from each node and a source flux q_u from outside the system.

The flow of a liquid ρ can now be written as

$$\begin{aligned}\dot{\rho_u}(t) &= \sum_v J_{uv}\rho_v(t) - (\lambda + J_{vu})\,\rho_u(t) + q_u H(t) \\ &= \sum_v J_{uv}\rho_v(t) - (\lambda + D_{uu})\,\rho_u(t) + q_u H(t)\end{aligned}$$

where $H(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 & \text{otherwise} \end{cases}$ is Heaviside step function.

The steady state solution to this is given by

$$\begin{aligned}\rho_{t=\infty} &= (\lambda\mathbf{I} + \mathbf{D} - \mathbf{J})^{-1}\,\mathbf{q} \\ &= \mathbf{G}\mathbf{q}\end{aligned}$$

where $\mathbf{D}$ is the diagonal in-degree matrix such that $D_{uu} = \sum_v J_{vu}$. The graph diffusion kernel is defined as

$$\mathbf{G} = (\lambda\mathbf{I} + \mathbf{D} - \mathbf{J})^{-1}.$$

If we denote $\mathbf{J} - \mathbf{D}$ as $\mathbf{L}$ (the Laplacian matrix), we arrive at $\mathbf{G} = (\lambda\mathbf{I} - \mathbf{L})^{-1}$.

Substituting $\lambda = \frac{1}{\beta}$, we obtain

$$\mathbf{G} = \beta(\mathbf{I} - \beta\mathbf{L})^{-1},$$

which is of same form as the spin-spin correlation, except that we are using the Laplacian matrix rather than the adjacency matrix.

For a large matrix, the inverse can be summed using the series

$$\mathbf{G} = \frac{1}{\lambda}\left(\mathbf{I} - \lambda^{-1}\mathbf{L}\right)^{-1} = \frac{1}{\lambda}\left(\mathbf{I} + \lambda^{-1}\mathbf{L} + \lambda^{-2}\mathbf{L}^2 + \lambda^{-2}\mathbf{L}^3 + \lambda^{-4}\mathbf{L}^4 + \dots\right),$$

with a finite sum leading to a approximation.

3.2.2.1 Exact correspondence for regular graphs

All vertices in a regular graph have the same degree. The graph diffusion kernel and the spin-spin correlation can be shown to have exactly the same form for the case of regular graphs. For a degree d regular graph, we have the diagonal degree matrix $\mathbf{D} = k\mathbf{I}$. This means that the graph diffusion kernel can be written as

$$\mathbf{G} = (\lambda\mathbf{I} + \mathbf{D} - \mathbf{J})^{-1} = (\lambda\mathbf{I} + d\mathbf{I} - \mathbf{J})^{-1} = \frac{1}{d+\lambda}\left(\mathbf{I} - \frac{1}{d+\lambda}\mathbf{J}\right)^{-1} = \beta\left(\mathbf{I} - \beta\mathbf{J}\right)^{-1}$$

where we have defined the inverse temperature $\beta = 1/(d+\lambda)$.

In practice, for most networks we will have an irregular graph, and the edges are often normalized by the degree to discount the effect of high-degree edges.

If we are interested in a symmetric graph diffusion kernel, we will use symmetric normalization, employing both the in-degree and out-degree, to obtain

$$j_{uv} = \frac{J_{uv}}{\sqrt{d_h\, d_v}},$$

where $\mathbf{J}$ is the original adjacency matrix and $\mathbf{j}$ is the normalized matrix to be used to calculate the graph diffusion kernel.

Using this normalization, for a regular graph, the diffusion kernel can be written as

$$\begin{aligned}
\mathbf{G} &= (\lambda\mathbf{I} + \mathbf{I} - \mathbf{j})^{-1} \\
&= (\lambda+1)^{-1}\left(\mathbf{I} - \frac{1}{1+\lambda}\mathbf{j}\right)^{-1} \\
&= (\lambda+1)^{-1}\left(\mathbf{I} - \beta'\mathbf{j}\right)^{-1} \\
&= \left(\lambda\mathbf{I} + \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{J}\mathbf{D}^{-\frac{1}{2}}\right)^{-1} \\
&= \left((\lambda+1)\mathbf{I} - \frac{1}{d}\mathbf{J}\right)^{-1} \\
&= (\lambda+1)^{-1}\left(\mathbf{I} - \frac{1}{d(\lambda+1)}\mathbf{J}\right)^{-1} \\
&= (\lambda+1)^{-1}\left(\mathbf{I} - \beta''\mathbf{J}\right)^{-1}.
\end{aligned}$$

Here we have defined the inverse temperatures $\beta' = 1/(\lambda+1)$, and $\beta'' = 1/(d(\lambda+1))$.

3.2.3 Summary of the functions considered

Table 3.1: Summary of the methods (i.e., graph kernels or distance measures) discussed in this study.

Physical analogy/rationale	Calculation method and name used in this study	Formula	Notes
Graph diffusion of heat/random walkers over the nodes	Can be computed exactly	$\mathbf{G} = (\lambda\mathbf{I} + \mathbf{D} - \mathbf{J})^{-1}$	Valid for all $\lambda > 0$. This method has been previously used in the context of link prediction, candidate gene prioritisation etc.
Spin-spin correlation	Linear approximation	$\chi = (\mathbf{I} - \beta\mathbf{J})^{-1}$	Valid for $\beta < \beta_{\text{Curie}}$. This approximation is widely used in the statistical physics community.
Spin-spin correlation	Saturating approximation (tanh1)	$\chi = \mathbf{I} + (\mathbf{1} - \mathbf{I}) \diamond \overline{\tanh}\left(\frac{1}{2}\mathbf{J}\left(\chi + \chi^T\right)\right)$	Novel method derived using a mean field approach. Should be valid for a greater range of temperature than the linear approximation.

Physical analogy/rationale	Calculation method and name used in this study	Formula	Notes
Spin-spin correlation	Saturating approximation (tanh2)	$\chi = \mathbf{I} + (\mathbf{1} - \mathbf{I}) \diamond \overline{\tanh}\left(\beta\mathbf{J} + \beta^2\mathbf{J}\chi\mathbf{J}\right)$	Novel method derived using a relaxation or variational approach. Should be valid for a greater range of temperature than the linear approximation.

Before evaluating the various functions that we have just derived on some examples of graphs and real applications, we review in Table 3.1 all the measures with their mathematical formulae, their physical analogies and their names that will be used henceforth. The graph diffusion kernel is a commonly used method in complex network analysis for the class of problems we discuss here, while the linear approximation of the Ising model is commonly used in theoretical studies in magnetism. We believe that the other two approximations for the spin-spin correlation in Ising models are novel, both with respect to the physics literature and also for problems in bioinformatics and network analysis.

3.2.4 Computational details

The graph diffusion and the linear spin-spin correlation functions were evaluated using matrix inversion, possible for the smaller data sets considered here. For very large datasets, one could use the recursive relations utilizing sums and repeated matrix multiplications to evaluate these functions, as has been explained in the derivation of these methods, with suitable stopping conditions. The two novel approximations with saturating correlations functions were implemented using the recursive relations. For all experiments, we simply evaluated these functions for 10 recursive steps. Often the longest shortest path in the giant components yields excellent convergence for the number of recursive steps; for the networks considered here, paths of more than 10 hops certainly are capable of spanning the network. This was confirmed by preliminary experiments suggesting that our results do not change even for earlier stopping conditions using 7 and 8 steps. The code for the graph diffusion and all the three Ising spin-spin correlation approximations was written in the R language.

For the exact computation and Gibbs sampling of the spin-spin correlation functions, we wrote custom C++ code that interfaces with R using Rcpp (Eddelbuettel and Fran, 2011). The correlation function was computed exactly by enumerating over all the possible states of the spin variables, and this enumeration was fast and convenient enough for the graph of 20 vertices evaluated on a desktop computer. For Gibbs sampling, we averaged the spin-spin correlation for 10 random restarts with 5000 samples each time, with a sample recorded after every 20 simulation steps. Each simulation step is a complete cycle of proposed switches of each and every spin variable.

3.3 Results

In this section, we first present results for approaches for calculating spin correlation functions and graph diffusion kernels for sample graphs where the truth is known and the graphs are sufficiently small to permit exact calculations by full summations over states. Secondly, we apply these methods to the practical problem of predicting missing links in a protein-protein interaction network.

3.3.1 The single clique, or long-range, weak interactions model

Let us consider a very simple case of N vertices with complete connections, with an edge between every distinct pair of vertices. In statistical physics, this is referred to as the long-range, weak interactions model as opposed to a regular grid model. It is long-range because there are interactions between all vertices, even if they are separated by large distances in a physical model. To bound the energy in the limit of infinite size, the average interaction is scaled as $J_{ij} = 1/N$, leading to weak interactions. For simplicity, we will keep a uniform $A_{ij} = 1$ for all non-diagonal entries in our adjacency matrix, scaled as stated by system size to $1/N$.

3.3.1.1 The mean field Ising model

A clique of size N leads to $N-1$ interactions for each vertex. Since the problem is invariant over the vertices, $m_i = m \forall i$. For $J_{ij} = 1$,

$$m = m_i = \tanh(\beta \sum_j J_{ij} m_j) = \tanh(\beta(N-1)m)$$

This can also be considered as a discrete time dynamical system (map or recurrence relation), with the right hand side giving the next value of m given the present value.

In the high temperature regime, for $(N-1)\beta < 1$, there is only one solution, $m = 0$. For temperatures colder than the Curie point, where $\beta > 1/(N-1)$, there is a phase transition, and two additional symmetric solutions appear. We can evaluate the stability of the three solutions using the second derivative of the free energy. The $m = 0$ solution is unstable, while the positive and negative magnetization solutions are stable.

3.3.1.2 The spin correlation calculated using linear response

The first order spin-spin correlation $\chi^{\{\mathbf{1}\}} = (\mathbf{I} - \beta\mathbf{J})^{-1}$. For convenience, consider the recurrent form

$$\chi_{uv} = \delta_{uv} + \beta \sum_{w} J_{uw}\chi_{wv}.$$

For the clique model, $\chi_{uv} = \chi \forall u \neq v$ and $\chi_{uu} = 1$. This reduces to

$$\chi = \beta(N-1)$$

While this does not show any phase transition, it is clear that the spin-spin correlation is greater at high β (low temperature) and for larger cliques (high N), assuming the individual edge weights (J_{ij}) remain constant.

3.3.2 Approximations to the correlation function evaluated for a small graph

In this section, we construct a small graph to explore the behaviour of the linear response and our tanh approximations for the correlation function. We explore the factors affecting the accuracy of the various approximations. The small graph permits exact calculations of the correlation function by a complete sum over states, with connectivity chosen to be representative of features of sub-regions of real graphs. The adjacency matrix for this graph is given in Figure 3-1.

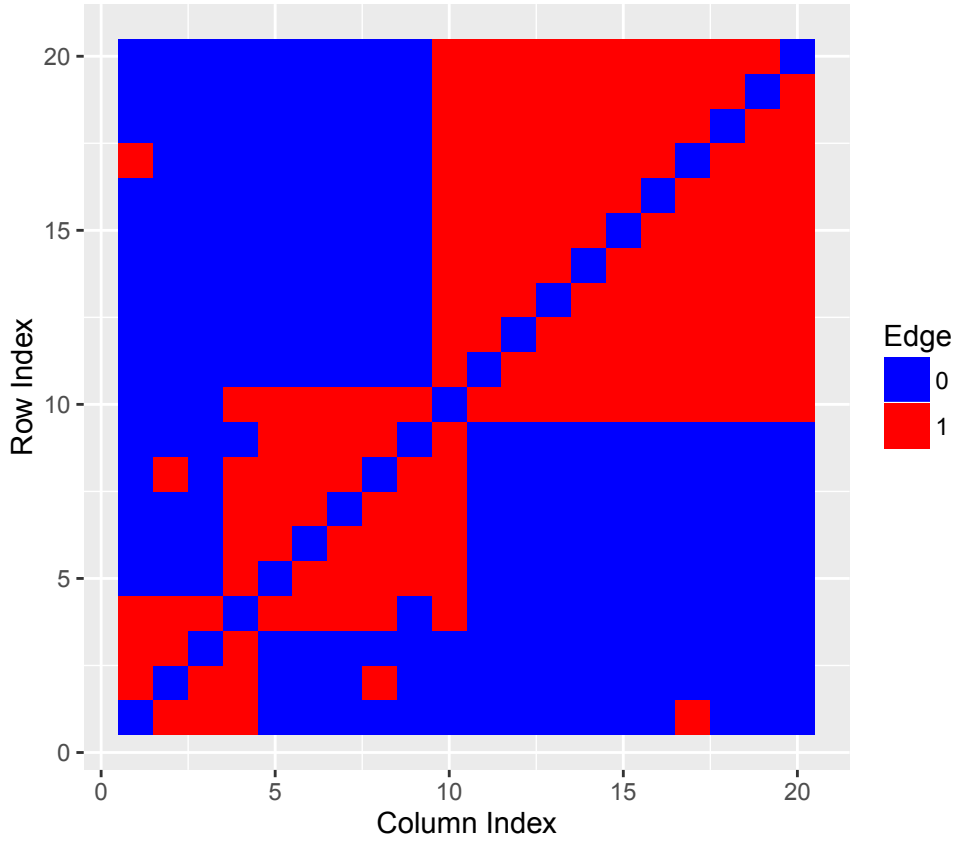

Figure 3-1: Adjacency matrix for a small artificially constructed graph used to evaluate the approximations to the correlation function.

Figure 3-2 visually displays the spin-spin correlation matrices, calculated exactly by enumerating over all states, by Gibbs sampling, and with the three approximations discussed in the Methods section. The value of β increases from left to right, corresponding to lower temperatures and prominent and larger clusters of nodes coalescing together. For the case of the Ising magnetic models, these would be the growing microdomains of magnetization. The various approximations to the spin-spin correlation matrix appear to behave similarly here. To examine numerical differences that might

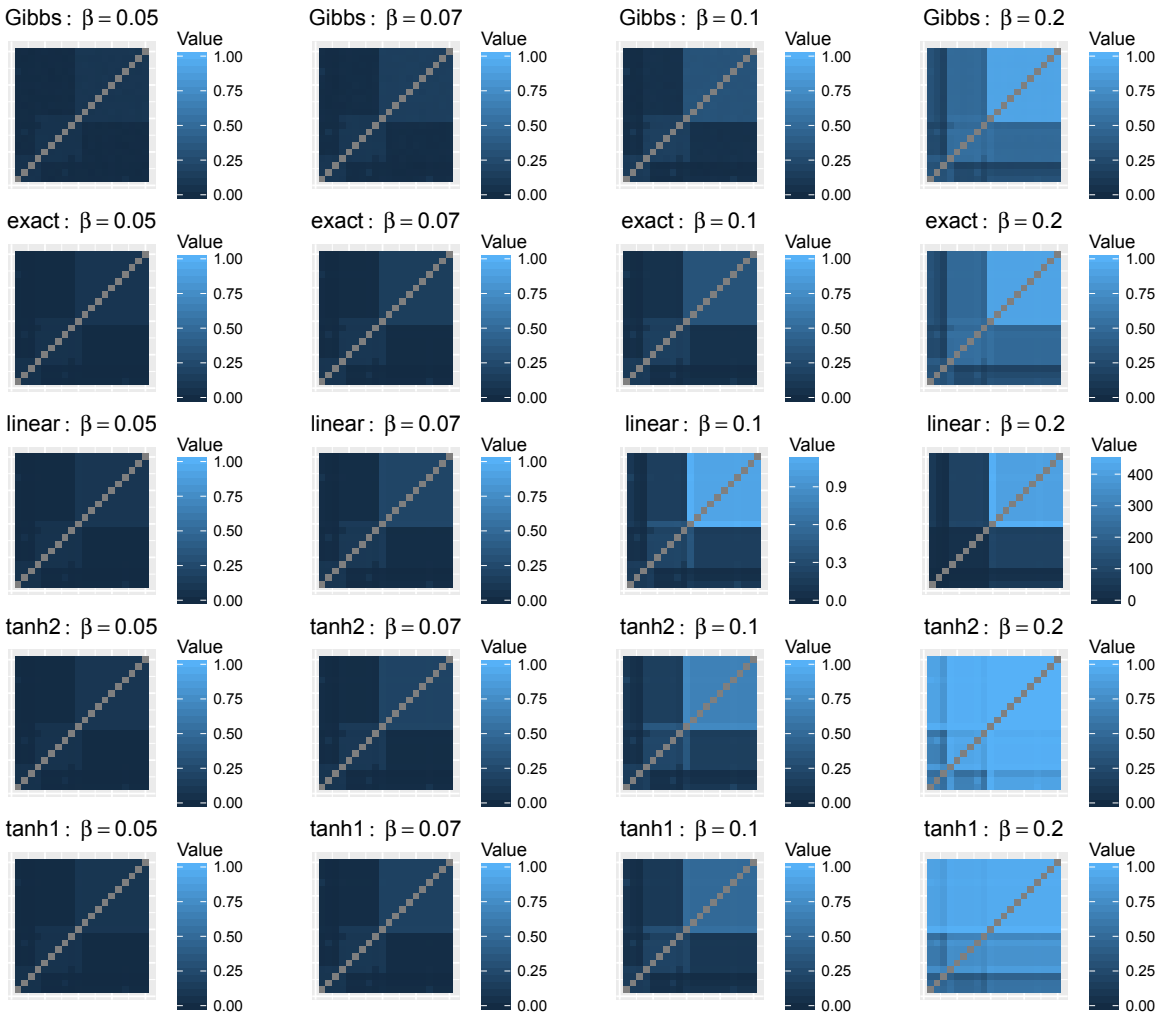

Figure 3-2: The behaviour of the different approximations to the correlation function along with the exact calculation for the small graph shown in Figure 3-1.

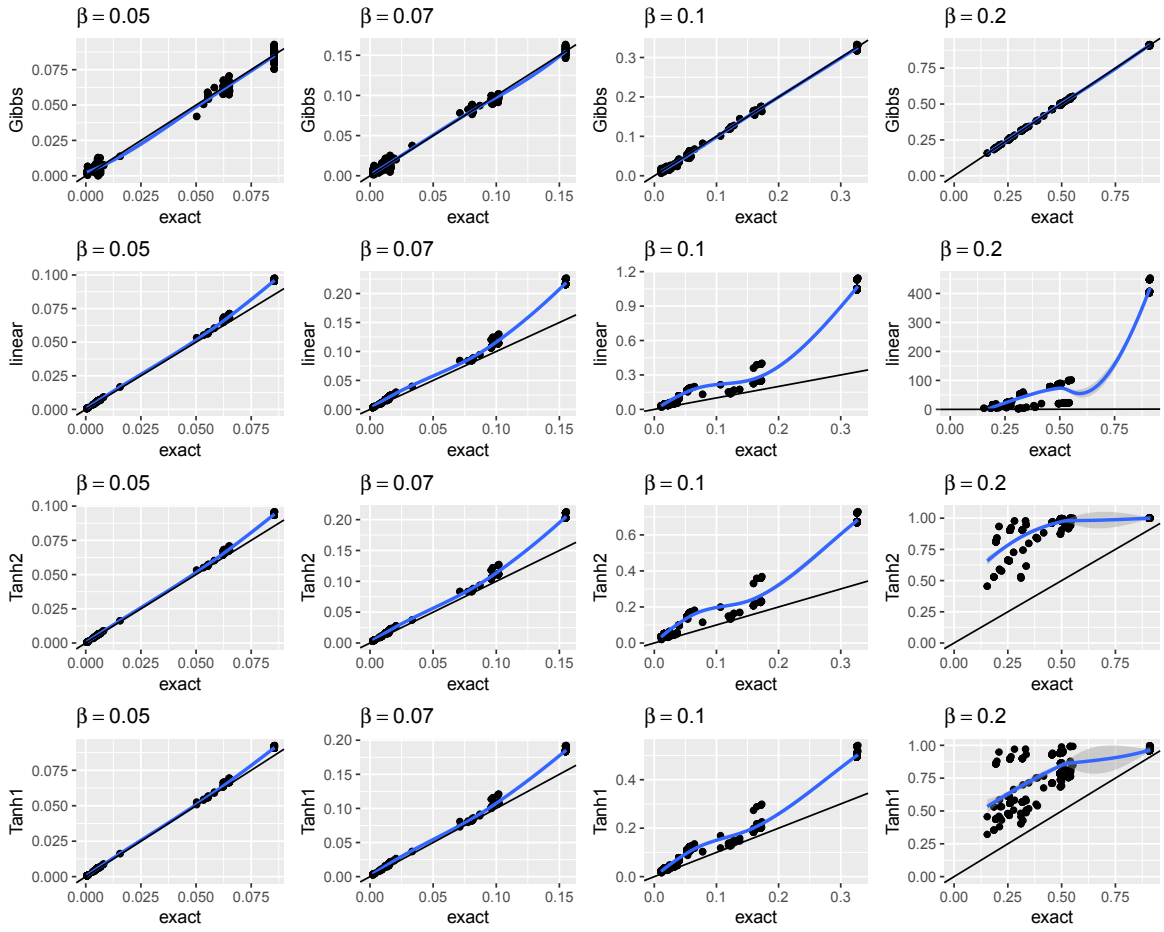

Figure 3-3: Accuracy of different estimates of the spin-spin correlation for the small graph shown in Figure 3-1. Each panel had the exactly computed correlation (by summing over all states) on the x-axis and the estimated value on the y-axis. Each column has a different temperature expressed as inverse temperature β. Dots represents pairs of nodes for which the spin-spin correlation is calculated. The black line is the $x = y$ curve that is expected if the estimate of the correlation is equal to that computed exactly. The blue line denotes the best Lowess fit for the exact vs. estimate.

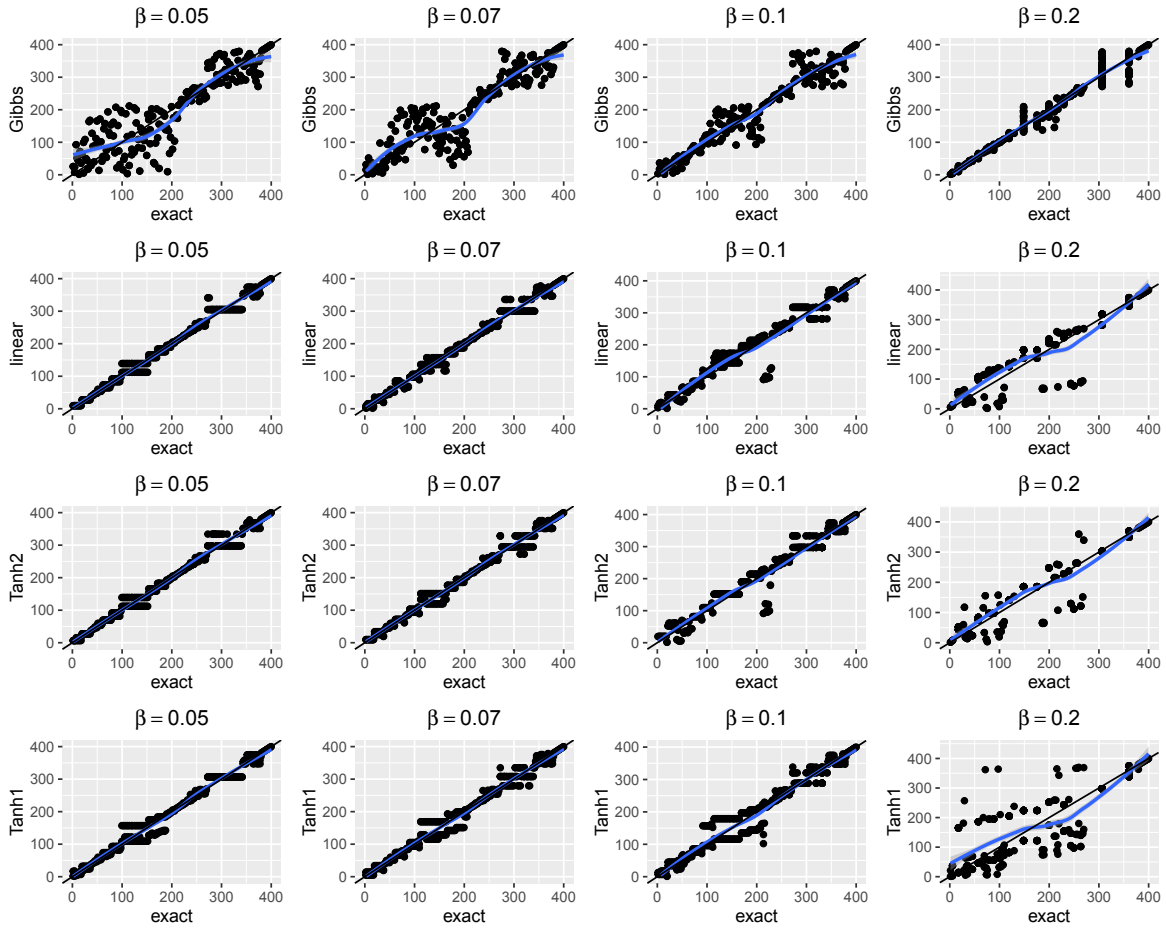

Figure 3-4: Comparison of node-pairs ranked from low to high spin-spin correlations using different estimates for the correlation for the small graph shown in Figure 3-1. The x-axis denotes the ranking by computing the correlation exactly while the y-axis denotes ranking using an estimate. Each column is for a different temperature expressed as inverse temperature β. Dots represents pairs of nodes for which the spin-spin correlation is calculated. The black line is the $x = y$ curve that is expected if the rank of the estimate of the correlation is equal to the rank from the exactly computed correlation. The blue line denotes the best Loess fit for the exact vs. estimated.

have been overlooked in the matrix images, Figure 3-3 plots the estimated values of the correlation function for each pair of nodes versus the exactly computed correlation. It is clear that the deviation between the exact and approximate correlations increases at lower temperatures (higher β). Also, it is clear that the linear response correlation approximation can give values greater than 1 even before its collapse due to the phase transition. At even lower temperatures, we observe more non-physical values of correlation from the linear response correlation function. Our two saturating correlation functions, while still suffering from increasing inaccuracy at lower temperatures, have greater validity in bounding the magnitude of the correlation function to values less than 1. Our approximations also show the generally correct trend of the exact correlation function.

For applications including prioritizing candidate genes to phenotypes or pathways and predicting missing links, we are often more interested in the relative order of the similarity/distances between pairs of nodes rather than the magnitude of the distance measure. Figure 3-4 compares the ranking of pairs of nodes in the graph using the various spin-spin correlation estimates with rankings from the exact correlation. For this example for the values of β considered, the general ranking of node pairs is maintained by the approximate correlations, even if the correlation functions themselve differ from the true values. The Gibbs sampling estimates for high temperatures (the panel on upper left hand corner of the image) shows higher variance than our analytic approximation. This is natural because while Gibbs sampling estimates are unbiased, they do have the natural thermal variance present at that temperature and can be slow to converge.

3.3.3 Resolution and behaviour on a string of cliques

In this section, we consider a slightly larger sample graph with block structure, with connections enriched within blocks and much lower frequency between blocks. We then visualize the graph diffusion kernel and the correlation approximations. The graph includes 5 cliques of various sizes arranged linearly. Consecutive cliques share a vertex, yielding a string of cliques. In addition, some noise is added by randomly removing some edges from within cliques and randomly adding edges between vertices in different cliques. The graph is shown in Figure 3-5, and the corresponding adjacency matrix is visualized in Figure 3-6.

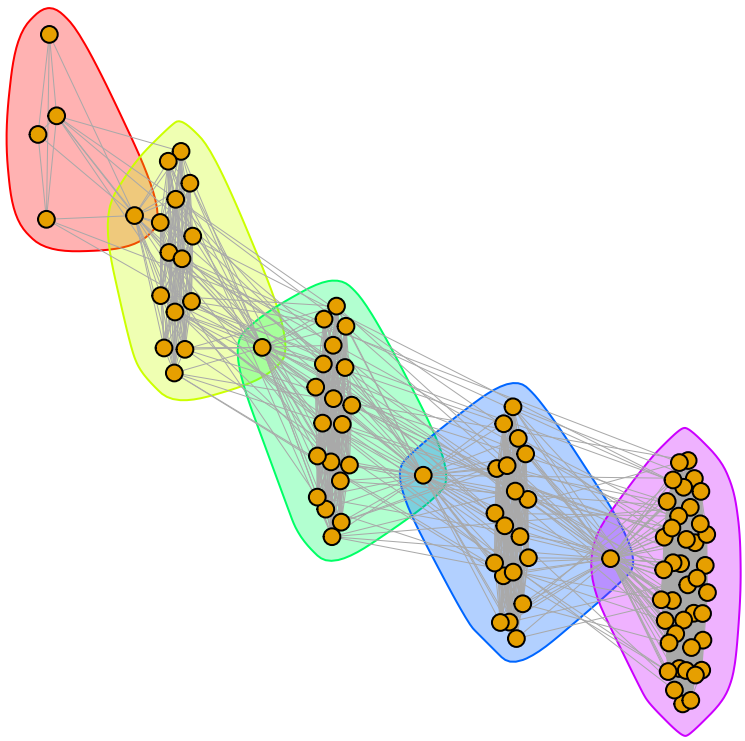

Figure 3-5: Graph of a string of cliques that is used to illustrate the behaviour of the algorithms. The colored regions denote the different cliques used to construct the graph. Subsequent colored regions share a vertex, which is a member of two cliques. A small number of inter-clique edges randomly are added and a small number of intra-clique edges are removed.

Since this a small graph, we can infer some conclusions about the behaviour of the various functions on this graph for different temperatures by visualising the diffusion kernel or the spin-spin correlation. Figure 3-7 shows the behaviour of the various methods as matrix images. The size of this graph makes it prohibitive to evaluate the exact correlation function by a complete sum over each possible state. However, as we have seen earlier, the Gibbs sampling estimate converges sufficiently quickly to represent the real correlation matrix.

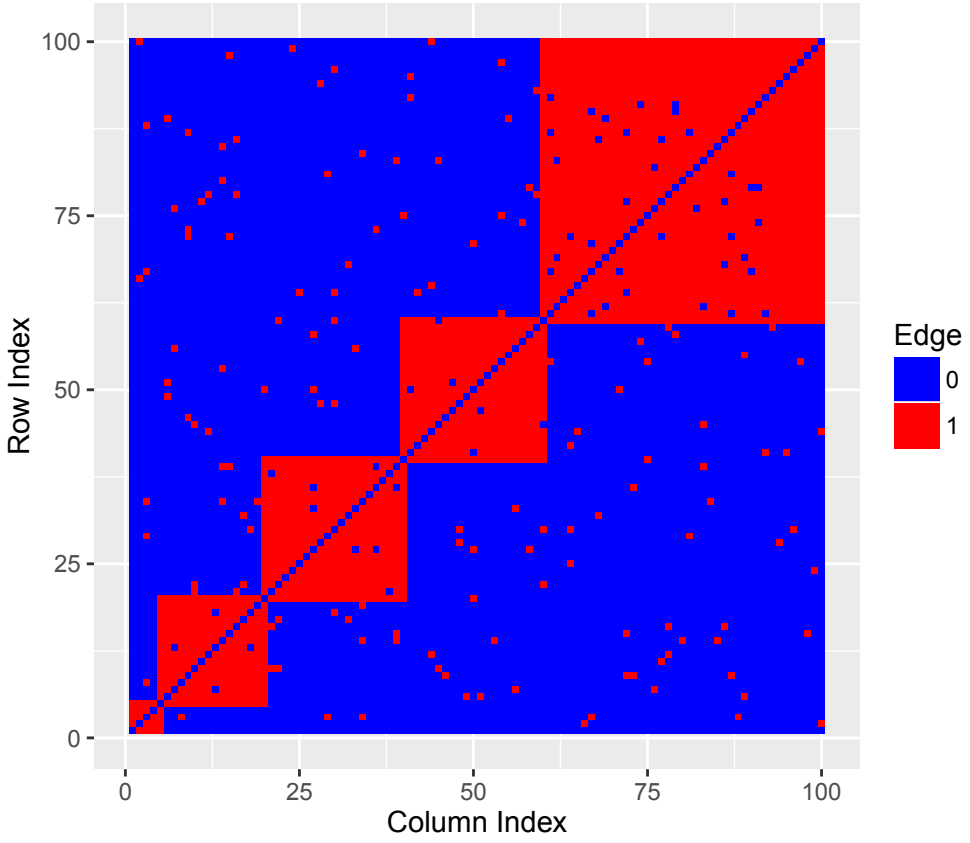

Figure 3-6: Adjacency matrix of the graph shown in Figure 3-5. The graph consists of a number of cliques of different sizes, with adjacent cliques sharing a vertex, with a small amount of noise added by removing some edges within cliques and adding extraneous edges between cliques.

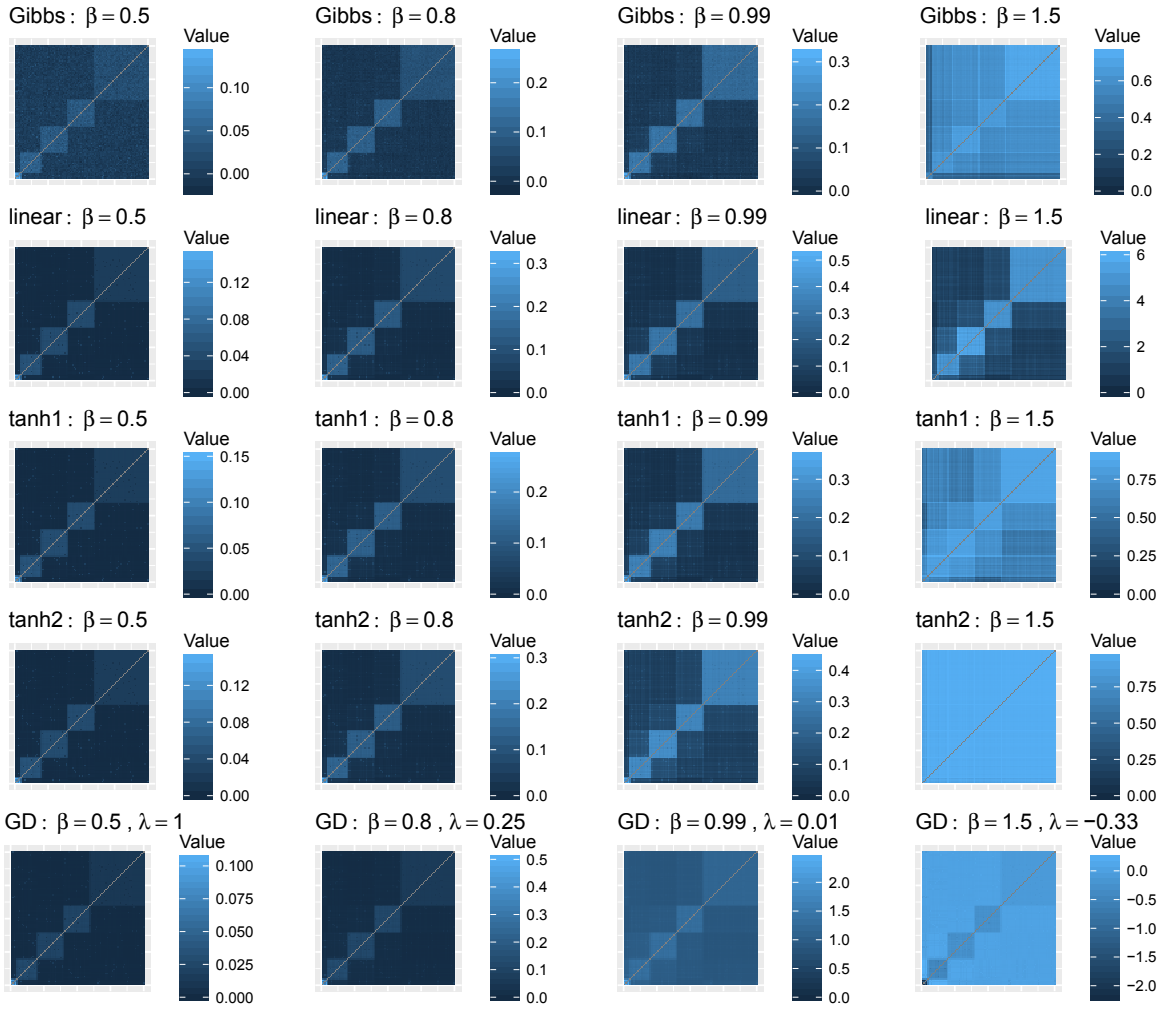

Figure 3-7: Analytic and Gibbs simulation of the Ising model and the corresponding Graph diffusion kernel. The linear and saturating tanh approximations along with the Gibbs simulation of the spin-spin correlation of the Ising model, and the Graph diffusion kernel computed with $\lambda = \frac{1}{\beta} - 1$ for the symmetric degree normalization ($J_{uv} = \frac{A_{uv}}{\sqrt{d_u}\sqrt{d_v}}$) toy graph shown in Figure 3-5. The calculations are performed for various values of β as shown in the subfigure titles.

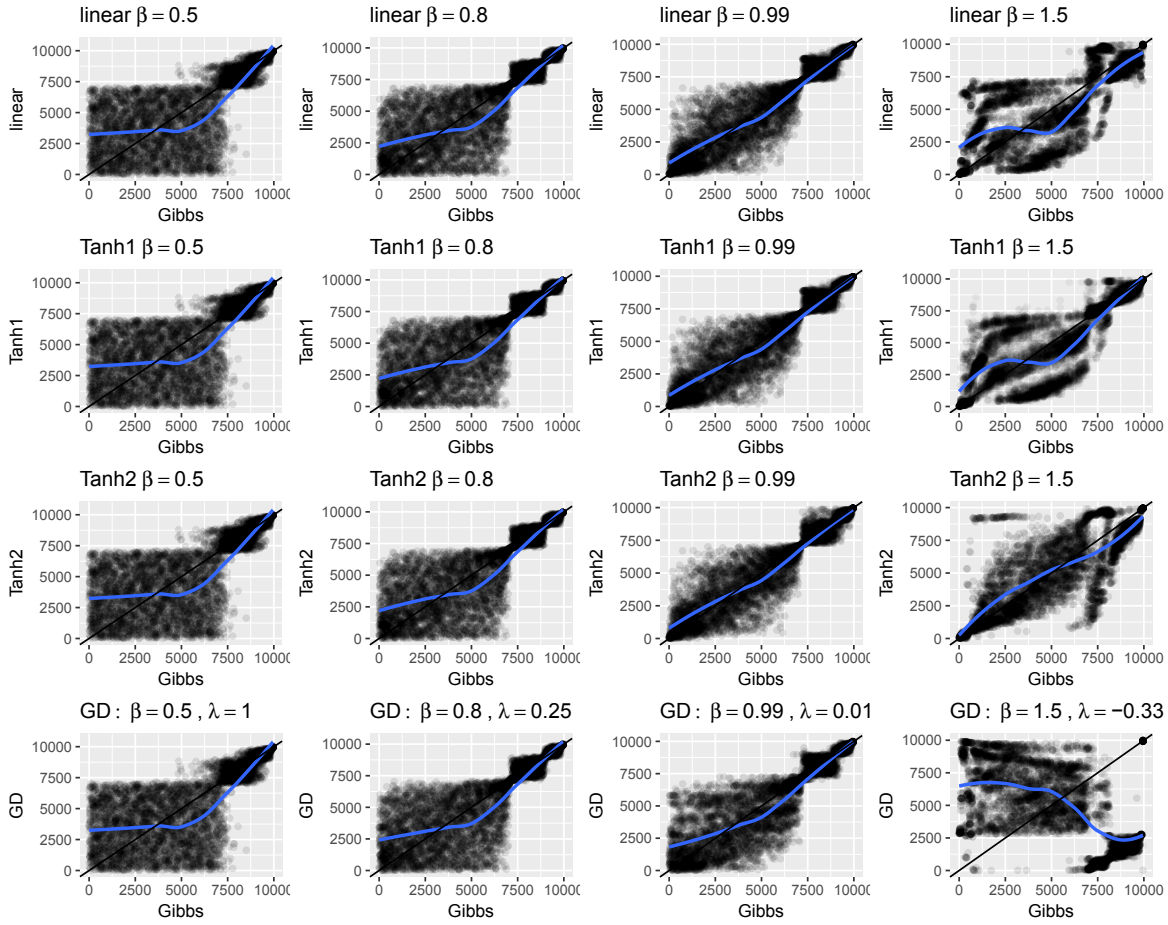

Figure 3-8: Comparison of the ranks of spin-spin correlation between node pairs using different estimates of the correlation for the string of clusters shown in Figures 3-5 and 3-6. The node pairs are ranked from low correlation to high correlation. The x-axis denotes the ranking using Gibbs simulation of the correlation, and the y-axis denotes ranking using a closed form estimate. Each column has a different temperature, expressed as the inverse temperature β. Dots represents pairs of nodes for which the spin-spin correlation is calculated. The black line is the $x = y$ curve that is expected if the rank of the estimate of the correlation is equal to the rank from the exactly computed correlation. The blue line denotes the best Loess fit for the Gibbs simulation vs. the closed form estimate.

In the high temperature regime (low β), the cliques are clearly visible separately, and the direct links between vertices are the most influential in determining the kernel. As the temperature decreases, we observe that the correlation function or diffusion kernel for nodes in adjacent cliques becomes higher as the algorithm groups larger numbers of nodes together. In terms of the Ising model, these represent growing microdomains of superparamagnetism. At lower temperatures, the magnitude of the correlation function also increases. Note that symmetric degree normalization gives symmetric diffusion kernels and spin correlation functions and similar behaviour for small degree and large degree nodes. Since this a small graph, the linear spin-spin correlation function and the graph diffusion kernel are calculated using direct matrix inversion rather than repeated multiplications. That is, however, a detail of implementation and not necessarily enlightening for the basic methods. This makes it possible to calculate the function even when the expansion is invalid at low temperatures. To make the comparison easier, we denote the different graph diffusion kernels with different values of β just as we have for the spin-spin correlations using the formula $\beta = 1/(1 + \lambda)$ as explained in Section 3.2.2.1.

At the highest β, the corresponding value of λ is negative and the the graph diffusion kernel is invalid, with the values of the kernel within cliques lower than those between cliques. Obviously, negative λs violates the physical model for graph diffusion. Even though the solution is mathematically invalid, we have presented it to understand the numerical behavior of the expression viewed independently as a distance measure over vertices in a graph. We also note that neighboring cliques start merging together at lower temperatures, with slightly different temperature for various methods.

As for the previous example, we plot the ranks of each pair of nodes calculated using the different methods in Figure 3-8. As mentioned previously, relative ranks of node pairs may be more important for prioritization problems (as opposed to physics problems where we are interested in bulk properties and macrostates rather than individual spins or particles). We compare each method to the ranks obtained using a Gibbs sampling of the spin-spin correlation. We generally observe that all methods give similar rankings in their regions of validity, such as when the graph diffusion kernel has a negative value of λ or the temperature is below the phase transition temperature for the Ising model. The spread in the individual plots in Figure 3-8 is partly due to the thermal noise present in the Gibbs estimate. As a visual aid, a Loess smoothing curve over the scatter points is drawn as a blue line.

While these numerical experiments give us an overview of the behaviour of these methods, it is

important to note that we are plotting the spin-spin correlations and the graph diffusion kernel over links that are actually present in the network. For predicting missing links, we are only interested in pairs with no direct link present in the graph over which the correlation is calculated. In the following sections we compare these methods for predicting missing links to explore performance on realistic predictive applications.

3.3.4 Predicting missing links in the yeast protein interaction network

In this section, we evaluate how well our methods perform on a link prediction task. We select *Saccharomyces Cerevisiae* PPI network because the awesome power of yeast genetics (Stark, 2001) has made it an extraordinarily well-characterized eukaryotic cellular system. While it is a simpler organism, evolutionary conservation between yeast and human has made it an important model for mammalian disease and development. Yeast is also a very useful industrial microbe and the subject of many synthetic biology efforts. We extracted the protein-protein interaction network of *Saccharomyces Cerevisiae* (common yeast) from Biogrid (Stark et al., 2006). After recursively trimming all the nodes of degree one, we obtained a graph of 5630 nodes and 83137 edges. We then use the graph diffusion kernel and our approximation of the spin-spin correlation to predict missing links. For each cross-validation, we hold back randomly selected 10% of the edges as the positive test set, and an equal number of pairs without edges as the negative test set.

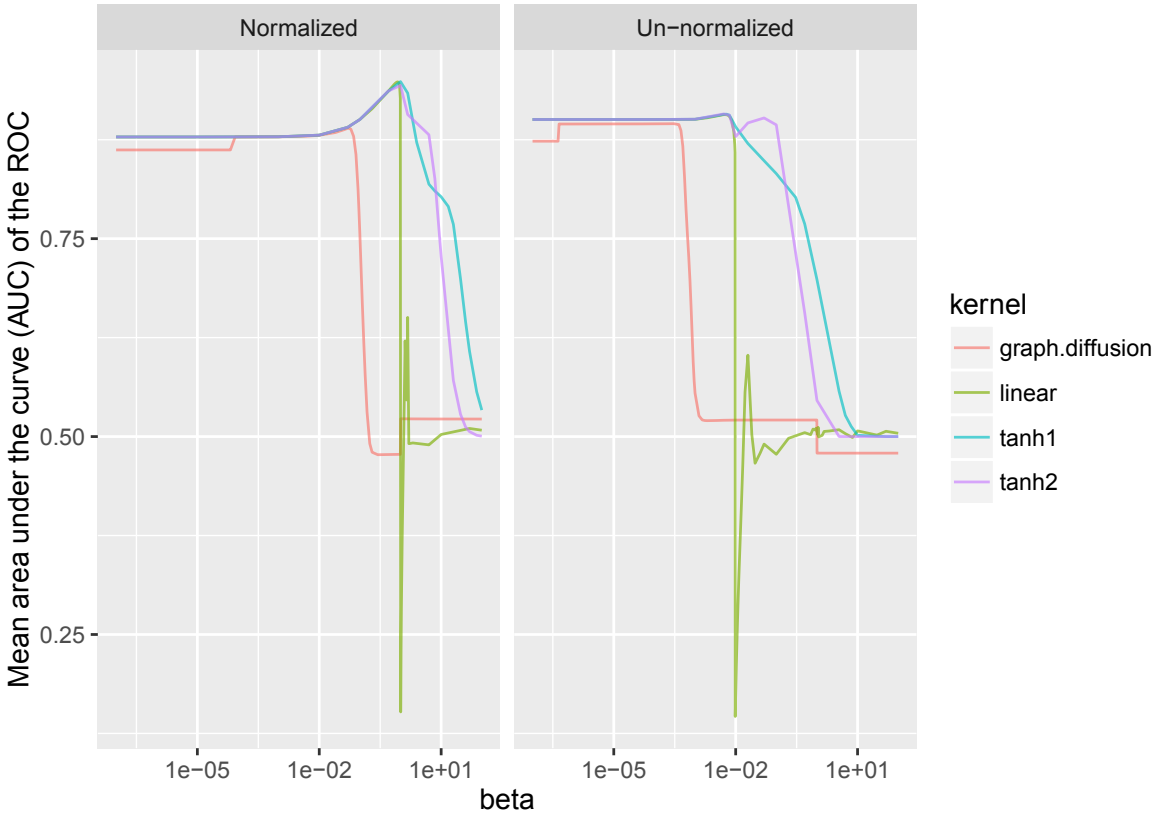

Figure 3-9: Comparison of the different kernels for predicting missing links in the Yeast PPI. Each curve is the mean of the area under of the curve for the ROC curves of 5 different cross-validation tests. The results computed using the symmetric degree normalized and unnormalized adjacency matrix are plotted in different panels.

We show the performance, as measured by the area under the curve of ROC curve, for all the

methods together in Figure 3-9. We see that the Ising correlation methods outperform the graph diffusion kernel on the link prediction task. Secondly, we observe that the best AUC is reached at about the same temperature for all the three approximations of the spin-spin correlation function, and this temperature is close to or right before the temperature of the phase transition. Thirdly, our saturating approximations (denoted tanh 1 and tanh 2) do not break down as catastrophically at the phase transition. We can argue that our approximations are thus more robust for these tasks and not as sensitive to temperature as the linear response method. Fourthly, we see that the degree normalized adjacency matrix gives a slightly higher AUC.

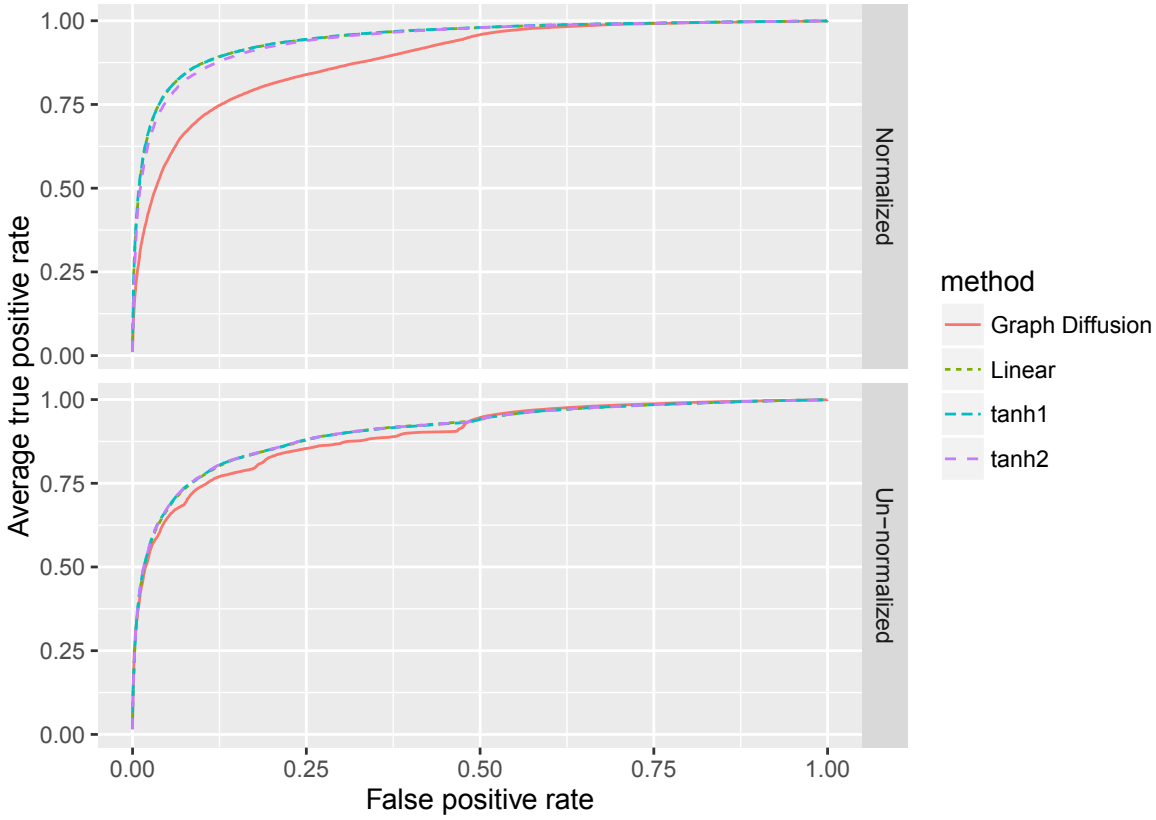

Figure 3-10: The ROC curves for predicting missing links in the Yeast PPI network at the individual best β for that specific kernel. The same β is used for all cross-validation sets. (Top) where the kernels are evaluated by symmetric normalization of the adjacency matrix, and (Bottom) where the kernel is evaluated on the un-normalized or raw adjacency matrix.

Figure 3-10 shows ROC curves at the respective optimal temperatures for each method. The three of the Ising spin-spin correlation methods perform very similarly but better than the graph diffusion kernel. This effect is more pronounced when we use the symmetrically degree normalized adjacency matrix to compute correlation and diffusion kernels.

3.3.5 Predicting missing links in the *Plasmodium falciparum* protein interaction network

We confirm the applicability of our methods on a similar network here. *Plasmodium falciparum* is the parasite that causes malaria, a major public health concern in the developing world. Since it is a smaller (in terms of the number of genes in the graph) organism than yeast, the computations are fast and it is a good way to validate the generality of our findings. We also show that small microbial PPI networks may have enough data to warrant the application of these generic systems level methods. Hypothesis generation using these network analysis approaches can aid in directing experimental attention to genes that may aid in discovery.

The method followed here mirrors that used for the yeast dataset. We extracted the protein-protein interaction network of *Plasmodium falciparum* from Biogrid. After recursively trimming all the nodes of degree one, we obtained a graph of 686 nodes and 1930 edges. We then use the graph diffusion kernel and the linear and saturating (tanh) versions of the spin-spin correlation to predict missing links. For each cross-validation, we hold back randomly selected 10% of the edges as the positive test set, and an equal number of pairs without edges as the negative test set. The performance of the graph diffusion kernel and our approximations of the spin-spin correlation is evaluated in terms of the AUC and plotted against different values of β in for 5 cross-validation tests.

Figures 3-11 and 3-12 also plot the smallest eigenvalue of $\mathbf{I} - \beta\mathbf{J}$ in addition to the AUC for different values of β. We see that the sharp fall in the AUC happens right at the point of the phase transition when the linear response breaks down, resulting in an impossible negative definite correlation matrix. The peak in the AUC also happens for β right below this phase transition.

Figure 3-13 shows similar behavior to that shown in Figure 3-9 for the yeast dataset. Here, however, the saturating approximations ($\tanh 1$ and $\tanh 2$) are not just more robust but actually outperform the linear response correlation function even at the optimal temperature.

Figures 3-14 and 3-15 shows the ROC and precision-recall curves respectively at the optimal temperature for each method. We can clearly see that the $\tanh 1$ and $\tanh 2$ approximations outperform the linear response approximation, which in turn outperforms the graph diffusion kernel. This order is most pronounced true on the lower left hand corner of the curve, which shows the most probable predictions of the missing links and is the most useful for generating candidate genes for experiments.

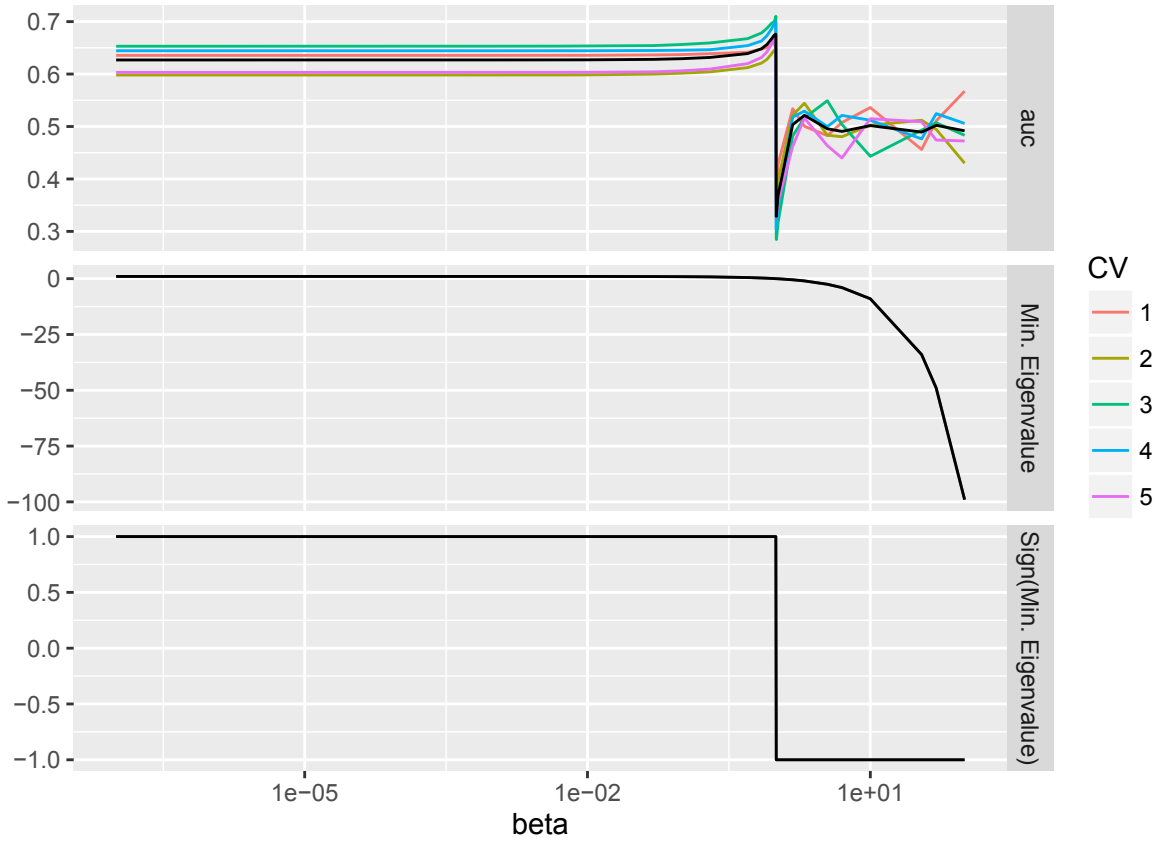

Figure 3-11: The performance of the linear estimate of the spin-spin correlation function on the normalized *Plasmodium falciparum* PPI network as measured by the area under the receiver operating characteristics curve. A fraction of the held-out edges are used as the positive test set, while an equal number of vertex pairs without an edge form the negative test set. The individual lines denote the different cross-validation sets while the black line represents the mean of all the cross-validation sets. The minimum eigenvalue of the matrix that is inverted ($\mathbf{I} - \beta\mathbf{J}$) and its sign is also plotted to show the transition temperature where the linear spin-spin correlation estimate becomes invalid. This is the point where the prediction performance (AUC) drops as well.

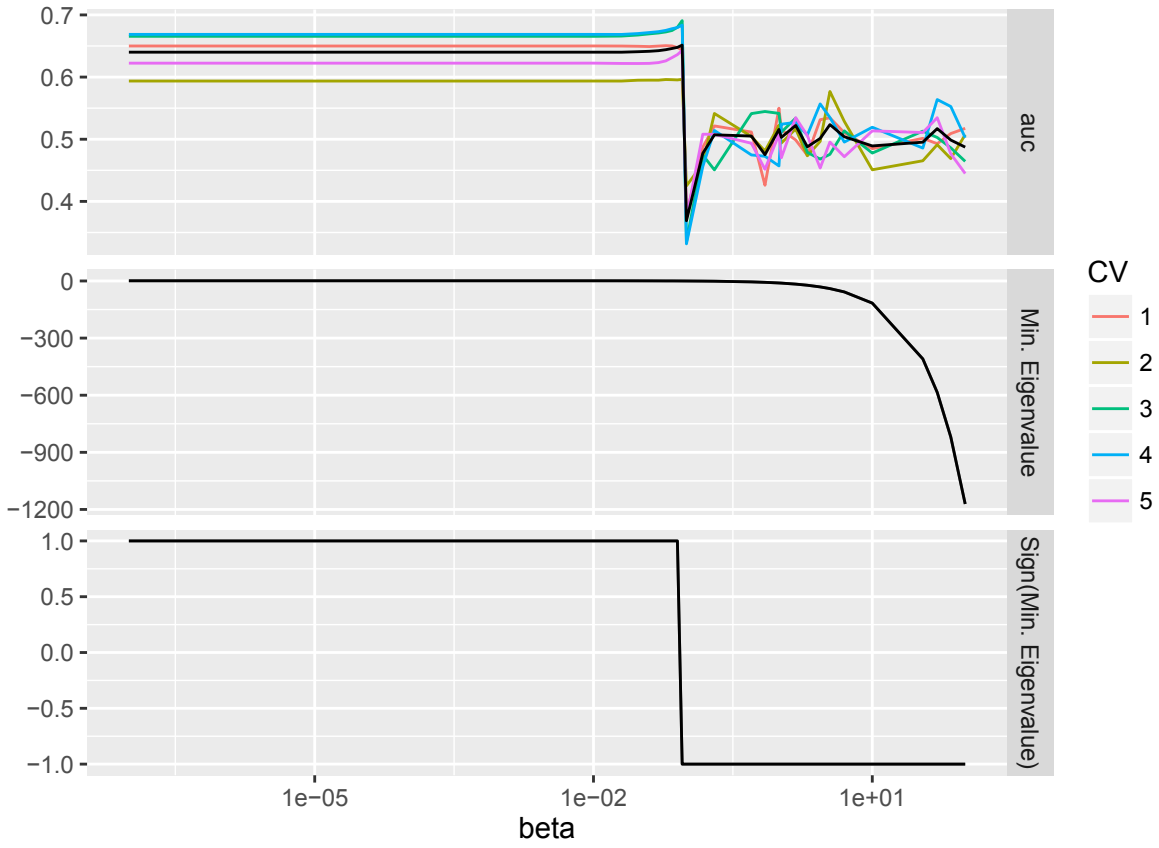

Figure 3-12: The performance of the linear estimate of the spin-spin correlation function on the unnormalized *Plasmodium falciparum* PPI network as measured by the area under the receiver operating characteristics curve. A fraction of the held-out edges are used as the positive test set, while an equal number of vertex pairs without an edge form the negative test set. The individual lines denote the different cross-validation sets while the black line represents the mean of all the cross-validation sets. The minimum eigenvalue of the matrix that is inverted ($\mathbf{I} - \beta\mathbf{J}$) and its sign is also plotted to show the transition temperature where the linear spin-spin correlation estimate becomes invalid. This is the point where the prediction performance (AUC) drops as well.

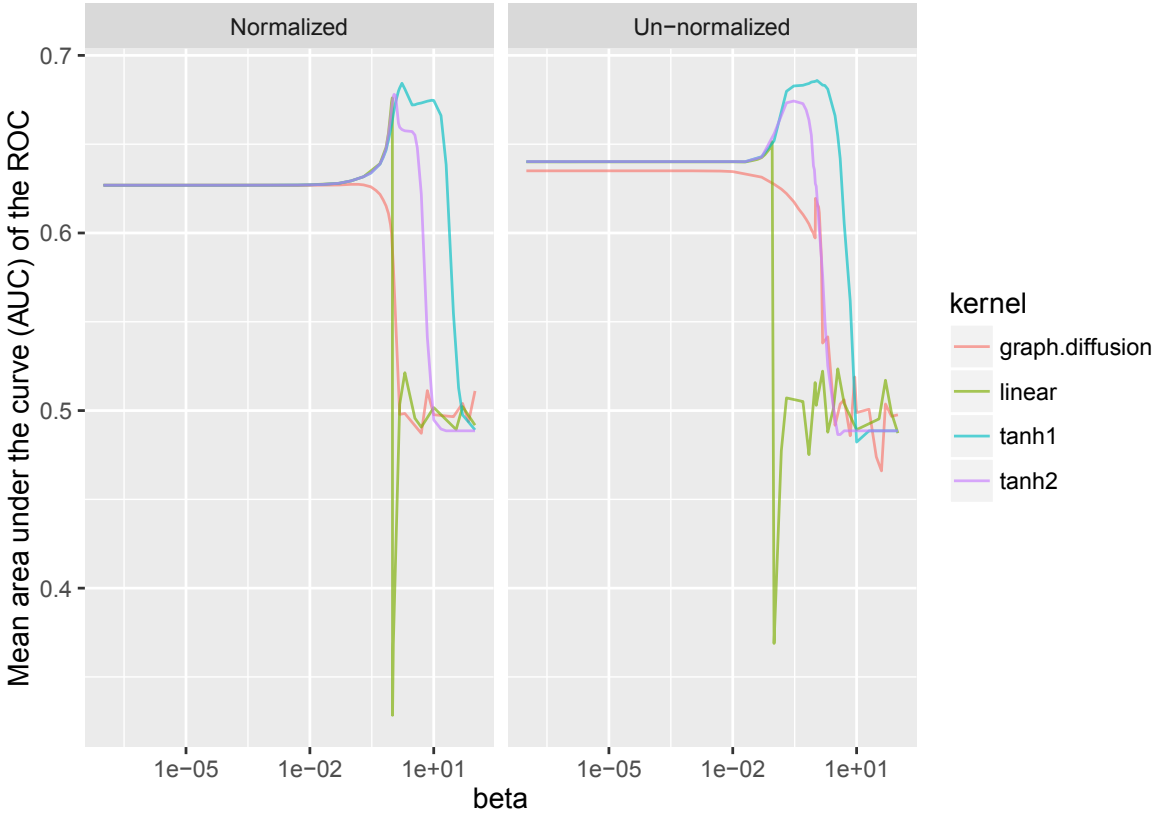

Figure 3-13: Comparison of the performance of all the different kernels for predicting missing links in the *Plasmodium falciparum* network. Each curve is the mean of 5 cross-validation tests. The performance is compared in the two panels for the kernels computed from the normalized and un-normalized adjacency matrices.

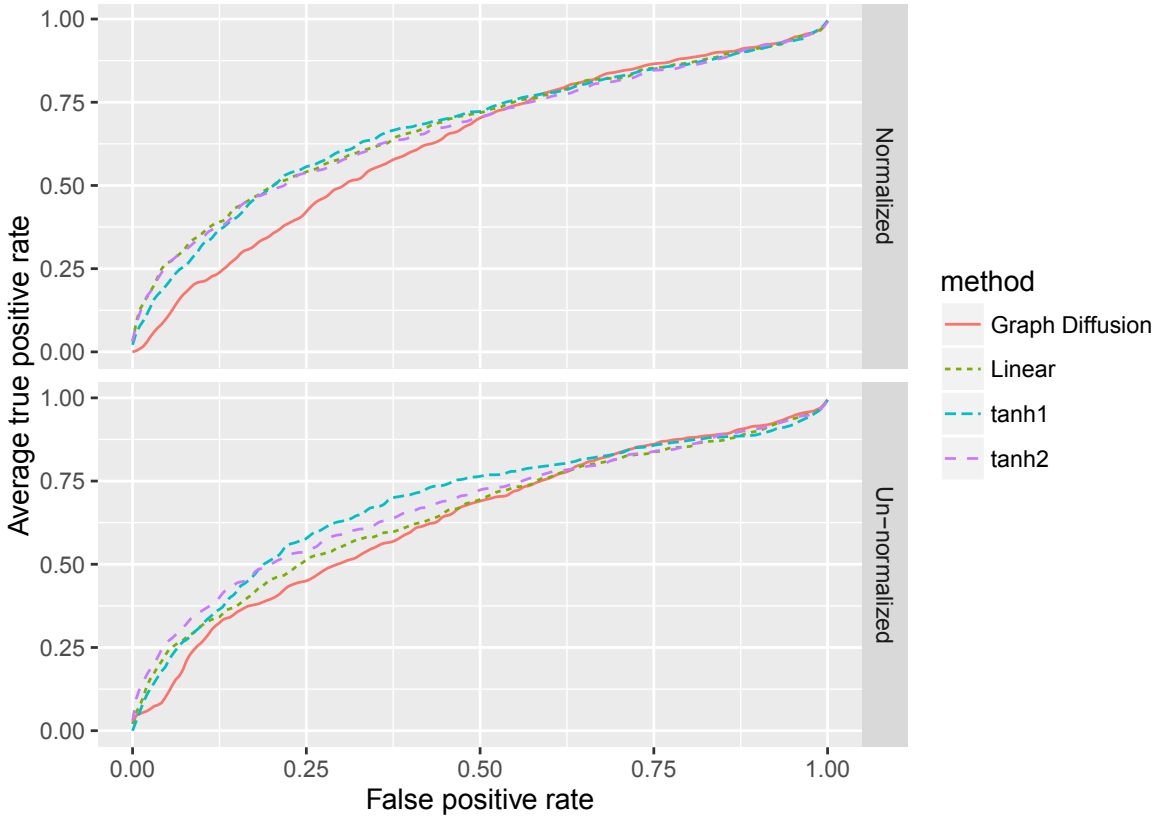

Figure 3-14: The ROC curves for predicting missing links in the *Plasmodium falciparum* PPI network at the individual best β for that specific kernel. The same β is used for all cross-validation sets. (Top) where the kernels are evaluated by symmetric normalization of the adjacency matrix, and (Bottom) where the kernel is evaluated on the un-normalized or raw adjacency matrix. The much smaller size of the *Plasmodium falciparum* PPI network leads to only about 200 edges in the test set, giving a noisier ROC curve compared to the yeast PPI network.

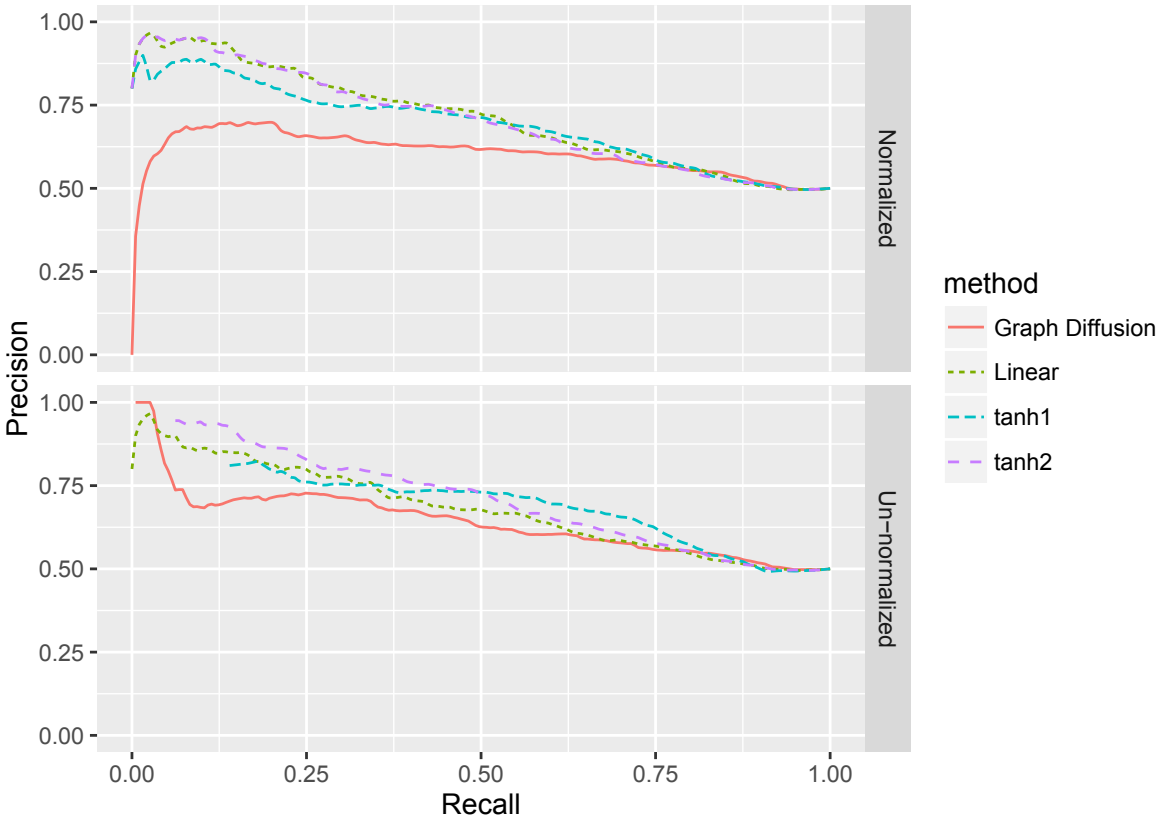

Figure 3-15: The precision-recall curves for predicting missing links in the *Plasmodium falciparum* PPI network at the individual best β for that specific kernel. The same β is used for all cross-validation sets. (Top) where the kernels are evaluated by symmetric normalization of the adjacency matrix, and (Bottom) where the kernel is evaluated on the un-normalized or raw adjacency matrix.

3.4 Discussion

We have shown how the spin-spin correlations from the Ising model and graph diffusion kernels from random walk models lead to identical expressions for regular graphs, and similar expressions for general networks. Both of these models can provide effective distance measure over nodes in a graph that can be used for link prediction. There are a number of other methods commonly used in the analysis of complex networks that are also related to these models. Specifically, the notion of random walks over a graph is employed in Markov clustering, and for a distance measure to a set of nodes for gene prioritization (Köhler et al., 2008). Also, the Ising model and its variants have been used for clustering and modeling propagation across networks among other problems. We discuss how the results from these disparate methods also have the same form and can be related to one another in the same way that spin-spin correlation and the graph diffusion are related.

3.4.1 Markov clustering

The Markov clustering algorithm (Van Dongen, 2008) is another graph clustering approach that is inspired by the idea of random walks on graphs. It has been applied to several areas such as protein interaction networks (Krogan et al., 2006; Pu et al., 2007) and detection of orthologous genes (Chen et al., 2006). The algorithm proceeds as follows. Suppose $\mathbf{V}$ is the symmetric adjacency matrix of the graph to be clustered. First, the adjacency matrix is added to the identity matrix and then column normalised to give a Markov transition matrix that we will call $\mathbf{J}$ in keeping with the convention of this paper.

$$J_{ij} = \frac{V_{ij} + \delta_{ij}}{1 + \sum_k V_{ik}}$$

A matrix inflation operator Γ_r is defined which raises each element to the power r and then column-normalizes the matrox,

$$(\Gamma_r M)_{ij} = \frac{M_{ij}^r}{\sum_k M_{ik}^r}$$

The Markov clustering proceeds by iterative rounds of expansion and inflation. Let us define a sequence of matrices $\mathbf{M_1}, \mathbf{M_2}, \mathbf{M_3}$ that may be initialized by $\mathbf{M_0} = \mathbf{J}$. Then, at each subsequent

step we compute the next matrix as

$$\mathbf{M_{t+1}} = \Gamma_r \mathbf{M_t}^e,$$

where the matrix raised to the (integer) power e is computed by successive matrix multiplications. Eventually, $\mathbf{M_t}$ converges to the idempotent steady state matrix $\mathbf{M}_\infty$. The conventional view is that the Markov matrix describes the probability of a random walker moving from one vertex to the other. Powers of a matrix (expansion) is the walker taking multiple steps, and the inflation operator is a trick to prevent the successive steps leading to a uniform diffusion across the whole graph by rewarding the larger elements in column and shrinking the smaller elements.

3.4.2 The mean field Potts model

We have mentioned that superparamagnetic clustering proceeds by finding the pair-wise spin-spin correlation for a Potts model using Monte Carlo simulation. In this section we simply write down the equivalent mean field equations for the average magnetization of a Potts model. Assume the spins s_i take values in the set $\{1, 2, 3, \ldots q\}$. Now, the average magnetization m_i at the vertex i is a $q \times 1$ vector where $\sum_{k\in\{1,2,3,\ldots q\}} m_{ik} = 1$. The average magnetizations for all the N vertices of a graph can be described by a $N \times q$ matrix $\mathbf{m}$ where $\sum_r m_{ik} = 1 \forall k$.

The self-consistent mean field equations can thus be derived in a manner similar to the Ising case,

$$m_{ik} = \frac{\exp(\beta \sum_j J_{ij} m_{jk})}{\sum_{t\in\{1,2,3,\ldots q\}} \exp(\beta \sum_j J_{ij} m_{jt})}$$

3.4.2.1 A particular approximation to the mean field Potts model

Consider a special case of the Potts model where possible spin states for a single vertex are equal to the total number of vertices in the graph, $q = N$ where N is the number of vertices. Let us try to solve the successive approximations to the average spins using the mean field method. Due to the symmetric nature of the problem, we start with the average magnetization matrix $\mathbf{m}^{(0)} = \mathbf{I}$. Let us calculate a matrix $\mathbf{w}$ where

$$w_{ik}^{(0)} = \sum_j J_{ij} m_{jk}^{(0)} = J_{ij}\delta_{jk} = J_{ik}.$$

Note that effectively, in terms of the random walk or Markov clustering, this is similar to the expansion step.

Now, the subsequent approximation to the average magnetization will be

$$m_{ik}^{(1)} = \frac{\exp(\beta w_{ik}^{(0)})}{\sum_t \exp(\beta w_{it}^{(0)})},$$

which is similar to the inflation step.

Note that

$$\exp(x) = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots.$$

Consider the case where $x > 1$. The terms t_i in this series keep growing for all $i < x$ and then start declining $i > x$ with a plateau around the point i close to x. For a low temperature limit, β will be high and therefore $\beta \sum_j m_{jr} J_{ij} = \beta w_{ir}$ will be greater than 1. Let us approximate the exponential $\exp(\beta w_{it}^{(0)})$ by some multiple of the largest term in the series to give $A(\beta w_{it}^{(0)})^r$ for some exponent r.

Therefore,

$$\begin{aligned} m_{ik} &= \frac{\exp(\beta w_{ik})}{\sum_t \exp(\beta w_{it})} \\ &\simeq \frac{A(\beta w_{ik}^{(0)})^r}{\sum_t A(\beta w_{it}^{(0)})^r} \end{aligned}$$

or,

$$\mathbf{m^{(n+1)}} \simeq \Gamma_r \mathbf{w^{(n)}}$$

Even if our approximation is far from exact, the effect of the exponentiation and normalization by the site-wise partition function is similar to the effect of the inflation operator in Markov clustering.

We therefore hypothesize that the MCL algorithm in effect calculates a matrix close to the average magnetization of a Potts model near a low-temperature phase transition.

The matrix multiplication of the solved matrix **M** with itself rather than with the inital Markov matrix **J** accelerates the stable magnetization within microdomains by inreasing the interaction strength. A physical analogue of this effect is cooperativity, in which greater spin correlation leads to greater ferromagnetic coupling between the vertices, possibly by motion towards each other in physical space to increase the interaction strength J_{ij}. Indeed, ferrofluid systems have been physically demonstrated (Khalil et al., 2012) and mathematically modeled (Gruber and Griffiths, 1986; Palm and Korenivski, 2009).

3.4.3 Superparamagnetic clustering

Blatt et al. (1996) introduced the idea of paramagnetic clustering. They first induce a graph over a collection of points in Euclidean space, with edge weights inversely proportional to the distance between points. Considering a Pott's model, the energy of the system is defined as

$$H(\mathbf{s}) = -\sum_{uv} J_{uv}\delta_{s_u s_v}$$

where each s_u can take a total of q discrete values ($\{1, 2, \ldots q\}$).

They use Monte Carlo simulations of the state space (spins at each node) and evaluate the spin-spin correlation at different temperatures. The temperature at which the super-paramagnetic phase transition occurs (as inferred from the jump in magnetic susceptibility) is selected. The spin-spin correlations at neighboring points are calculated and a thresholding of these correlations is used to separate the clusters.

In this study, we calculated the spin-spin correlations using analytical approximations rather than Monte Carlo simulations. This makes our work suitable for large scale networks for which Monte Carlo simulations might be slow to converge and computationally much more expensive. While the original superparamegnetic clustering algorithm used a Potts model, we argue that in fact the Ising model is sufficient for clustering into more than two communities. This is because the spin-spin correlations are decoupled across different magnetic domains, even if there are only two choices for the direction of the spin. Therefore, our spin-spin correlation function could be used in a similar fashion to the super-paramagnetic clustering algorithm for clustering or community detection.

3.5 Conclusions

The spin-spin correlation on an Ising model on graphs has a similar form to the graph diffusion kernels commonly used in complex network analysis problems. The spin-spin correlation estimate commonly used in statistical physics models is derived from the linear response of a spin due to an applied external field at a different site. This is valid at temperatures above the Curie point, below which a net non-zero magnetization develops without external magnetic fields. This important property is used to predict phase transitions observed in real materials. For the link prediction problems we consider, we show that the phase transition inverse temperature (β) is the inverse of the largest eigenvalue of the adjacency matrix of the graph. We also present two novel analytical approximations to the correlation that saturate to the physically valid limit of 1 at temperatures near and below the phase transition.

Applying our spin-spin correlation approximations to a link prediction problem in a protein-protein interaction network, we show that the spin-spin correlation outperforms graph diffusion kernels. Also, the optimal temperature where the correlation is the best at predicting missing links is near the phase transition temperature. We interpret this to mean that at a certain temperature, the spins are strongly correlated within meaningful clusters or microdomains. The novel spin-spin correlation expressions we derived in this work are robust with respect to the temperature at which the graph structure is correctly "understood" in terms of the spin-spin correlation matrix. For one of the datasets (the *Plasmodium* PPI network), we also see a marked improvement in the optimal performance in addition to the robustness in temperature. We believe that for certain problems, the optimal temperature could be different in different regions of the graph. For the linear response function, the correlation could catastrophically break down in one region when we optimize the temperature to understand the community structure in another region. However, since the new expressions of the correlation degrade more gracefully, this could help in a better overall performance at the globally optimal temperature.

We have also shown theoretical links between graph diffusion, the Potts model, super-paramagnetic clustering, and Markov clustering. Although link prediction and clustering address two different problems, it is clear that for many of these methods that assume higher link probability within communities compared to between communities, similar mathematical expressions underlie node similarity and co-membership. Studies have explored link prediction as a means to improve clustering (Burgess et al., 2016), and conversely have used community information to improve link

prediction (Soundarajan and Hopcroft, 2012). In fact, methods described at the start of this chapter using biological information to assist network analysis can be interpreted in this fashion.

This work thus provides a unified view for widely used methods for network analysis derived from statistical physics and stochastic processes. Our novel approximations motivated by mean field theories improve the performance over more standard linear response approximations, with virtually no increase in computational cost. Network analysis problems including finding additional members of communities, priotizing candidate genes, and clustering graphs are becoming increasingly important across many fields. Our methods provide new insight and may show better performance across many related problems and fields.

Chapter 4

Abscisic acid response in *Brassica napus* guard cells

4.1 Abstract

Drought is a major threat to food crops and is poised to become increasingly so with climate change. Abscisic acid (ABA) is the main hormone signaling drought or water deficit stress in plants. The present study aims to investigate the response to ABA in the guard cells of *Brassica napus*, an important oil crop. Stomatal closure is mediated by ABA-induced guard cell changes in turgidity, which limits the loss of moisture from leaves. We sequence mRNA derived from guard cell protoplasts treated with ABA and quantify the changes in gene expression. We find far-reaching changes in gene expression, affecting a host of physiological processes including stomatal movement, changes in metabolism, seed germination, and light response. We also find some evidence supporting fast and slow responses to ABA. The transcription factors and regulatory networks mediating these responses are in agreement with what is known from *Arabidopsis thaliana*. We integrate network data to suggest additional genes and interactions that are important in the response to ABA in guard cells. We also find evidence of the continuing evolution of the ABA response in Brassica since its divergence from the common ancestor it shares with Arabidopsis.

4.2 Introduction

4.2.1 Problem description

This study aims to elucidate drought response pathways in *Brassica napus* guard cells. We use mRNA sequencing, *Arabidopsis-Brassica* homolgy, interaction network data, and the knowledge of cis-acting regulatory components to understand the response to ABA in guard cells. Drought results in water deficit in the plant, and drought often co-occurs with higher temperatures, further aggravating the deficit due to increased evaporation rates. It is well established that ABA is the hormone that signals drought stress. Guard cells in leaves control the stomatal openings that are the major means of exchange of water vapor, carbon dioxide, and oxygen with the surrounding air. By constricting the stomatal opening, the guard cell modulates the amount of water loss from the leaves, while balancing the carbon dioxide exchange and subsequent photosynthesis.

4.2.2 Motivation

Understanding the genetic and molecular basis of how crops respond to droughts could enable further development and genetic engineering of drought resistant crops, which will be essential for the social and economic well-being of most of the world. Globally, the total water consumption due to crop production was estimated to be 6490×10^9 m^3/year in 2007, more than 70% of the world's total water consumption (Hoekstra and Chapagain, 2006). High income countries with larger amounts of meat consumption require a larger portion of this water for animal agricultural use. The per-capita US water consumption of 2480 m^3/person/year is twice the global average of 1240 m^3/person/year (Hoekstra and Chapagain, 2006).

Climate change projections predict increasing frequencies of extreme climate events, including droughts (Dai, 2012; Trenberth et al., 2013). Climate change threatens water supplies for agriculture, with effects on food security (Wheeler and von Braun, 2013). Food insecurity is a major political risk factors (Schwartz and Randall, 2003; Barnett and Adger, 2007) and it is likely that droughts may have precipitated recent conflicts (Kelley et al., 2015; Gleick and Gleick, 2014) and wars (Couttenier and Soubeyran, 2014; Fjelde, 2015). Some studies predict that the US southwest may enter a long period of increased aridity (Seager et al., 2007; Cook et al., 2004). To sustain agriculture in the southwest, simply increasing water storage or savings from domestic use is not likely to be sufficient, and agricultural water use will have to be reduced (MacDonald, 2010).

Drought and heat stress adversely affect crop productivity (Ciais et al., 2005) and plant growth. Climate change forecasts predict that 20-60 Mha of irrigated cropland will be forced to rely on rain-fed agriculture (Elliott et al., 2014), and crops will need to withstand the vagaries of nature, including drought and rising temperature.

Brassica napus is an important food crop as the source of an edible oil popularly known as Canola or rapeseed oil. *Brassica napus* was developed (Jonnson, 2009; Qiu et al., 2006; Kondra and Stefansson, 1965) to lower the total content of anti-nutrients (components that interfere with absorption of minerals), such as erucic and eicosenoic acids and glucosinolates, and has found acceptance as a major food source.

Brassica napus belongs to the *Brassica* genus, which includes many other important food crops such as mustards, turnip, broccoli, and cabbages. The larger *Brassicaceae* family also includes the *Arabidopsis* genus and the model plant *Arabidopsis thaliana*. Insights drawn from *Brassica napus* can thus be used to understand parallel mechanisms in *Arabidopsis thaliana*, and candidate genes identified through Brassica may be often be tested using Arabidopsis orthologs. Therefore, understanding the mechanism of drought response in *Brassica napus* is of practical and immediate importance for agriculture and the economy, and it can also be used to help decipher the biology of drought resilience in other plants of agricultural or scientific interest.

ABA is the main phytohormone signaling drought stress in plants. One of its most important activities is to control stomatal closing, thereby reducing water loss. Unfortunately, closing the stomata also impedes the transfer of CO_2 needed for photosynthesis. However, Yang et al. (2016) showed that over-expressing the ABA receptors in Arabidopsis thaliana showed increased water use efficiency. This suggests that there are ways to modulate stomatal opening to reduce water loss while retaining high rates of photosynthesis. A deeper understanding of the mechanism and signaling controling stomatal closure in guard cells could lead to engineering crops that do not just survive periods of drought but maintain their accumulation of biomass. In addition, more plant growth in water scarcity can also help to capture atmospheric carbon.

4.2.3 Previous work and known biology

4.2.3.1 Mechanisms of drought response

Conditions of drought or water deficit are sensed in both the root and the leaves. The water deficit signal sensed in the root is communicated to the leaves and elicits the adaptive response (Schulze and Hall, 1982). Experiments in apple (Gowing and Davies, 1990) and maize (Blackman and Davies, 1985) demonstrate that there are chemical signals from the roots to the leaves that result in stomatal closure and re-opening. Other experiments in Arabidopsis (Christmann et al., 2007; Assmann et al., 2000), birch (Saliendra et al., 1995), fir (Fuchs and Livingston, 1996), and *Hymenoclea salsola* (Comstock and Mencuccini, 1998), however, have also uncovered evidence of a hydraulic mechanism from the root to the shoot through the xylem that transmits the drought signal. In a study of wilting of the common bean *Phaseolus vulgaris L.* (Qin and Zeevaart, 1999), it was shown that ABA synthesis was controlled by the gene PvNCED1 and that the ABA synthesis was more pronounced in the leaves than the roots. Bauer et al. (2013) has shown that the guard cell has the complete functional machinery for ABA synthesis, and that ABA synthesis in the guard cell can amplify the received ABA signal and is also responsive to decreases in relative humidity. This suggests that the drought sensing and signaling may also originate in the leaves.

Plants employ a number of strategies to survive and grow during periods of reduced water availability. Attempting to reduce evaporative water loss through stomatal closure is generally considered a drought avoidance mechanism (Fang and Xiong, 2015). Leaf wilting (Poorter and Markesteijn, 2008) and rolling (Begg, 1980) under high solar radiation to decrease the evaporative surface area are similar mechanisms. Other drought avoidance traits could be decreasing the hydraulic conductance of the root-to-stem loss of water into the ground (North and Nobel, 1992; Lo Gullo et al., 1998).

Drought escape mechanisms aim to hasten development to complete the life cycle of the plant before maximum drought severity. Early flowering is one of the phenotypes associated with this strategy. There is evidence that ABA stimulates earlier onset of flowering in response to longer days (Riboni et al., 2013).

4.2.3.2 The role of ABA

ABA has long been recognized as a chemical signal mediating the drought response in plants (Jones and Mansfield, 1970). In addition to its role in drought response, ABA is also involved in mediating

the response to heat stress (Heikkila et al., 1984; Larkindale and Knight, 2002). The stress response in leaves has been shown to have ABA-dependent and ABA-independent pathways (Shinozaki and Yamaguchi-Shinozaki, 1996). In particular, a number of genes have a cis-acting ABA response element (ABRE) (Choi et al., 2000) that mediates the ABA-dependent stress response. The corresponding dehydration response element (DRE) mediates the ABA-independent stress response. The drought response element also mediates the cold stress response. Some genes are activated by both the ABA-dependent and ABA-independent pathways, and there is considerable cross-talk between them (Narusaka et al., 2003; Nakashima et al., 2014).

Water uptake by the roots may also increase in response to ABA levels in the root (Hose et al., 2000). Water deficit sensed in the leaves results in reduced stomatal water loss (Comstock and Mencuccini, 1998) though reduced stomatal aperture. This reduces evaporative water loss and extends plant survival, although it may also reduce photosynthetic activity impeding the uptake of CO2 (Cornic, 2000; Panek and Goldstein, 2001; Medrano et al., 2002). It is tempting to suppose that rising atmospheric concentrations of CO2 might offset the effects of this lower stomatal conductance in response to drought, but unfortunately it is becoming clear that this is not the case (Leakey et al., 2006; Pataki et al., 1998).

4.2.3.3 Phylogeny of *Brassica napus*

Brassica napus was formed from the polyploid hybrid speciation of *Brassica rapa* and *Brassica oleracea* (Allender and King, 2010; Song and Osborn, 1992). The amount of synonymous nucleic acid substitution rates (Chalhoub et al., 2014) and low degree of chromosomal rearrangements (Parkin et al., 1995) point to its relatively recent speciation event about ten thousand years ago. The 19 *Brassica napus* chromosomes are derived from the 10 chromosomes (subgenome A) of *Brassica rapa* and the 9 chromosomes (subgenome C) of *Brassica oleracea*. The *Brassica* species, along with the much more well studied plant *Arabidopsis thaliana*, are members of the Brassicaceae family. Studies point to an estimated age of 10-20 million years for the *Arabidopsis*/*Brassica* split and to a number of genome duplication events giving rise to the Brassica species from its common ancestor with Arabidopsis (Yang et al., 1999; Ermolaeva et al., 2003; Blanc et al., 2003).

Numerous studies of drought stress (Rizhsky et al., 2004; Seki et al., 2002) and ABA response (Leonhardt et al., 2004) in *Arabidopsis thaliana* and its genetic similarity with *Brassica napus* provide a framework to compare and interpret our results in *Brassica napus*. We will be able to investigate

whether the genetic components are conserved across lineages. Since the *Brassica napus* genome has many times more genes than *Arabidopsis thaliana*, it is likely to have more complex or redundant pathways, and we can investigate whether the gene copies resulting from polyploidy have had time to evolve different functions. In addition, we are able to use the knowledge of the regulatory interactions in *Arabidopsis thaliana* to suggest some causal mechanisms of the ABA response in *Brassica napus*.

4.2.3.4 Other ABA response studies

Previous studies have profiled the changes in gene expression in response to ABA. Hoth et al. (2002) profiled *Arabidopsis thaliana* mRNA levels in whole seedlings, comparing wild type with a mutant with ABA-insensitive phenotype, using massively parallel signature sequencing (MPSS). Since the mRNA in that study was extracted from whole seedlings, the genes identified as differentially expressed could have been involved in any of several ABA response processes, including germination, flowering, and stomatal movement, rather than those specific to guard cells or stomatal movement. Guard cell specific ABA response genes were identified by Leonhardt et al. (2004), who separately quantified expression levels from isolated *Arabidopsis thaliana* guard cells and mesophyll cells after ABA treatment of the whole plant. While this experiment was able to find the guard cell specific genes that were regulated, they could have been regulated either by ABA sensed directly in the guard cell or by an indirect mechanism wherein ABA sensed in a different cell regulates gene expression downstream in the guard cell.

Other gene expression studies have also investigated the role of ABA in disease susceptibility and interactions with other signaling molecules (Anderson et al., 2004) in the whole plant, or elucidated the role of ABA-response element transcription factors in the mesophyll cells (Yoshida et al., 2015, 2010). Gene expression profiles in other organisms have also been studied in other organisms such as rice in response to ABA treatment of seedlings (Rabbani et al., 2003).

4.2.4 Approach used in this study

The present study looks at the genome-wide transcription response of ABA in the guard cells of *Brassica napus*. We focus on the plastic drought avoidance traits in the leaf, specifically the control of stomatal conductance.

We extracted mRNA from guard cell protoplasts of *Brassica napus* after 15 minutes and 60 minutes of ABA treatment. mRNA was then fragmented and sequenced with the Illumina platform, generating single end short reads of approximately 100 basepairs. Short reads were then aligned to the *Brassica napus* genome and differential gene expression is quantified.

As mentioned above, earlier studies primarly involved application of ABA to either whole plants (or seedlings) or leaves. By applying ABA to isolated guard cell protoplasts, we study in isolation the genes regulated directly by ABA in the guard cell specifically, excluding intermediary signaling from mesophyll cells or the cell wall barriers. By using the protoplasts, we ensure that ABA is able to robustly and promptly initiate the signaling pathways without any hindrance. We extract mRNA at only 15 minutes and 60 minutes of application. We aim to uncover ABA dynamics at this time scale, such as differentiating between genes with early and delayed responses to ABA, or genes that may return to baseline expression after an early response due to negative feedback. In addition, earlier studies on ABA response in other species were performed using microarrays rather than direct sequencing; therefore, they were limited to the genes pre-selected for the microarray. For example, the study by Leonhardt et al. (2004) used a microarray with about 8100 genes, which is a third of the estimate of the total number of protein coding genes in *Arabidopsis thaliana*. In contrast, by employing RNA sequencing and aligning to the genome, we are able to potentially capture the whole transcriptome and all the annotated genes rather than a subset. The resolution of these genes, however, is limited by the number of reads captured, sequence similarity of paralogs, and sequencing noise.

The *Brassica napus* genome is essentially a number of copies of the *Arabidopsis thaliana* genome with various degrees of divergence and fragmentation. This is due to the evolutionary history of genome duplication, hybridization, and gene loss in *Brassica napus* since its split from the common ancestor with *Arabidopsis thaliana*. Since genome duplications are not thought to have occured in the Arabidopsis lineage, we use *Arabidopsis thaliana* genes as a proxy for the ancestral origin for families of *Brassica napus* paralogs. We use these sets of paralog families to integrate mRNA expression with DNA regulatory elements to explore the evolution of ABA response in *Brassica napus*. We also compare the regulation of gene expression in *Brassica napus* with the known biology of *Arabidopsis thaliana* and draw conclusions regarding the evolution of the ABA response.

The integration of pathway models and interaction data with gene expression data enables us to find up-regulated pathways. The promoters of the corresponding genes contain DNA regulatory elements

that suggest the most likely transcription factors mediating the ABA response and the genes that are activated downstream of ABA signaling. These results could be used to select gene candidates for further experiments.

4.3 Results

4.3.1 Temporal dynamics of the ABA response

Guard cell protoplasts were extracted from 5-7 day old leaves from *Brassica napus* plants grown under 60% relative humidity. Sets of extracted guard cell protoplasts were treated with ABA for 15 minutes, 60 minutes, and an ABA-free solution each. The entire process of the three treatments was repeated three times for a total of nine samples. mRNA was extracted from the treated protoplasts of each sample, and sequenced on the Illumina platform to give 100 bp reads. These reads were aligned against the *Brassica napus* genome of the Darmor-bzh line (Chalhoub et al., 2014) using the Tophat-Bowtie pipeline (Kim et al., 2013). Differential expression for 15 minutes of ABA and 60 minutes of ABA was estimated using DESeq2(Love et al., 2014).

In this section, we explore whether there is a time-varying ABA response of interest.

4.3.1.1 Genes are similarly regulated under both the early (15 minutes) and late (60 minutes) response, with a larger effect at 60 minutes

By measuring the gene expression at two different times, we identify the temporal response to ABA. Possible responses are short lived (around 15 minutes), delayed (around 60 minutes), growing with time, or stable from 15 to 60 minutes. We performed exploratory analysis comparing response at 15 and 60 minutes of ABA vs. baseline to categorize genes by their temporal response.

The relationship between mRNA expression at 15 minutes and 60 minutes for individual genes is shown in Figure 4-1, restricted to genes that are significant at either time vs. baseline ($< 5\%$ FDR). Expression levels at the two time scales is correlated. The log2 fold changes are correlated with an R^2 of 0.53 and a highly significant p-value $< 2.22e-16$.

Most genes show the same direction of regulation at the two time points, with a larger change at 60 minutes. The slope of the best fit line is 2.39, corresponding to a 4-fold increase from 15 minutes to 60 minutes. This would be approximately the expected increase assuming a constant rate of mRNA production starting instantaneously at ABA application with a zero initial mRNA concentration and no degradation. Thus, the dominant pattern for differential expression is consistent with immediate up-regulation in response to ABA, with a constant production rate and no degradation in the first hour of response.

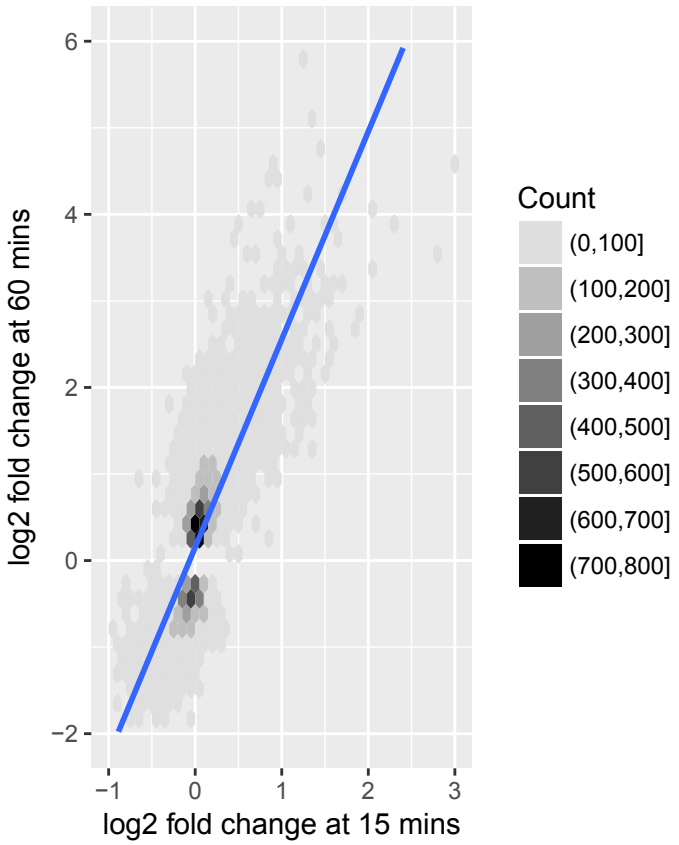

Figure 4-1: Scatter plot of the correlated gene expression values at 15 and 60 minutes of ABA treatment. We plot the log2 fold changes at 15 minutes vs. 60 minutes of ABA treatment for all genes that are significantly regulated ($<$ 5% FDR) at either time. Each cell in the plot shows the density of genes with a particular combination of log2 fold changes at each time point. The straight line is the linear best fit. Genes that are not significantly differentially expressed at either time are excluded; the best fit estimate reflects stable regulation rather than background variation.

Note that a similar pattern is observed for genes whose expression levels decrease, with the decrease greater at 60 minutes than 15 minutes, but there are fewer genes in this category than in the consistently increasing category. The magnitude of the increased expression could be significantly larger because a larger protein machinery is required to afford the cell proetection from the resulting stress. On the other hand, the housekeeping and photosynthetic processes only slow down rather than shutting down which might explain the comparatively smaller decrese in expression. Choosing only the subset of genes with decreasing expression, the slope of the best fit line is between 1.1251077 and 1.6376437 (with and without the intercept), suggesting that the decay rate in 60 minutes is slightly lower than the first 15 minutes.

We next take the larger fold change at 60 minutes of treatment than 15 minutes of treatment as a null hypothesis and investigate genes whose differential regulation falls outside this pattern, suggesting additional mechanisms of gene regulation. We tabulate the number of genes up-regulated, down-regulated, or unchanged at each timepoint, using a 5% FDR threshold for differential expression vs. baseline, in Table 4.1.

Table 4.1: Contingency table showing the number of genes with significant positive and negative differential expression ($< 5\%$ FDR) for 15 minutes and 60 minutes of ABA treatment.

	Down regulated at 60 min.	Insignificantly regulated at 60 min.	Up regulated at 60 min.
Down regulated at 15 min.	64	7	0
Insignificantly regulated at 15 min.	4282	89104	7177
Up regulated at 15 min.	0	4	402

Almost all genes identified as differentially expressed at 15 minutes are also differentially expressed at 60 minutes with concordant direction. The log-ratios seem are greater 60 minutes, giving the test at 60 minutes higher statistical power and identifying more genes as differentially expressed. It is therefore difficult to distinguish between genes with a delayed response and those with an immediate response but that simply lack significance at 15 minutes.

The other, smaller group of genes that show statistically significant regulation at 15 minutes but whose differential expression at 60 minutes is either in the opposite direction or statistical significant represent an interesting temporal response that is not explained by the general pattern of sustained activation. We study these genes further in the Section 4.3.1.2.

We also observe that a larger number of genes are up-regulated than down-regulated. This may have a physiologically interpretion that the ABA response and stomatal closing require additional machinery that the cell needs to generate while maintaining the baseline activities from its quiescient state prior to the onset of drought.

4.3.1.2 Temporal dynamics suggest roles for genes in the early response to ABA

For most genes we do not detect a statistically significant differential expression at 15 minutes, and many genes that are significantly differentially expressed at 15 minutes show the same direction of regulation (up or down) at 15 minutes and 60 minutes. There are, however a few genes showing significant differential expression at 15 minutes, with either opposite differential expression or no significant effect at 60 minutes. The levels of these genes for the 3 time points, corrected for batch effects, are plotted in Figure 4-2.

Table 4.2: Genes with significant differential expression at 15 minutes with no regulation or regulation in the opposite direction at 60 minutes of ABA treatment.

	Temporal Pattern	*B. napus* gene	*A. thaliana* gene	Common name	FDR at 15 min.	FDR at 60 min.
BnaA05g12320D	Only down-regulated at 15 mins	BnaA05g12320D	AT2G30040	MAPKKK14	0.00986	0.195
BnaA08g02360D	Only down-regulated at 15 mins	BnaA08g02360D	AT1G49850		0.03198	0.683
BnaA09g11950D	Only down-regulated at 15 mins	BnaA09g11950D	AT1G64090	RTNLB3	0.02868	0.071
BnaA09g27780D	Only down-regulated at 15 mins	BnaA09g27780D	AT1G27730	STZ	0.00735	0.814
BnaC03g39060D	Only down-regulated at 15 mins	BnaC03g39060D	AT3G15353	ATMT3	0.02381	0.718
BnaC05g21480D	Only down-regulated at 15 mins	BnaC05g21480D	AT1G27730	STZ	0.00029	0.889

	Temporal Pattern	*B. napus* gene	*A. thaliana* gene	Common name	FDR at 15 min.	FDR at 60 min.
BnaC06g40000D	Only down-regulated at 15 mins	BnaC06g40000D	AT1G79660		0.02328	0.485
BnaAnng27240D	Only up-regulated at 15 mins	BnaAnng27240D	AT4G19230	CYP707A1	0.00192	0.074
BnaC01g11650D	Only up-regulated at 15 mins	BnaC01g11650D	AT4G19230	CYP707A1	0.01032	0.536
NA	Only up-regulated at 15 mins	BnaC05g31990D			0.01387	0.088
BnaC06g32150D	Only up-regulated at 15 mins	BnaC06g32150D	AT1G71010	FAB1C	0.01980	0.513

Table 4.3: Genes with opposite directions of significant regulation in the first 15 minutes of ABA treatment and the last 45 minutes of ABA treatment.

B. napus gene	*A. thaliana* ortholog	Common name	log2 fold change for 0 to 15 mins.	log2 fold change for 15 to 60 minutes	FDR for 0 to 15 mins.	FDR for 15 to 60 mins.
BnaA08g02360D	AT1G49850		-0.52	0.41	0.03	0.02
BnaA09g27780D	AT1G27730	STZ	-0.43	0.48	0.01	1.07e-04
BnaAnng27240D	AT4G19230	CYP707A1	0.39	-0.32	0.02	0.02
BnaC01g11650D	AT4G19230	CYP707A1	0.46	-0.43	2.90e-04	7.99e-05
BnaC03g39060D	AT3G15353	ATMT3	-0.77	0.54	0.02	0.05
BnaC05g10770D	AT1G14500		-0.81	0.42	3.29e-09	3.76e-03
BnaC05g31990D			0.81	-0.60	0.01	0.02

B. napus gene	*A. thaliana* ortholog	Common name	log2 fold change for 0 to 15 mins.	log2 fold change for 15 to 60 minutes	FDR for 0 to 15 mins.	FDR for 15 to 60 mins.
BnaC06g40000D	AT1G79660		-0.42	0.30	0.02	0.04

Table 4.2 lists the genes showing significant differential expression at 15 minutes in a different direction from 60 minutes of ABA treatment. Although only a few genes are identified here, either because of the limited statistical power at 15 minutes due to a smaller fold change or because only a few genes are involved in the initial transient response, they hint at an initial transient response to ABA. Previous literature points to the role of some of these genes. FAB1C is known to be responsible for fast closure of the stomata and its mutation causes slow stomatal closure (Bak et al., 2013). These observations support our data showing that FAB1C is up-regulated for this initial response and it subsequently returns to its basal expression, with less contribution to the later ABA response. In the context of ABA signaling in seed maturation and dormancy, CYP707A1 is known to be expressed in mid-maturation and is then down-regulated in late maturation (Okamoto et al., 2006). Perhaps a similar feedback mechanism could be present in the guard cells and responsible for the initial up-regulation and then return to baseline expression of both CYP707A1 and FAB1C.

Since we only test for the presence of regulation rather than its absence, the selection of genes for null regulation in Table 4.2 may not be statistically sound. The genes are listed mainly for the purpose of suggesting candidate genes that may have an early-only response. To gain statistical confidence for an early response, we calculate the differential expression between the 15 minutes and 60 minutes of ABA treatment, and select genes that show the opposite direction of differential expression compared to the first 15 minutes. These are listed in Table @ref(tab:opp.genes.table). We see that most of the genes are selected in this method as well. We have selected genes with a corrected p-values (i.e., false discovery rate) of less than 0.05 for both tests of differential regulation. Since these are two correlated tests, the genes are actually selected conservatively, with an overall false discovery rate between 0.025 (if the tests were independent) and 0.05 (if the tests were perfectly correlated).

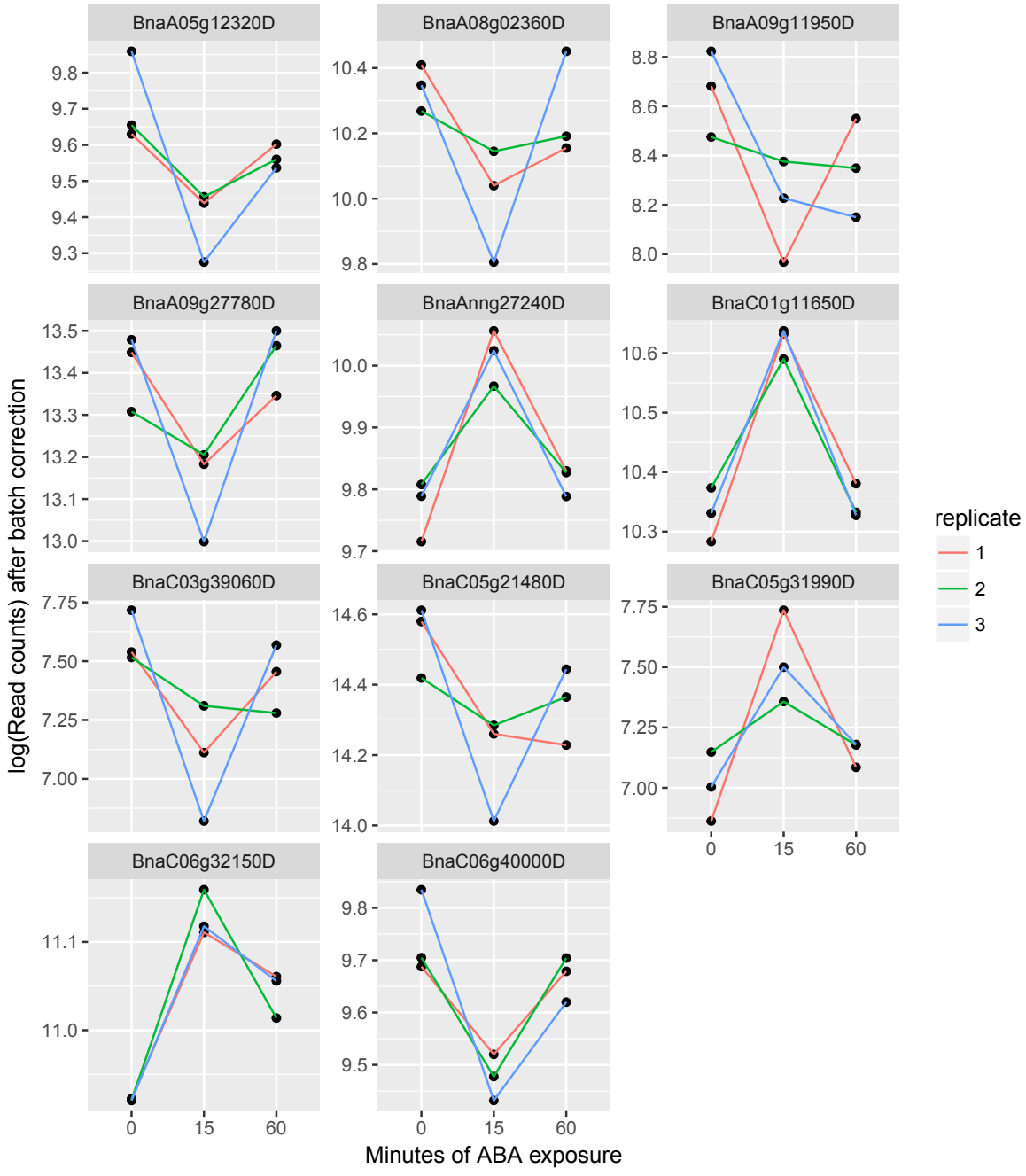

Figure 4-2: Batch corrected read counts for genes showing significant differential expression at 15 minutes in a different direction from 60 minutes.

4.3.2 Gene regulation is conserved within paralogous families

Since *Brassica napus* is the product of a number of genome duplication and fractionation events, its genome contains paralogous families each descended from a common ancestral gene. In order to understand the evolution of the drought response in *Brassica napus*, we investigated whether the responses of individual genes within a single family had diverged. If the regulatory response to ABA in paralogous gene families had completely diverged, we would expect that the fold changes within gene families would not be correlated. We therefore tested whether the ABA response (as measured by the fold change at 60 minutes) of genes within each paralogous family clustered around different means, or whether they all followed a single distribution common to the family. We tested this hypothesis using one-way analysis of variance.

Brassica napus genes were mapped to their corresponding closest *Arabidopsis thaliana* homologs. Since the genome triplication and polyploid hybridization events in *Brassica napus* happened after its split from the *Arabidopsis* lineage, multiple *Brassica napus* genes map to the same *Arabidopsis thaliana* ortholog. For the purposes of this study, these groups of paralogous *Brassica napus* genes are considered to be gene families resulting from duplications of the same ancestral gene. After filtering out genes with very low read counts, hence with inaccurate estimates of fold-change, and including only groups with more than two valid *Brassica napus* genes, we conducted the one-way analysis of variance, with results shown in Table 4.4. The p-value is highly significant.

With a large number of genes and paralogous families, the highly significant p-value could arise from a small effect. We therefore estimated the effect size (η^2), calculated as the ratio of the variance explained by the paralogous families to the total variance. We see that 68.56% of the variance is explained by paralogous group membership. We conclude that while there has been measurable divergence of gene expression in these paralogous families, the regulation is still substantially correlated within gene families.

Table 4.4: Analysis of variance to test whether the differential expression is correlated within paralog gene families.

Source	Sum of Squares (SS)	Degrees of Freedom	Mean SS	F-statistic	log(p-value)	Effect Size (η^2)
Paralogous Group	5857	16575	0.3533609	4.546911	-7022.886	0.6855728
Error	2686	34565	0.0777145			

Source	Sum of Squares (SS)	Degrees of Freedom	Mean SS	F-statistic	log(p-value)	Effect Size (η^2)
Total	8543	51140				

4.3.3 The ABA responses in *Arabidopsis thaliana* and *Brassica napus* are similar

The ABA response and its role in drought response is well studied in both model plants such as *Arabidopsis thaliana* and agricultural crops. We explored whether the results of our experiment agree with the known ABA responses in the model organism *Arabidopsis thaliana*.

We visualized the comparison of the ABA response in *Arabidopsis thaliana* guard cells and *Brassica napus* protoplasts by plotting the changes for each gene in *Brassica napus* at 15 minutes and 60 minutes (Figure 4-3) with the corresponding fold change in expression for its ortholog in *Arabidopsis thaliana*, taken from Wang et al. (2011). Expression fold changes of genes in *Arabidopsis thaliana* that were not significantly differentially expressed were not reported in Wang et al. (2011), creating an empty horizontal band in the figure. It is clear that the correlation is highly statistically significant at both time points, and the R^2 is twice as high at 60 minutes compared to 15 minutes.

Despite its statistical significance, the overall R^2 is small, suggesting a moderate degree of correlation. However, in addition to differences between the species (*Arabidopsis thaliana* vs. *Brassica napus*), additional sources of variation are guard cells vs. protoplasts, microarray vs. RNA-seq experimental technologies, and the 3 hour vs. 15 or 60 minute exposure to ABA. Due to these differences, it is understandable that even genes that occupy a similar role in the transcriptional response may show quantitatively different fold changes. We therefore continued by investigating qualitative patterns of differential expression in the two experimental systems.

Table 4.5: Cross-tabulation of the direction of statistically significant differential expression of *Brassica napus* genes in response to 15 minutes of ABA treatment against the regulation of their corresponding orthologous genes in *Arabidopsis thaliana*.

	Down-regulated in Brassica (15 min.)	Insignificantly regulated in Brassica (15 min.)	Up-regulated in Brassica (15 min.)
Down-regulated in Arabidopsis	**16**	1352	**4**
Up-regulated in Arabidopsis	**0**	1809	**135**

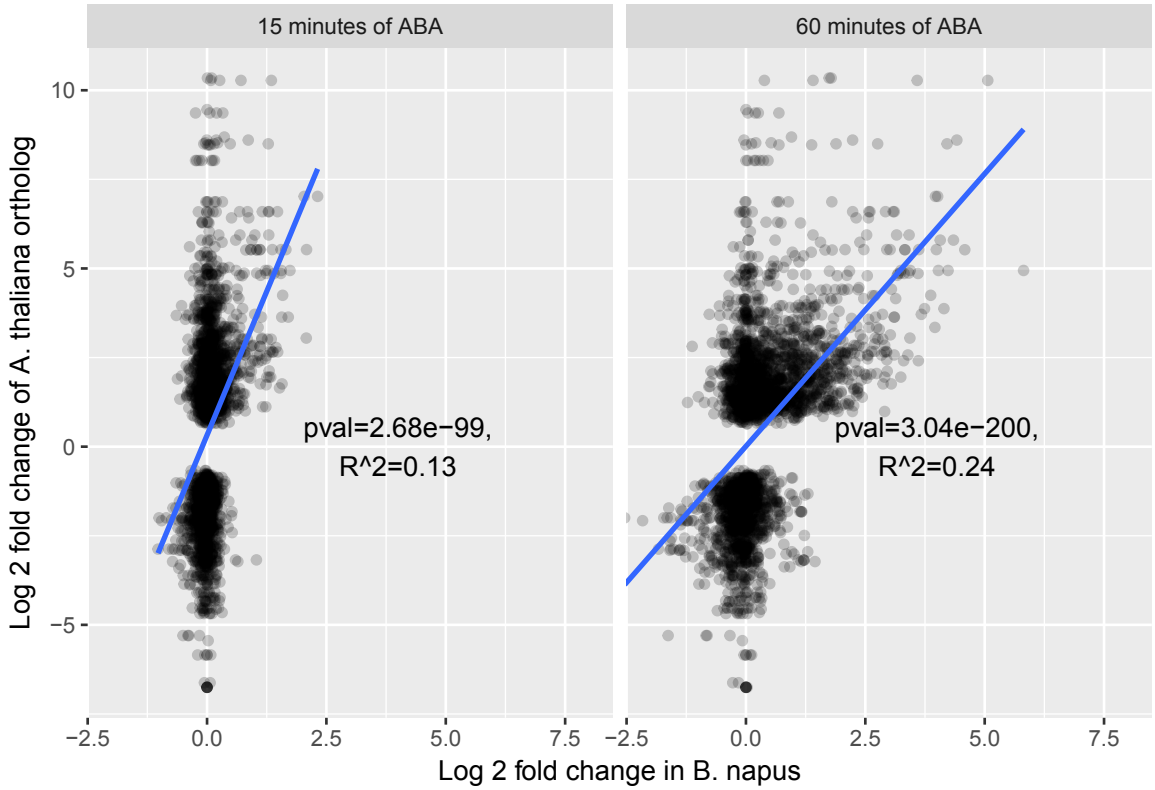

Figure 4-3: Statistically significant correlation of the ABA response in *Arabidopsis thaliana* guard cells and *Brassica napus* protoplasts. The log2 fold change observed in *Brassica napus* after 15 minutes and 60 minutes of ABA treatment is plotted against that of the corresponding *Arabidopsis* ortholog after 3 hours of ABA treatment as reported in Wang et al. (2011). Only significantly differentially expressed genes were reported for the *Arabidopsis* experiment, resulting in the missing horizontal band in the figure. The p-value for the statistical significance of the correlation and R^2 values for each time point are overlayed in the plots.

Table 4.6: Cross-tabulation of the direction of statistically significant differential expression of *Brassica napus* genes in response to 60 minutes of ABA treatment against the regulation of their corresponding orthologous genes in *Arabidopsis thaliana*.

	Down-regulated in Brassica (60 min.)	Insignificantly regulated in Brassica (60 min.)	Up-regulated in Brassica (60 min.)
Down-regulated in Arabidopsis	**186**	1064	**122**
Up-regulated in Arabidopsis	**48**	978	**918**

Table 4.7: Tests of association for the differential expression in response to ABA for *Brassica napus* and *Arabidopsis thaliana* guard cells. For each test, the 2*imes*2 contingency table for significant differential expression in each species was constructed. The p-value was calculated using the Fisher's exact test, and the effect size was measured by Cramer's V. The tests are then repeated by selecting, for each paralogous gene family, the *Brassica napus* paralog with the most significant p-value for differential expression at the corresponding time point.

		Fisher's exact test	
	Brassica napus gene sets	p-value	Effect size (Cramer's V)
2	Regulation at 15 minutes	2.035e-18	0.8815
3	Regulation at 60 minutes	6.679e-93	0.6128
4	Regulation at 15 minutes (only the most regulated paralog)	2.035e-18	0.8815
5	Regulation at 60 minutes (only the most regulated paralog)	4.995e-46	0.5944

To investigate co-regulation by ABA of homologous genes in *Brassica napus* and *Arabidopsis thaliana*, we tabulated the number of *Brassica napus* genes regulated in each direction against their corresponding *Arabidopsis thaliana* orthologs in Tables 4.5 and 4.6 for 15 and 60 minutes of ABA treatment in *Brassica napus*, respectively. While we identify fewer significant genes at 15 minutes, it is clear that their regulation is consistent with the regulation observed in *Arabidopsis thaliana*. More genes are significant after 60 minutes of ABA treatment, but many of these have discrepancies with the regulation in *Arabidopsis thaliana*.

The statistical significance of the agreement between the regulation in *Arabidopsis thaliana* and *Brassica napus* is evaluated by Fisher's exact test, selecting only the counts for genes showing significant differential expression (the columns in bold in Tables 4.5 and 4.6). To examine whether some of the variation between the *Arabidopsis* and *Brassica* gene expression can be explained by *Brassica* polyploidy, we also conducted tests of association using only the paralog that is the most significantly regulated (i.e., with the smallest p-value for differential expression). The results of these statistical tests, along with effect sizes, are reported in Table @ref{tab:bna-at-tests}. Since we have tested only the 2×2 tables excluding genes not significantly regulated in *Brassica napus*, we conclude that while some paralogs may have lost regulation due to lower positive selective pressure, they are distinct from the genes observed to be regulated in opposite directions from *Arabidopsis thaliana*.

The preceding exploratory analysis yielded a small number of genes with opposite response in *Arabidopsis* vs. *Brassica*. Of the genes identified at 15 minutes, only 4 show discordant regulation. These are shown in Table 4.8. These genes may are interesting candidates for further investigation of the evolution of ABA response. NLP1 is an enzyme in the putrescine synthesis pathway (Piotrowski et al., 2003). It is also known that in drought conditions, the levels of putrescine and spermidine are tightly controlled with interconversion between these species (Alcazar et al., 2011). It is possible, therefore, that NLP1 might be over-expressed in certain drought stress conditions (such as that observed in our *Brassica napus* cells), and under-expressed in others. AtMYB30 is a close paralog of AtMYB96 that is involved in regulating wax production, which could provide protection from the harsh, dry environment in drought (Seo et al., 2011). It is possible that the identified gene BnaC02g37590D functions as the equivalent of AtMYB96. HSFA7A is a heat shock protein, with plausible involvement with the drought stress pathway (Port et al., 2004; Nishizawa-Yokoi et al., 2011).

Table 4.8: Genes selected for discordant gene expression in *Atabidopsis thaliana* and *Brassica napus* guard cell ABA responses.

B. napus gene	*A. thaliana* gene	Common name in *A. thaliana*	ATl2fc	log2 fold change in *B. napus* (15 mins.)	log2 fold change in *B. napus* (60 mins.)
BnaA06g25930D	AT2G17150	NLP1	-1.319	0.3356	1.299
BnaA08g21990D	AT1G19620	NA	-3.177	1.035	1.208
BnaC02g37590D	AT3G28910	ATMYB30	-0.9752	0.3724	0.7381
BnaC04g28450D	AT3G51910	AT-HSFA7A	-3.224	0.6457	1.157

4.3.4 Proline and isoprene polymerization pathways are up-regulated

We tested the hypothesis that the metabolism of guard cells changes in response to ABA. We inferred metabolic pathways in *Brassica napus* from the corresponding Arabidopsis orthologs and the metabolic pathways in BioCyc (Caspi et al., 2016).

Table 4.9: BioCyc pathways enriched for *Brassica napus* genes differentially expressed at 60 minutes of ABA treatment

BioCyc pathway ID	pvalue	Pathway
CITRULBIO-PWY	1.455e-14	citrulline biosynthesis
HEXPPSYN-PWY	1.414e-18	hexaprenyl diphosphate biosynthesis
PWY-5783	1.414e-18	octaprenyl diphosphate biosynthesis
PWY-5805	1.414e-18	nonaprenyl diphosphate biosynthesis I
PROSYN-PWY	4.963e-16	proline biosynthesis I
PWY-6922	6.732e-24	L-N-delta-acetylornithine biosynthesis
PWY-3341	2.232e-18	proline biosynthesis III

We searched for the pathways enriched for genes differentially expressed in response to ABA according to the statistical tests described in the Methods (Section 4.5.3.2). A total of 611 pathways were tested, and the p-value cutoff at the family wise error rate was $1.2770332 \times 10^{-11}$ (details in the methods). The enriched pathways are listed in Table 4.9. The various isoprenyl polymerization pathways are actually the same enzyme; and the proline, acetylornithine, and citrulline synthesis pathways share the same up-regulated enzymes.

As an example, we show the proline biosynthesis pathway in Figure 4-4 with the log2 fold change shown next to the gene names. Proline has an established role in drought stress response. Proline accumulation in response to drought has been observed in roots and leaves (Sofo et al., 2004). Proline can be protective of proteins from the stresses of heat and increasing amounts of inorganic solutes characteristic of water loss (Samaras et al., 1995). It has also been hypothesized that proline accumulation is a mechanism of storing energy to be released once the stress is relieved. Proline

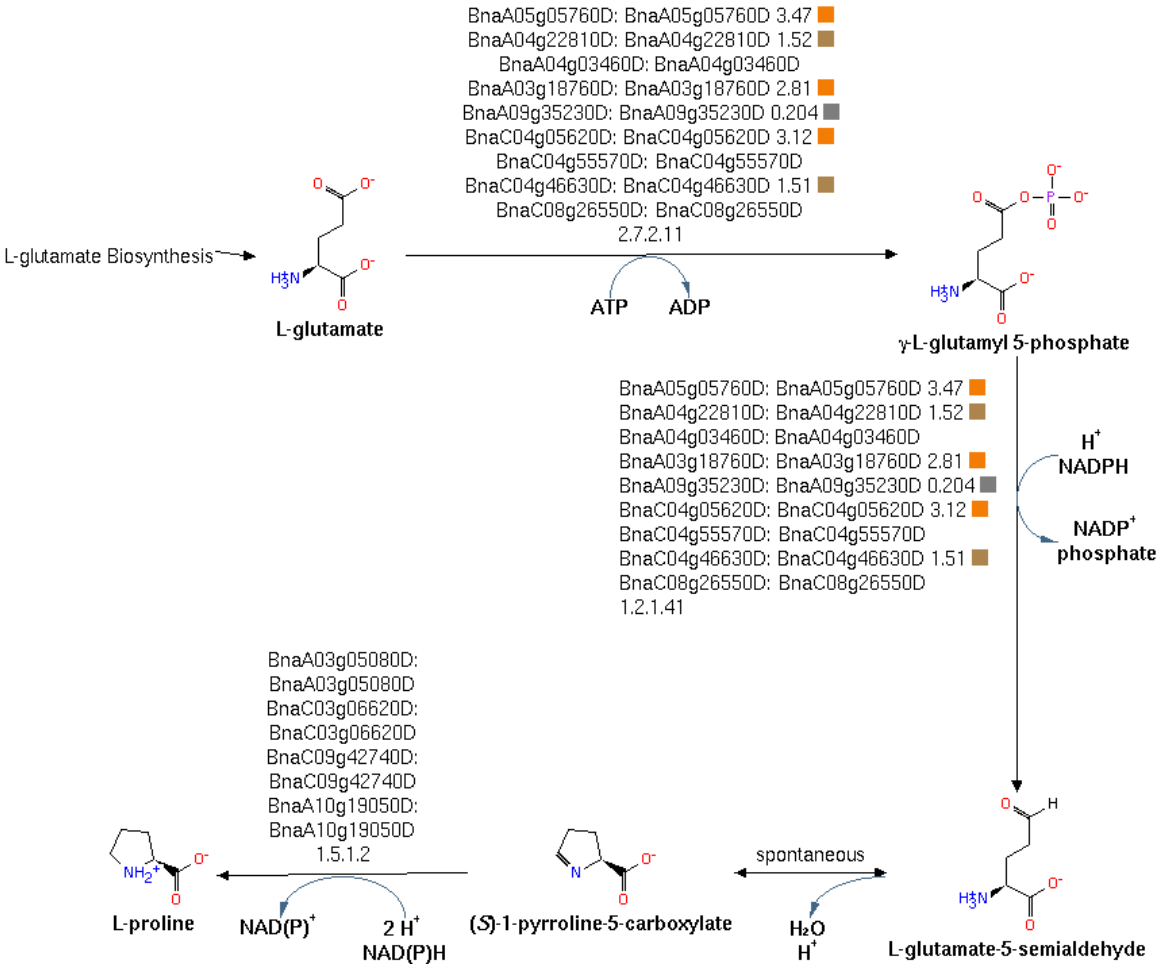

Figure 4-4: Proline biosynthesis is enriched for ABA responsive genes. The enzymes catalyzing each reaction are listed next to the reaction along with the log2 fold change in expression observed at 60 minutes of ABA treatment. The colored squares visually represent the fold change, with brighter colors for greater regulation.

accumulation was found to be correlated with post-stress recovery in sorghum (Blum and Ebercon, 1976). Proline accumulates under drought stress in both drought resistant and drought sensitive cultivars of barley (Hanson et al., 1979). The transformation of the Arabidopsis delta-1-pyroline-5-carboxylate synthetase (the enzyme up-regulated under ABA) into petunia (Yamada et al., 2005) and tobacco (Kishor et al., 1995) was shown to confer drought resistance. In *Arabidopsis thaliana*, the delta-1-pyroline-5-carboxylate synthetase gene was expressed under dehydration but the delta-1-pyroline-5-carboxylate reductase gene was not regulated (Yoshiba et al., 1995). Over-expression of N-acetyl-L-glutamate synthase in Arabidopsis was found to confer drought tolerance (Kalamaki et al., 2009). The dual role of proline as an osmoprotectant and energy source upon recovery may may confer additional drought resistance in particular systems. Application of exogenous proline to increase drought tolerance and yield was tested in *Brassica juncea*, with mixed effects observed (Wani et al., 2016).

Similarly, isoprene polymetrization is also associated with the drought response. It is known that isoprene emissions are related to heat, light, and CO_2 concentrations, and they are the major form of volatile organic carbon compounds released by the ecosystem. Isoprene emission was found to be protective of the photosynthetic apparatus in tobacco (Ryan et al., 2014). Citrulline is an osmoportectant in the leaves and acetylornithine is produced as an intermediate in citrulline synthesis.

4.3.5 Known ABA signaling genes are up-regulated at both 60 minutes and 15 minutes

The ABA synthesis and signaling network is comprehensively discussed by Hauser et al. (2011). They list the transcription factors, kinases, ion channels, and other genes involved in the multiple pathways (Shinozaki and Yamaguchi-Shinozaki, 2006) downstream of ABA. We expect that the expression of some of these genes may also be regulated by ABA in *Brassica napus* guard cells. We therefore tested whether this known ABA network is significantly regulated. The *Brassica napus* orthologs of the ABA signaling *Arabidopsis thaliana* genes were used to test for the enrichment of ABA responsive genes.

We cross-tabulated the number of genes regulated and the membership in the ABA signaling network in Tables 4.10 and 4.12. The statistical significance of the association between membership in the signaling and differential expression was tested and the effect sizes estimated in Tables 4.11 and 4.13.

Table 4.10: Cross-tabulation of membership in the known ABA signaling network against differential expression at 15 minutes of ABA treatment in *Brassica napus*.

	Down-regulated in Brassica (15 min.)	Insignificantly regulated in Brassica (15 min.)	Up-regulated in Brassica (15 min.)
Gene not part of the ABA signaling network	68	100153	385
Gene part of the ABA signaling network	3	410	21

Table 4.11: Tests of statistical significance and the estimated effect size for the association of the known ABA signaling genes and those differentially expressed under 15 minutes of ABA treatment in *Brassica napus*.

Test of association	Statistic	Degrees of freedom	P value	Effect size measure	Effect size
G-test	76.50	2	< 2.22e-16	Contingency coefficient	0.05
Pearson's χ^2 test	238.58	2	< 2.22e-16	Cramer's V	0.05

Table 4.12: Cross-tabulation of membership in the known ABA signaling network against. differential expression at 15 minutes of ABA treatment in *Brassica napus*.

	Down-regulated in Brassica (60 min.)	Insignificantly regulated in Brassica (60 min.)	Up-regulated in Brassica (60 min.)
Gene not part of the ABA signaling network	4293	88857	7456
Gene part of the ABA signaling network	53	258	123

Table 4.13: Tests of statistical significance and the estimated effect size for the association of the known ABA signaling genes and those differentially expressed under 60 minutes of ABA treatment in *Brassica napus*.

Test of association	Statistic	Degrees of freedom	P value	Effect size measure	Effect size
G-test	235.59	2	< 2.22e-16	Contingency coefficient	0.06
Pearson's χ^2 test	356.64	2	< 2.22e-16	Cramer's V	0.06

There is a statistically significant association of the genes in the signaling pathway and those differentially regulated by ABA treatment for both 15 minutes and 60 minutes of treatment. However, the effect sizes are quite small, at 0.049 for 15 minutes and 0.059 for 60 minutes. Only a small core subset of genes in the ABA signaling pathway are transcriptionally regulated in response to ABA. To test whether the number is unexpectedly low, we also calculate the statistical significance and effect size of the enrichment of ABA signaling genes in the differentially expressed genes in *Arabidopsis thaliana*.

Table 4.14: Cross-tabulation of membership in the known ABA signaling network vs. differential expression in *Arabidopsis thaliana* due to ABA treatment.

	Down-regulated in Arabidopsis	Statistically insignificant regulation in Arabidopsis	Up-regulated in Arabidopsis
Gene not part of the ABA signaling network	407	18522	594
Gene part of the ABA signaling network	13	103	21

Table 4.15: Pearson's chi-square (Pearson) and the G (Likelihood ratio) for the independence of the known ABA signaling genes and those differentially expressed under 3 hours of ABA treatment in *Arabidopsis thaliana*.

Test of association	Statistic	Degrees of freedom	P value	Effect size measure	Effect size
G-test	58.64	2	1.8463e-13	Contingency coefficient	0.07
Pearson's χ^2 test	106.13	2	$<$ 2.22e-16	Cramer's V	0.07

The correlation with *Arabidopsis thaliana* differential expression, with a Cramer's V of 0.073, is

only slightly larger than that at 60 minutes for *Brassica napus* (0.059). We thus conclude that the ABA signaling network and its transcriptional regulation is conserved from *Arabidopsis thaliana* to *Brassica napus* and the regulation of the genes in this network is more pronounced at 60 of treatment than at 15 minutes, reflecting the increased transcriptional response with time. Since the signaling nomodifications, protein binding, and transport, with many genes not transcriptionally regulated. Thus the overall small but significant transcriptional effects are consistent with evolutionary conservation of ABA signaling.

4.3.6 Regulatory interactions and the observed differential expression

In this section, we examine the observed differential expression in *Brassica napus* in the light of the known regulatory interaction network in *Arabidopsis thaliana*. We use the AGRIS database, a collection of known transcription factors and their targets from various *Arabidopsis thaliana* studies (Davuluri et al., 2003). We examine the extent of the agreement of our results with various high-throughput and low-throughput studies and what this may mean for the utilization of the regulatory network during the ABA response.

4.3.6.1 ABA-induced differential expression does not employ the whole of the known regulatory interactome from high throughput studies

Any stimulus, including ABA response, may involve only a few direct protein targets. These targets convey the signal through cascades involving protein binding, post translational modifications including phosphorylation and dephosphorylation, and transcriptional factor-DNA binding. The observed changes in mRNA expression of any gene in response to the stimulus involve regulatory interactions between transcription factor proteins and their promoter elements in DNA.

Given mRNA data, we focus primarily on changes to transcription due to transcription factor binding. The qualitative effect of a known regulatory interaction on a target is defined by the type of interaction (activation or repression) and the expression of the transcription factor (up-regulated or down-regulated). For example, if a transcription factor is up-regulated and it is known to activate the expression of its target, we expect to see up-regulation of the target, while if an up-regulated transcription factor is known to repress the expression of its target, we expect to see the target down-regulated. We tabulated the stated type of the interaction vs. the observed expression of the target under ABA treatment. We then performed tests of the observed patterns.

The effects represented by the 2×2 tables were summarized as odds ratios. The odds ratio is defined as the ratio of the fraction of up-regulated genes among those predicted to be up-regulated, divided by the fraction of up-regulated genes among those predicted to be down-regulated. An odds ratio much larger than one implies that our estimate predicts of the observed direction of regulation. When we use all the known interactions from the Agris database, the cross-tabulation does not show this effect, and in fact the opposite is true (with an odds ratio of 0.46).

Table 4.16: Cross-tabulation of the calculated effect of the all the interactions (rows) vs. the actual differential expression (columns) in response to ABA in *Brassica napus*.

	Gene observed up-regulated for 60 mins. ABA treatment	Gene observed down-regulated for 60 mins. ABA treatment
Target gene predicted to be up-regulated based on regulatory interactions	402	170
Target gene predicted to be down-regulated based on regulatory interactions	410	80

This implies that either many of these known interactions are not functional for the ABA response or that some subset of these studies may not be in agreement with our results due to the nature of the experiments that were used for inferring the regulatory interactions.

4.3.6.2 Activation/Repression interactions from low-throughput studies are consistent with the direction of the differential expression

The accuracy and quality of interactome networks derived from high throughput systematic studies and whether they are of similar confidence as individual studies has been debated in literature (Bader et al., 2004; Mrowka et al., 2001). Although the case of protein-protein interactions is more well studied, similar issues may arise in transcription factor studies. For this reason, we re-evaluated the agreement with previously published studies by stratifying according to the study size, with the hypothesis that smaller studies are higher quality. The issue of study size is further explored in Section 4.3.6.4.

Considering regulatory interactions from only the low-throughput studies, fewer than 50 *Arabidopsis thaliana* interactions reported, we see more interactions where the direction of differential expression is consistent with the reported type of regulation.

Table 4.17: Cross-tabulation of the calculated effect of the interaction (rows) vs. the actual differential expression (columns) in response to ABA in *Brassica napus*, using only the interactions derived from low-throughput ($N < 50$) studies.

	Gene observed up-regulated for 60 mins. ABA treatment	Gene observed down-regulated for 60 mins. ABA treatment
Target gene predicted to be up-regulated based on regulatory interactions	103	2
Target gene predicted to be down-regulated based on regulatory interactions	12	8

Table 4.18: Fisher's exact test for the correlation of the theoretical effect of transcriptional regulatory interactions with the actual differential expression of the target.

P value	Alternative hypothesis	odds ratio
3.978e-06 * * *	greater	32.47

P value	Alternative hypothesis	odds ratio

For small studies, the odds ratio of 32.47 is much larger than 1 and highly statistically significant. We conclude that the observed differential expression is highly consistent with regulatory interactions reported by low throughput studies.

4.3.6.3 Differentially expressed regulatory subnetwork

The set of regulatory interactions from low-throughput studies among the differentially expressed genes of *Brassica napus* is visualized in Figure 4-5. Most of the genes represented in this regulatory network are up-regulated, reflecting the observed bias of up-regulation in ABA transcriptional response.

Some transcription factors and targets in Figure 4-5 have confirmed roles in drought stress response. For example, in addition to its role in heat stress, the overexpression of Heat Shock Factor 3 (HSF3) was found to improve drought resistance and water productivity (Bechtold et al., 2013). One target of HSF3 is Galactinol Synthase 1 (GolS1), which is involved in the synthesis of raffinose oligosaccharides and galactinol (Panikulangara et al., 2004), which are osmoprotectants and an antioxidants (Nishizawa et al., 2008). Both of these genes are found to be up-regulated in our experiment.

Signaling pathways for flowering timing (a drought avoidance strategy) may also be shared with signaling for stomatal movement. We see in our experiment that FLO2 is down-regulated in response to ABA. FLO2 is thought to repress the expression of SOC1, which is up-regulated. Both FLO2 and SOC1 are well known for their role in drought escape through early flowering (Riboni et al., 2013), but recently it has been found that SOC1 also affects stomatal opening (Kimura et al., 2015).

We conclude that integrating low-throughput, high-confidence regulatory interactions with gene expression response reveals downstream mechanisms of ABA signaling. Rediscovery of known response components suggests that this analysis can generate new hypotheses to further define ABA signaling pathways.

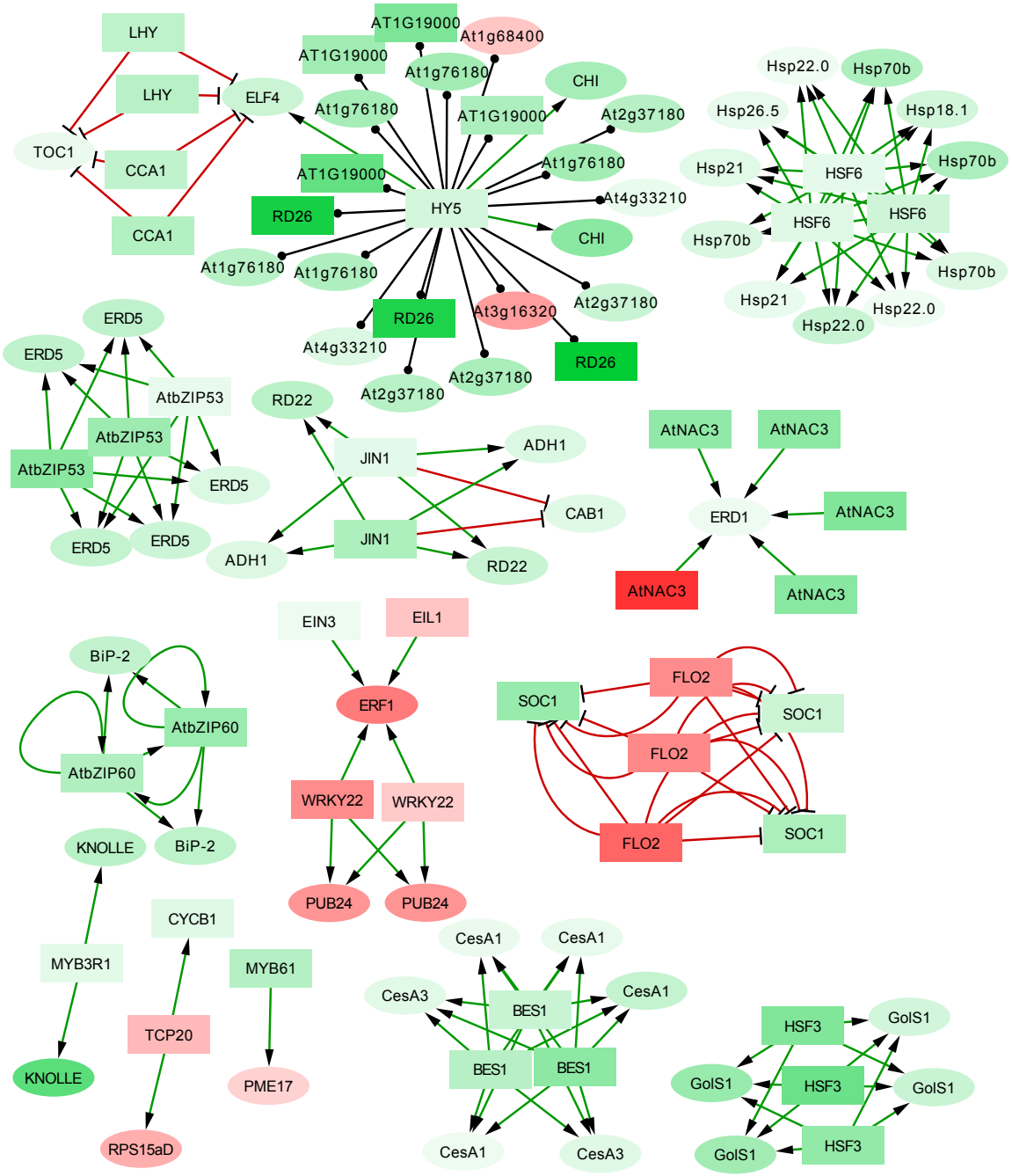

Figure 4-5: Regulatory interactome likely involved in the *Brassica napus* guard cell ABA response. Caption continued on the next page

Figure 4-5: (Continued caption) *Brassica napus* regulatory interactions compiled from the ARGIS low throughput studies ($N < 50$) among differentially expressed genes. Genes are colored by their direction of regulation, with green for up-regulated genes and red for down-regulated genes. The edges and arrow heads ends denote the type of regulatory interactions: activation (green edges with pointed arrows), repression (red edges with flat heads), and unknown (black edges with dotted ends). Since Arabidopsis gene names are more well known, Brassica genes are labelled according to their *Arabidopsis thaliana* orthologs.

4.3.6.4 Agreement with observed differential expression decreases with study size

We have shown above that the regulatory interactions compiled from small studies are consistent with the observed guard cell gene expression profiles in response to ABA. In this section, we explore the high-throughput studies that seem to be inconsistent with our differential expression results. These studies may be of lower confidence or they may be measuring regulatory interactions in conditions that are different from our experiments.

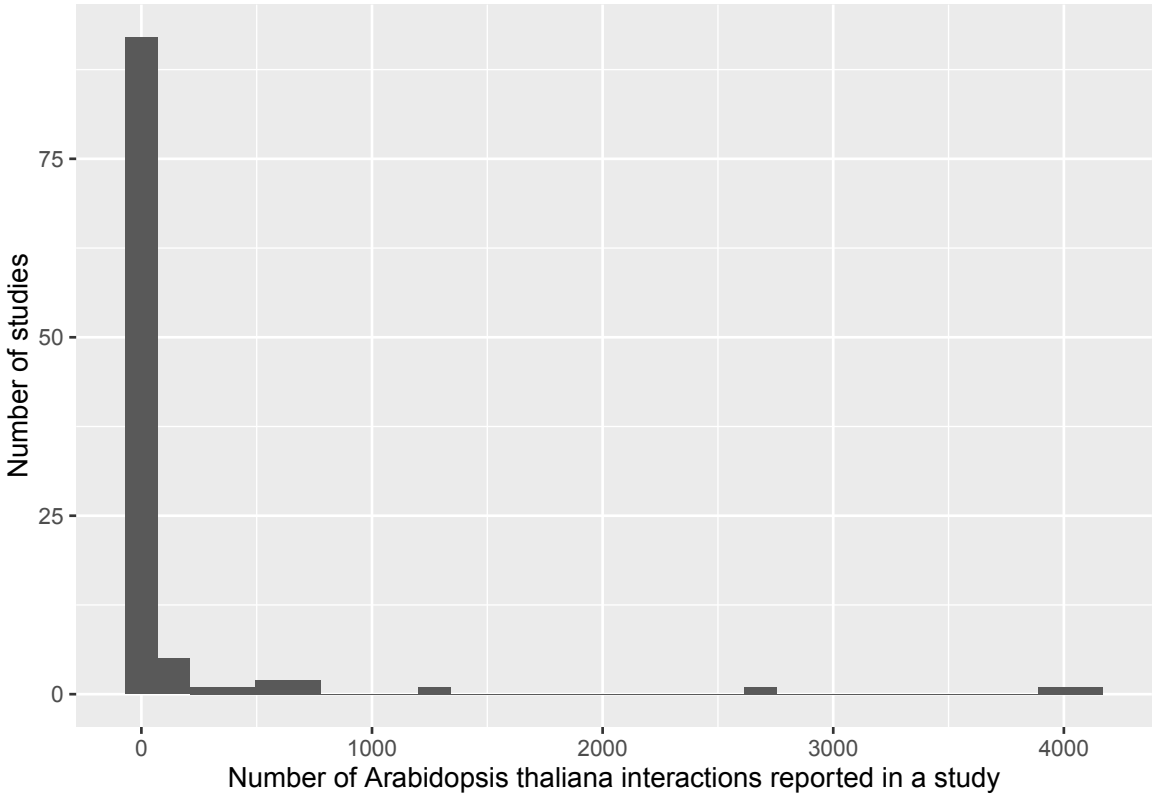

Figure 4-6: Histogram showing the number of *Arabidopsis thaliana* transcriptional regulatory interactions reported per study in the AGRIS database (Davuluri et al., 2003).

First, we provide an exploratory visualization of the numbers of interactions reported in various studies. The histogram in Figure 4-6 shows that most studies report only a few interactions, with a few studies reporting hundreds to thousands of interactions. We use a simple metric, the fraction of edges consistent with our differential expression data, to characterize each study. An edge is considered consistent if the direction (i.e., up or down) of expression of the target predicted by the transcription factor's (up or down) regulation and interaction type (repression or activation) is the same as the observed differential expression of the target. This metric does not account for the class imbalance of up/down regulation or the activation/repression edges. Figure 4-7 shows

that many low-throughput studies are consistent with the observed differential expression, and the fraction consistent varies across a wide range. The 4 high throughput studies with a large number of signed interaction edges show about 0.3 to 0.6 fraction consistency with our differential expression observations.

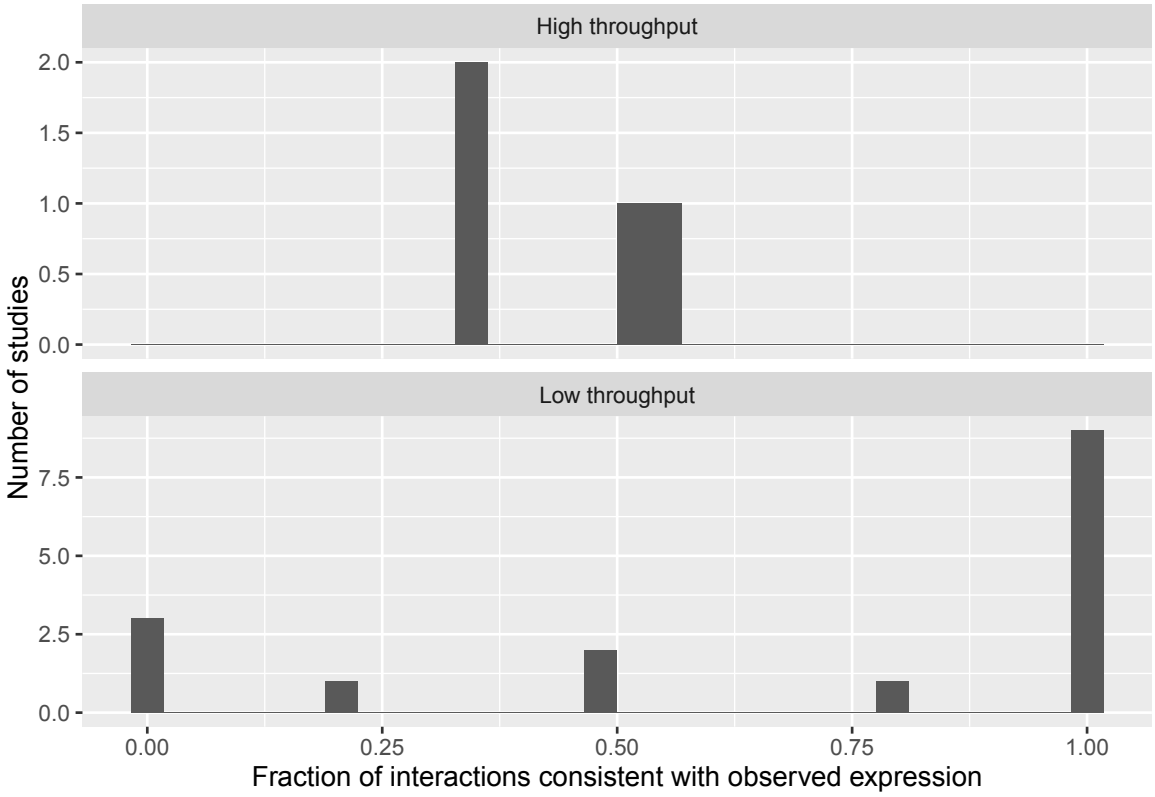

Figure 4-7: Histogram for the fraction of *Brassica napus* interaction edges that are consistent with our observed ABA differential expression per study. The panels separately show the histograms for studies with less than 50 reported interactions (labelled as low throughput) and for studies reporting 50 or more interactions (labelled as high throughput).

Since the number of up-regulated and down-regulated genes are unbalanced, the effect of individual studies on the statistical significance of the directional prediction of the targets' differential expression is best evaluated by the Fisher's exact test rather than the fraction of consistent edges. Figure 4-8 shows the effect of increasing data by including larger studies vs. the effect of potentially less accurate or consistent large studies.

We have only evaluated the consistency of the interactions with the differential expression in one condition, and it is possible that activation vs. repression could vary in other conditions. Nevertheless, the lower agreement with larger study sizes suggests that high-throughput studies of transcriptional regulation are less accurate than low-throughput studies.

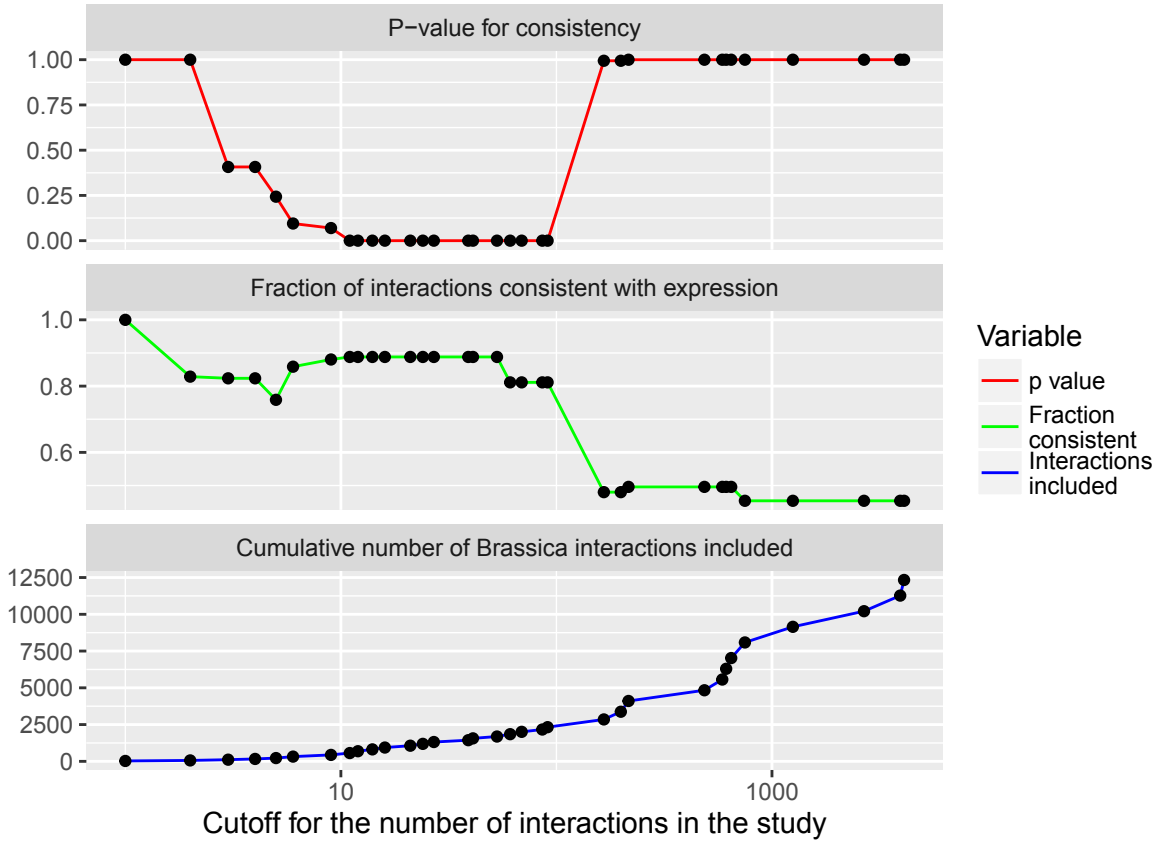

Figure 4-8: The effect of including larger studies on the consistency with the observed gene expression profiles. The x-axis denotes the cutoff for the study size, and for each point all studies larger than this size are disregarded. For each cutoff, we plot the fraction of interactions that are consistent with the observed gene expression, the statistical significance (p value) of this consistency calculated using Fisher's exact test, and the cumulative number of translated *Brassica napus* interactions between differentially expressed genes added. An *Arabidopsis thaliana* interaction from the AGRIS database may map to multiple *Brassica napus* interactions due to Brassica polyploidy; or it may result in zero *Brassica napus* interactions being included in the study if either the factor or target are not differentially expressed).

4.3.6.5 Differentially expressed targets of known ABA signaling transcription factors suggest candidates for a role in ABA signaling

Table 4.19: Regulatory interactions with known ABA transcription factors among differentially expressed genes.

Common name of the transcription factor	Common name of the target	*B. napus* transcription factors	*B. napus* targets	Interaction Type	Is the target a known member of the ABA signaling pathway?
AtMYC2	ADH1	BnaC08g07580D, BnaA05g18020D	BnaC06g37860D, BnaA07g33310D	Activation	No
AtMYC2	RD22	BnaC08g07580D, BnaA05g18020D	BnaC07g29150D, BnaA06g39340D	Activation	No
AtMYC2	CAB1	BnaC08g07580D, BnaA05g18020D	BnaC03g59520D	Repression	No
HY5	At1g76180	BnaA10g21200D	BnaC06g36880D, BnaA07g21490D, BnaA07g32420D, BnaA02g36030D, BnaC06g21970D	Unknown	No
HY5	At2g37180	BnaA10g21200D	BnaC04g08090D, BnaA05g07290D, BnaC04g08100D, BnaA05g07300D	Unknown	No
HY5	At4g33210	BnaA10g21200D	BnaA01g03630D, BnaC01g04970D	Unknown	No

Common name of the transcription factor	Common name of the target	*B. napus* transcription factors	*B. napus* targets	Interaction Type	Is the target a known member of the ABA signaling pathway?
HY5	At1g19000	BnaA10g21200D	BnaC08g18480D, BnaA08g22290D, BnaC08g36920D, BnaA09g44370D	Unknown	No
HY5	RD26	BnaA10g21200D	BnaA01g16400D, BnaC07g40860D, BnaA03g48570D	Unknown	Yes
HY5	At3g16320	BnaA10g21200D	BnaC05g37040D	Unknown	No
HY5	CHI	BnaA10g21200D	BnaA09g34840D, BnaC08g26010D	Activation	No
HY5	ELF4	BnaA10g21200D	BnaA05g05560D	Activation	No

Not all regulatory interactions may actually be functional in all physiological conditions, and not all of them may be involved in any particular response. If a transcription factor is known to be involved in ABA signaling, this is prior information that leads to a higher confidence that its regulatory interactions are likely to be functional in the ABA response. Therefore, we attempt to use our knowledge of the important drivers of the ABA response to predict other targets that may in fact be regulated using the known mechanisms and have an important role in the response to ABA by guard cells.

Within the high-confidence (i.e., low-throughput) set of regulatory interactions, we select a subset where the transcription factors are known members of the guard cell ABA signaling pathway and where both the factor and target are differentially expressed. These selected transcription factors, their targets, and interactions are listed in Table 4.19. This table suggests candidate genes among

the targets that are likely to be involved in ABA signaling. Among the targets identified, RD22 and RD26 are named for their response to desiccation and have known roles for their response to water deprivation. ADH1 is known to be regulated by both dehydration and hypoxia. CAB1 is a light-responsive protein involved in photosynthesis. CHI is involved in the response to UV light. These results suggest shared elements between the various stress response pathways and a role for ABA in the regulation of these pathways. ELF4 is an early flowering gene. AT1G76180 (ERD14) is a Ca2+ binding protein involved in the early response to dehydration and cold.

4.3.7 Regulation inferred from cis-acting transcription factor binding sites

A transcription factor often regulates the expression of its target by binding to a cis-regulatory element in the target gene's promoter region or a more distant enhancer. A transcription factor or a class of transcription factors often binds in a sequence-specific manner, recognizing a specific polynucleotide sequence. The existence of these binding sites in the promoter regions of *Brassica napus* genes permit inference of putative regulatory interactions between *Brassica napus* transcription factors and their targets. These provide testable hypotheses of the regulatory network and refine the networks generated by mapping *Arabidopsis thaliana* interactions to all the corresponding *Brassica napus* orthologs, which we used in the previous section.

4.3.7.1 Promoters of differentially expressed genes are enriched for the binding sites of putative causal transcription factors

To search for possible transcription factors responsible for ABA gene expression response, we identified known binding sites in the putative promoter regions of *Brassica napus* genes. For each known Arabidopsis transcription factor binding site sequence in our set, we calculated the enrichment of the occurrence of the sequence in the promoters of the up-regulated or down-regulated genes versus all other genes as background, with statistical significance measured by the hypergeometric p-value corrected for multiple testing.

Table 4.20: Binding site sequences over-represented in the putative promoter regions of genes up-regulated by ABA, along with the corresponding transcription factors.

Binding site sequence (BSS)	Number of promoter regions of up-regulated genes with BSS (out of 7669)	Total number of promoter regions with BSS (out of 101040)	Adjusted pval	*Arabidopsis thaliana* transcription factor(s) that bind to the BSS	Orthologous *Brassica napus* transcription factors	Is the transcription factor part of the ABA signaling network?
ABRE binding site motif	2625	19775	2.897e-216	ABF4 (AT3G19290)	BnaCnng41320D, BnaA01g26200D, BnaC01g43800D, BnaA03g35190D, BnaA05g20870D, BnaC05g33570D	Yes
ABFs binding site motif	2003	13791	6.162e-200	ABF1 (AT1G49720), ABF2 (AT1G45249)	BnaC06g02640D, BnaA06g03040D, BnaC06g00420D, BnaA10g28780D	Yes, Yes
ABRE-like binding site motif	6442	71043	4.318e-181	NA (NOTAVAILABLE)	NA	No
CBF2 binding site motif	960	6780	8.816e-81	ATCBF2 (AT4G25470)	BnaA08g30910D	No

Binding site sequence (BSS)	Number of promoter regions of up-regulated genes with BSS (out of 7669)	Total number of promoter regions with BSS (out of 101040)	Adjusted pval	*Arabidopsis thaliana* transcription factor(s) that bind to the BSS	Orthologous *Brassica napus* transcription factors	Is the transcription factor part of the ABA signaling network?
GBF1/2/3 BS in ADH1	960	6780	8.816e-81	AtGBF1 (AT4G36730), ATBZIP54 (AT4G01120), GBF3 (AT2G46270)	BnaC03g61840D, BnaC01g02130D, BnaA01g01100D, BnaA08g15400D, BnaC-nng01910D, BnaA09g00170D, BnaC04g01070D, BnaC03g25660D, BnaA05g01520D	No, No, No
ERF1 BS in AtCHI-B	5106	57711	5.664e-68	ATERF1 (AT3G23240)	BnaA07g06760D, BnaC07g08360D, BnaA01g23940D	No
TGA1 binding site motif	1633	15985	4.986e-38	TGA1 (AT5G65210)	BnaC09g06840D, BnaAnng04720D, BnaC02g43620D, BnaA09g07120D, BnaA06g24140D, BnaC03g49070D	No
RAV1-B binding site motif	6168	77449	3.646e-15	AtRAV2 (AT1G68840)	BnaC02g18650D, BnaA02g14040D	No

Binding site sequence (BSS)	Number of promoter regions of up-regulated genes with BSS (out of 7669)	Total number of promoter regions with BSS (out of 101040)	Adjusted pval	*Arabidopsis thaliana* transcription factor(s) that bind to the BSS	Orthologous *Brassica napus* transcription factors	Is the transcription factor part of the ABA signaling network?
HSEs binding site motif	2281	28000	0.001116	AT-HSFC1 (AT3G24520)	BnaC07g07130D, BnaA03g37460D, BnaC03g43990D, BnaA07g05580D	No
DREB1&2 BS in rd29a	204	2036	0.001853	ATCBF2 (AT4G25470)	BnaA08g30910D	No
AtMYB2 BS in RD22	4689	59738	0.004914	ATMYB2 (AT2G47190)	BnaA05g00710D, BnaC04g51450D	Yes
E2F-varient binding site motif	211	2182	0.01041	NA (AT2G36011)	NA	No
VOZ binding site	155	1541	0.01253	ATVOZ1 (AT1G28520)	BnaC05g21930D, BnaC03g58740D, BnaA08g18270D, BnaA09g27210D	No
ARF binding site motif	7433	97264	0.03437	ARF1 (AT1G59750)	BnaC01g28340D, BnaA01g35830D	No

Table 4.21: Binding site sequences over-represented in the putative promoter regions of genes up-regulated by ABA, along with the corresponding transcription factors.

Binding site sequence (BSS)	Number of promoter regions of down-regulated genes with BSS (out of 4547)	Total number of promoter regions with BSS (out of 101040)	Adjusted pval	*Arabidopsis thaliana* transcription factor(s) that bind to the BSS	Orthologous *Brassica napus* transcription factors	Is the transcription factor part of the ABA signaling network?
ERF1 BS in AtCHI-B	2766	57711	5.794e-06	ATERF1 (AT3G23240)	BnaA07g06760D, BnaC07g08360D, BnaA01g23940D	No

These binding sites are also enriched in the promoter regions of *Arabidopsis thaliana* guard cell ABA responsive genes (Wang et al., 2011). Specifically, while the ABRE targets are regulated in an ABA-dependent manner, the DREB targets are also up-regulated during drought and cold stress in an ABA-independent manner (Narusaka et al., 2003; Agarwal and Jha, 2010).

4.3.7.1.1 ABA and drought-related roles for implicated transcription factors

We predict that the transcription factors implicated by enrichment of their binding sites function in ABA and drought responses. CBF2/DREB1C disrupted mutants have a higher tolerance for drought and cold stresses (Novillo et al., 2004). GBF1/2/3 genes are associated with light response, and their shared binding motif is also enriched in drought response genes in Arabidopsis (Huang et al., 2008), suggesting cross-talk between stress and circadian clock pathways. ERF1-overexpressing Arabidopsis plants have increased resistance to drought and salt stress (Cheng et al., 2013). ERF1 also functions in pathogen response pathways that are mediated by ethylene and jasmonic acid signaling. TGA1 was found to be up-regulated in response to drought in Arabidopsis (Ding et al., 2013). RAV1 is

known to down-regulate the expression of ABA related genes ABI3, ABI4, and ABI5 (Feng et al., 2014). Plants overexpressing MYB2 have greater sensitivity to ABA (Abe et al., 2002), and its expression is regulated by dehydration (Urao et al., 1993) and ABA (Xin et al., 2005). VOZ1 and VOZ2 are implicated in flowering time regulation (Yasui and Kohchi, 2014) (a drought escape strategy) by repressing FLC and their double mutant increased drought tolerance (Nakai et al., 2013).

4.3.7.1.2 Transcription factors regulating the up-regulated genes are themselves up-regulated

There are 170 distinct *Brassica napus* transcription factors whose binding sites are enriched in the promoter regions of up-regulated genes. Out of these 170 genes, 29 genes are themselves up-regulated as opposed to 1944 out of the total 101040 genes, giving a statistically significant hypergeometric p-value of $4.11e-20$.

4.3.7.1.3 Transcription factors regulating the up-regulated genes are known members of the ABA signaling network

Out of the 170 *Brassica napus* transcription factors whose binding sites are enriched in the promoter regions of up-regulated genes, 23 are known members of the translated ABA signaling pathway. Out of the total of 101040 genes, 435 are members of the ABA pathway. This leads to a statistically significant hypergeometric p-value of $4.86e-29$.

4.3.7.2 Loss/gain of specific binding sites after the genome duplication events affects differential expression

Brassica napus is the result of genome triplication, fractionation, and polyploid hydrid speciation events since its ancestor and the *Arabidopsis thaliana* lineage diverged (Cheng et al., 2014). As discussed above, binding site sequences of known transcription factor families are enriched in the promoter regions of differentially expressed genes. The same sequence motifs are also present in the promoter regions of the corresponding *Arabidopsis thaliana* orthologs. The actual mechanism of regulation of gene expression in response to drought and ABA probably also involves other cis and trans acting regulatory elements in addition to the transcription factor binding sites that we have found. This mechanism is present in *Arabidopsis thaliana*, and the enrichment of certain sequence motifs that we have found could simply be a carryover from an earlier mechanism that is no longer dynamically evolving.

Since the genome triplication event, some of these known binding sites in the gene promoter regions would have been lost due to mutations. We investigate whether the regulation of expression of *Brassica napus* genes is affected by the continuing loss or gain of these binding sites. If gene expression is indeed related to the relatively recent gain or loss of these binding sites, then this is additional evidence (like a mutation experiment might provide) that these enriched sites are a causal link in the regulation of gene expression. Also, this is evidence that drought response has been evolving in some fashion in the *Brassica* genus.

For each binding site sequence motif, we construct a fixed effects generalized linear model that relates the observed log2 fold change in expression of a gene to the number of copies of the binding site sequence found in the putative promoter region of the gene. We model *Arabidopsis thaliana* orthologs of the genes as fixed effect categorical variables. By controlling for the *Arabidopsis thaliana* ortholog, we have controlled for the variance explained within gene families due to a common ancestral gene. The remaining variance that our method models is only due to the gain or loss of binding site sequences since the genes copies diverged.

Table 4.22: Binding site sequences that influence the difference between the regulation (i.e., fold change in response to ABA) of individual *Brassica napus* genes coming from the *Arabidopsis thaliana* ortholog. Only those sequences that have a p-value of less than 0.05 are shown here, based on a T test of the coefficient representing the binding site sequence in a linear model predicting the fold change.

	Pr(>|t|)	Adjusted p val.
ABRE-like binding site motif	1.595e-15	8.456e-14
CBF2 binding site motif	7.261e-06	0.0001283
GBF1/2/3 BS in ADH1	7.261e-06	0.0001283
ERF1 BS in AtCHI-B	0.0005842	0.00774
ABRE binding site motif	0.003044	0.03226
MYB3 binding site motif	0.004178	0.03691
DPBF1&2 binding site motif	0.008482	0.06422
HSEs binding site motif	0.02602	0.1697
ABFs binding site motif	0.03116	0.1697
RAV1-B binding site motif	0.03202	0.1697

Table 4.22 shows the result of our fixed effects regression. The statistical significance of the number of copies of a specific binding site sequence in the promoter regions as a predictor is estimated by the p-value of including its coefficient in the model. Since we tested a total of 53 sequence motifs, we apply the Benjamini-Hochberg multiple testing correction to obtain an adjusted p-value (false discovery rate) shown in the model. We show the binding site sequences with the smallest p-values in the table.

4.3.8 Most genes are under negative selection

In the previous section we investigated how the evolution of ABA transcriptional response is related to changes in the binding site sequences in the promoter region. Here, we examine DNA mutation rates between *Arabidopsis* and *Brassica napus*.

Point mutations in the protein coding regions of the genomes can lead to substitutions of one nucleotide in place of another; we do not consider insertions or deletions here. Single-nucleotide changes in the DNA sequence that change the amino-acid sequence of the product, rather than introducing a stop codon, are termed non-synonymous substitutions; if the amino acid sequence remains the same (due to codon degeneracy), the change is termed a synonymous substitution. The substitution rate is defined as the number of times each nucleotide position in a sequence has undergone a substitution. Since a new subsititution will mask the history of the nucleotides that were once present at that position, the substitution rates between two homologous genes must be estimated rather than just counted as the fraction of mismatches. For the purpose of our analysis, we calculated the nucleotide mismatch rates only in the regions of protein sequence alignment. The regions of indels and unaligned regions due to early stopping were ignored. We have used the method of Li (1993) for calculating the nucleotide substitution rates from the nucleotide mismatches.

The synonymous (Ks) and non-synonymous (Ka) nucleotide substitution rates are calculated for all *Brassica napus* genes based on their corresponding *Arabidopsis thaliana* orthologs. Although we do not have access to the common ancestor of *Brassica napus* and *Arabidopsis thaliana*, we use this interspecies comparison as a proxy for the substitution rates since the *Arabidopsis*-*Brassica* split for simplicity. Figure 4-9 shows the ratio of the estimated non-synonymous and synonymous substitution rates for all genes. If there is a fitness loss for mutations of the protein sequence, we expect the non-synonymous substitutions to be less than the synonymous substitutions, which are more free to accumulate, since their effect on fitness should be less pronounced in general. Accordingly, K_a/K_s is less than 1 for most genes. Thus, most of the genes in *Brassica napus* are still under negative selection, with pressure to retain their ancestral amino acid sequence, despite the high polyploidy. Possible explanations are that the plant uses the multiple gene copies to finely control gene expression; or the different gene copies may have evolved to fulfill a more specialized function. The ancestral genomes, *Brassica rapa* and *Brassica oleracea*, have both experienced gene loss, and genes not under evolutionary pressure to conserve function may be lost as evolution continues. Since the speciation of *Brassica napus*, time may not yet have been sufficient to show larger rates of non-synonymous

substitutions for genes that are not under negative selection.

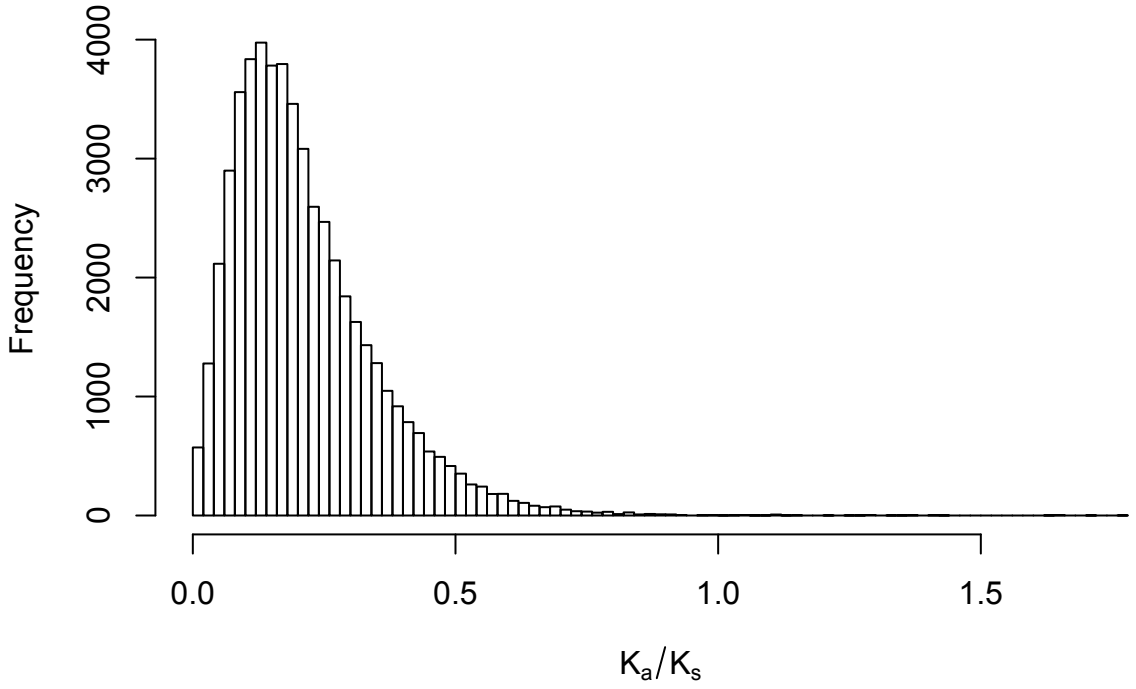

Figure 4-9: The distribution of the ratio of non-synonymous to synonymous nucleotide substitution rates (K_a/K_s) of *Brassica napus* genes from the corresponding *Arabidopsis thaliana* genes.

The mean synonymous substitution rate K_s of 0.4797091 is consistent with estimates that the Arabidopsis-Brassica split occurred 10-20 million years ago.

The calculation of the point substitution rates (k_a and k_s), and therefore the measure of conservation pressure on the protein sequence (k_a/k_s) according to our definitions depend only on the aligned portions of the amino acid sequence. We also counted the total number of amino acids that were part of insertions and deletions (including those due to late start codons or early stop codons). Since the indel mechanisms in the nucleotide sequence are different from the mechanisms of point mutations, and the effect of amino acid indels is different from the effect of amino acid point mutations, we used a separate measure for indels. Table 4.23 shows that the total length of amino acid indels is correlated with the point mutation rates, especially with the non-synonymous mutation rate (k_a).

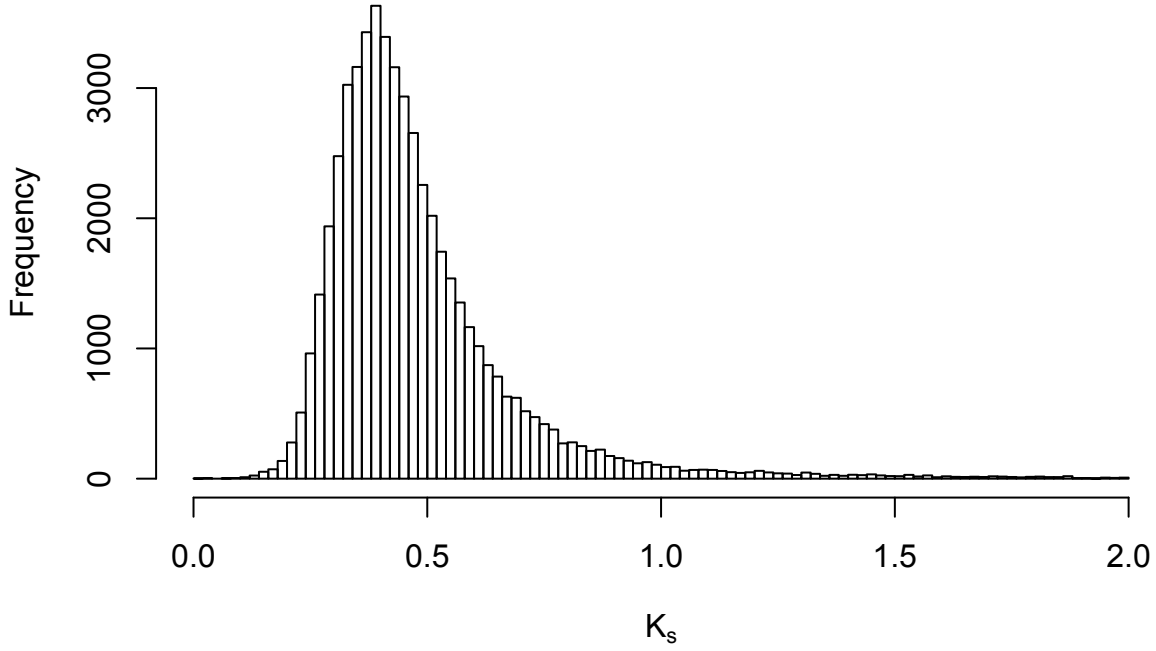

Figure 4-10: The distribution of the synonymous nucleotide substitution rate (K_s) of *Brassica napus* genes from the corresponding *Arabidopsis thaliana* genes.

Table 4.23: The correlation between the synonymous subsitution rate k_s, the non-synonymous substitution rate k_a, amino acid convervationary pressure k_a/k_s, and the number of amino acids in indels calculated between corresponding *Brassica napus* and *Arabidopsis thaliana* orthologs. All correlations are evaluated as Kendall's τ, and p-values correspond to rejecting the null hypothesis of no correlation ($\tau = 0$). Each cell represents the correlation between the quantities described in the corresponding row and column names.

	k_a	k_s	k_a/k_s
k_s	τ=0.26, p-value<2.2e-16		
k_a/k_s	τ=0.71, p-value<2.2e-16	τ=-0.04, p-value=1.31e-36	
Amino-acids in indels	τ=0.31, p-value<2.2e-16	τ=0.13, p-value<2.2e-16	τ=0.27, p-value<2.2e-16

4.3.9 Evolution of differential expression is related more to changes in the nucleotide sequence than the amino acid sequence

We explored the mechanism of the evolution of the regulatory response to ABA in *Brassica napus*. Specifically, we are interested in genes whose transcriptional response has changed since the Brassica-Arabidopsis split. We compare homologous gene families and calculate the variance in the fold changes of the individual members within gene families. Gene families with higher standard deviation in their fold changes must have members that diverged from their ancestor. We investigated whether this divergence is related to the synonymous nucleotide substitution rates (K_s), non-synonymous substitution rates (K_a), or the degree of amino acid conservation pressure (K_a/K_s).

For each gene family, we calculated the standard deviation in the $\log 2$ fold changes and fit linear models to the means of K_s, K_a, and K_a/K_s.

Table 4.24: Linear regression model showing the relation between differential expression divergence and nucleotide mutation rates. Parameters for the model $\sigma_{\log 2(\text{Fold Change})} \sim \mu_{k_a} + \mu_{k_s} + \mu_{k_a/k_s} + \mu_{extindels}$ i.e., the standard deviation of the log2 fold change of a gene family predicted from the mean nucleotide substitution rates (k_s, k_a, and their ratio k_a/k_s) and the number of amino acids in indels between members of the *Brassica napus* paralogous family and the corresponding *Arabidopsis thaliana* ortholog. The statistical significance is given by the p-value of the t-test.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1409	0.007377	19.1	1.732e-80
μ_{k_a}	-0.1278	0.04994	-2.56	0.01047
μ_{k_s}	0.0833	0.01441	5.782	7.526e-09
μ_{k_a/k_s}	0.07432	0.02636	2.82	0.004813
$\mu_{extindels}$	-5.618e-06	3.855e-06	-1.457	0.1451

Using the t-statistic as a test statistic for each coefficient, we observe that K_s is the most significant predictor. The divergence of the nucleotide sequence is therefore more strongly associated with the divergence of the ABA-responsive regulation of gene expression than the amino acid sequence.

Dependence on nucleotide rather than amino changes could be due to changes in histone affinities, the binding of other proteins to the coding region of genes, or due to the association of intragenic nucleotide mutation rates with the promoter region mutation rates.

4.4 Conclusions

This study has used experimental measurement of Brassica guard cell genes that are transcriptionally regulated by ABA. In addition to statistical analysis of the gene expression data, we have combined the expression data with known information about metabolic pathways, gene regulatory interactions involving transcription factors and DNA regulatory elements, and evolutionary comparisons to Arabidopsis. Our analyses are consistent with existing knowledge and have suggested new components of ABA signaling pathways. Here, we briefly summarise the conclusions from this analysis.

We have generally found qualitatively similar gene expression response at 15 and 60 minutes of ABA application. The extent of regulation generally increases from 15 minutes to 60 minutes, consistent with a mechanism in which the transcription rate is constant with negligible degradation. Only a few genes show statistically signficant regulation with a different dynamic pattern. These early response genes were identified and listed in Table 4.2.

Comparisons with Arabidopsis show considerable divergence of gene expression in paralogous gene families, but the level of correlation within families is still high. While statistically correlated, we see many differences between the measured guard cell ABA response in *Brassica napus* and *Arabidopsis thaliana*. Despite the low statistical power, and hence a smaller number of genes identified at 15 minutes of ABA treatment, these genes showed a much higher concordance with the *Arabidiopsis* response, with only 4 genes showing opposite direction. These 4 genes were up-regulated in *Brassica napus* genes, while their *Arabidopsis* orthologs were down-regulated; roles for these genes were discussed.

Among the metabolic pathways, proline synthesis was found to be up-regulated, consistent with published studies. A statistically significant part of the ABA signaling pathway is up-regulated, but most genes in the pathway do not change their expression, similar for both *Arabidopsis* and *Brassica*. We found that regulatory interactions reported in individual small-scale in *Arabidopsis* were more consistent with the observed *Brassica napus* gene expression profiles. We selected these interactions to generate the regulatory interaction graph shown in Figure 4-5. The regulatory interactions included in this figure are likely to be actively involved with the ABA response. We found certain transcription factor binding sites whose occurence in proximity to the translation start sites of *Brassica napus* genes was enriched for the regulated genes. Most of the transcription factors known to bind to these purported cis-acting regulatory elements were also regulated, and some of them are well known for their role in ABA signaling.

Finally, we showed that these non-coding DNA regulatory elements have diverged within paralogous families, and evolutionary divergence has affected the expression of their target genes. This is evidence of the functional evolution of the drought response facilitated due to the gene copies present in polyploid genomes.

4.5 Materials and Methods

4.5.1 Plant Material and Growth Conditions

All *Brassica napus* plants used in this study were from the double-haploid line DH12075. Brassica seeds were sown on Sunshine Redi-earth Plug & Seedling Mix (Sun Gro Horticulture, Canada) and then stratified for at least 2 d at 4°C. The plants were grown at 60% relative humidity in 16 h light at 21°C and in 8 h dark at 18°C.

4.5.2 Isolation of Guard Cell Protoplasts

Brassica leaves (~70g) 5-7 weeks old were excised and their central veins removed before blending for 3 × 1 min with a Waring blender in cold water. After filtering through a nylon mesh (pore size 200 µm), the epidermal fragments were washed thoroughly with water and transferred to a flask containing 100 mL of 0.7% Cellulase R-10 (Yakult Pharmaceutical, Tokyo, Japan), 0.05% Macerozyme R-10 (Yakult), 0.10% polyvinylpyrrolidone 40, 0.25% BSA, 0.5 mM ascorbic acid, and 55% basic medium (0.5 mM CaCl2, 0.5 mM MgCl2, 5 mM MES hydrate, 0.5 mM ascorbic acid, 10 µM KH2PO4, 0.53 M D-sorbitol, pH 5.5). The epidermal peels were incubated in a shaking water bath (175 RPM) at 22°C for 40-50 min in the dark to digest all epidermal and mesophyll cells. To adjust the osmolality in preparation for the second enzyme digestion, 150 mL of basic medium were added, and the epidermal peels were incubated for an additional 10 min prior to being collected using a nylon mesh (pore size 200 m) and washed two times with basic medium. The epidermal fragments were then transferred into a flask containing 50 mL of 1.1% Cellulase RS (Yakult), 0.0075% Pectolyase Y-23 (Duchefa Biochemie, Haarlem, Netherlands), 0.25% BSA, 0.5 mM ascorbic acid, and 100% basic medium. After incubating in a shaking water bath (100 RPM) at 22°C for 1-1.5 h in the dark, the solution containing free guard cell protoplasts was filtered through a single layer of nylon mesh (pore size 20 µm). Basic medium was also poured through the mesh to rinse the epidermal peels for a total volume of 400 mL. The protoplast solution was centrifuged at 350g for 5 min, after which the supernatant was removed. The pellet was re-suspended in a small volume of basic medium and then layered carefully on top of an equal volume of gradient solution containing 35% basic medium and 65% Histopaque (Sigma-Aldrich, St. Louis, MO, USA). Following centrifugation at 430g for 5 min, the guard cell protoplasts at the interface of the two solutions were isolated. Guard cell number and purity were determined using a hemacytometer. Protoplasts with

a purity of ~99% were used for subsequent experiments.

4.5.3 Statistical tests

4.5.3.1 Tests of association of categorical variables

We have tested contingency tables for dependency between counts of categorical variables like sets of differentially expressed genes and membership in signaling networks. The hypothesis tested was generally that membership of a gene in one classification is associated with its membership in another classification. The null hypothesis in these tests is that categorical classifications are independent, in which case the probability in each cell of the table is simply a product of the marginals.

For the 2×2 tables, the distribution of the table entries under the null hypothesis is the hypergeometric distribution. We therefore used Fisher's exact test (Fisher, 1922). The effect size in 2×2 tables can be calculated as the odds ratio, which is simply the ratio of the product of diagonal terms divided by the product of the off-diagonal terms.

For 2-way tables larger than 2x2, where at least one of the classifications involves three or more categories, we used continuous approximations to the discrete counts. The two statistics commonly used here are the Pearson's χ^2 statistic given by $\sum_i \frac{(O_i - E_i)^2}{E_i}$ and the G statistic given by $2 \sum_i O_i \ln(\frac{O_i}{E_i})$.

The G statistic can be thought of as the mutual information contained in the counts of the contingency table given the marginals. This is also the Kullback-Leibler divergence between the observed distribution and that expected assuming independence of the two classifications. The Pearson's χ^2 statistic measures the variance from the expected counts, but it can also be derived from the saddle node approximation of the G-statistic (Hoey, 2012). We report both tests in the results. While the G-test has favorable theoretical properties, the χ^2 is a more familiar and commonly used statistical test. Both tests gave similar results for our data. We used the vcd R package (Meyer et al., 2016) for calculating statistical significance for both tests.

The strength of association is measured using Cramer's V, which is calculated as $V = \sqrt{\frac{\chi^2}{n(k-1)}}$, where n is the total sample size (i.e., the sum of all entries in the contingency table), and k is the smaller of the number of rows and columns (Cramer, 1947).

4.5.3.2 Enrichment of metabolic pathways

Metabolic pathways were downloaded from BioCyc/PlantCyc and translated from *Arabidopsis thaliana* to their *Brassica napus* orthologs.

For each gene, we used the p-value of differential expression to calculate an equivalent z-score. The z-score is calculated as the value of a standard normal distribution that gives the p-value for that gene for a two-tailed test. The z-scores of the pathway genes were compared with those in the complementary set for a difference of means using a t-test.

However, the t-test may give significant p-values smply because the genes in the pathway are correlated. To account for this, we permuted the samples and calculate p-values for each gene set (pathway) for each permution. The best (i.e., smallest) p-value for each permutation among all the gene sets were tabulated. The median of the best p-values for among all permutations was used as the cutoff for calling significantly enriched pathways. Since this is the best p-value acheived in the case of randomly permuted samples, this is the cutoff for the family wise error rate.

4.5.4 Analysis of differential expression

All the RNA-seq reads (100 bp single ended) were aligned to the *Brassica napus* genome of the Darmor line (Chalhoub et al., 2014) with TopHat2 (Kim et al., 2013). Reads mapping to genes were counted with HTSeq (Anders et al., 2015). Multi-mapped reads were discarded by HTSeq due to low mapping quality. This lowers the statistical power to detect differential expression for genes with many close paralogs because sequencing reads may align to different paralogs. However, removing these ambiguously mapped reads means that we are confident that we are correctly distinguishing the different paralogs, and differentially expressed genes are stringently called.

We were able to map about 100 million reads for each sample. There were 101040 annotated gene models in the genome, out of which we had at least one read uniquely mapped to 78105 genes. It is possible that the other genes are not expressed in the guard cells of our line, or their expression is lost due to polyploidy.

The differential expression values was calculated with DESeq2 (Love et al., 2014).

Batch effect on replicates were noted and accounted for by including the replicate information in a

linear model design matrix. The expected expression level of a gene i in sample j was modeled as

$$q_{ij} = s_j \sum_r x_{jr}\beta_{ri},$$

where s_j is the sample normalization, x_{jr} is the effect r on sample j and β_{ir} is the effect r on gene i. We arranged our samples block-wise with time as { t=0 (Replicate 1), t=0 (Replicate 2), t=0 (Replicate 3), t=15 (Replicate 1), t=15 (Replicate 2), t=15 (Replicate 3), t=60 (Replicate 1), t=60 (Replicate 2), t=60 (Replicate 3), }. The design matrix x had the form

$$x = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

where the first 3 rows of the design matrix correspond to the effect of the 3 time points, while the latter 3 rows correspond to the effect of the 3 replicates. This design matrix separates the effects of the conditions and replicates that we observe in Figure 4-11.

The actual read counts K_{ij} for gene i and sample j are assumed to be sampled from a negative binomial distribution with the expected expressions q_{ij} and a gene-specific dispersion α_i. The probability mass function of the read counts is given by

$$\Pr(K_{ij} = k) = \frac{\Gamma(k + 1/\alpha_i)}{k!\Gamma(1/\alpha_i)} \left(\frac{q_{ij}}{q_{ij} + 1/\alpha_i}\right)^k \left(\frac{1}{1 + q_{ij}\alpha_i}\right)^{1/\alpha_i}.$$

The dispersion for a gene α_i is a shrinkage estimate based on all the observed genes, which decreases with increasing mean read counts.

The p-values for differential expression were obtained by the Wald's test (Wald, 1943). Genome wide significance was evaluated by adjusting for multiple testing using the Benjamini Hochberg correction (Benjamini and Hochberg, 1995) and an independent filtering step based on the mean expression level. An adjusted p-value cutoff of 0.05 (FDR) was used to call a gene differentially expressed in all downstream analyses.

4.5.5 Identification and enrichment of cis-acting regulatory elements

A list of 53 *Arabidopsis* transcription factor binding sites was downloaded from AtcisDB (Davuluri et al., 2003; Yilmaz et al., 2011). Since the gene models in the genome did not have transcription start sites, we defined a region of 5000bp upstream and 2000bp downstream of the translation start site of each gene to search for cis-regulatory elements. Exact matches of the binding sites in these regions were counted.

For calculating the enrichment of a particular binding site among up-regulated genes (and similarly for among down-regulated genes), we assumed a null model of random sampling of the binding sites among all the genes. This leads to a hypergeometric distribution, and we calculate a corresponding p-value for each binding site sequence. False discovery rates were calculated from the p-values to correct for multiple testing.

4.5.6 Evaluating the significance of the binding site gain/loss

We use a fixed effects linear model to evaluate the significance of the loss or gain of a binding site on the differential expression.

For a gene i, we denote the fold change in expression at 60 minutes of ABA treatment as $(\beta_{i,60}/\beta_{i,0})$. For a particular binding site sequence, let the number of binding site sequences occurring in the promoter region be denoted as $n_{i,\mathrm{BSS}}$, and let its mapped *Arabidopsis thaliana* ortholog be coded as the categorical variable O_i. If there are a total of M *Arabidopsis thaliana* genes, then O_i is an M vector of all zeros except one 1 for the corresponding Arabidopsis gene. We model the log fold change as

$$\log 2 \left(\frac{\beta_{i,60}}{\beta_{i,0}} \right) = \beta_{\mathrm{BSS}} n_{i,\mathrm{BSS}} + \beta_{\mathrm{Orth}} \mathrm{O}_i.$$

The parameter β_{Orth} is simply a vector of the mean $\log 2$ fold changes for each gene family, where a gene family are all the *Brassica napus* genes corresponding to the same *Arabidopsis thaliana* gene. The parameter β_{BSS} captures the effect of the presence of a binding site sequence on the fold change after correcting for the common ancestry of the genes within a gene family. We do not necessarily expect the $\log 2$ fold changes to be linearly dependent on the binding site sequence presence. However, a significant non-zero value of β_{BSS} should signal the dependence of the fold change on the presence

of binding site sequence with an effect that is detected in a linear model.

The statistical significance is evaluated as the p-value of the F-test for the null hypothesis of $\beta_{\mathrm{BSS}} = 0$ and the alternative hypothesis of $\beta_{\mathrm{BSS}} \neq 0$.

We modeled the gene ancestry (for which O_i is a proxy) as a fixed effect and the statistical significance was evaluated using the lfe R package (Gaure, 2013a,b). Like other statistical analyses in this work, multiple testing correction was applied to arrive at false discovery rates as the adjust p-values.

4.5.7 Nucleotide substitution rates

Amino acid sequences of the translated gene products were aligned using ClustalW2 (Larkin et al., 2007). The aligned protein sequences were then used to align the nucleotide sequences using transAlign (Bininda-Emonds, 2005). The synonymous and non-synonymous nucleotide substitution rates were calculated in the seqinr R package (Charif and Lobry, 2007) using the model of Li (1993). Any values of k_s and k_a greater than 2 were discarded as missing values for subsequent analyses assuming that these might be incorrect ortholog assignments or alignments since we do not expect to observe substitution rates this high across the length of any gene.

4.5.8 Paralogous groups and *Arabidopsis thaliana-Brassica napus* orthologs

Orthologs between *Arabidopsis thaliana* and *Brassica napus* were mapped according to Cheng et al. (2012), for the purposes of translating metabolic networks, the regulatory interaction network, and the cross-species comparison of ABA response. All *Brassica napus* genes mapping to the same *Arabidopsis thaliana* gene were considered as a paralogous gene family. Out of the 101040 *Brassica napus* genes, we mapped the corresponding *Arabidopsis thaliana* orthologs for 60573 (i.e., 59.95%) of these genes. Considering only those genes that had any transcripts mapped to them in our RNASeq data, the fraction of genes with *Arabidopsis* orthologs rises to 69.73%. Some of the genes for which *Arabidopsis* orthologs were not found could be pseudogenes which have acquired too many mutations to be easily mapped to their cross-species ortholog. Among those identified as significantly regulated (false discovery rate less than 0.05), $9910/11925 = 83.10\%$ have *Arabidopsis* orthologs.

The size of a paralogous gene family (number of *Brassica napus* genes mapping to the same *Ara-*

bidopsis thaliana ortholog) varied from 1 to 16, with a mean size of 3.0810275.

4.6 Supplementary Information

4.6.1 Differential expression diagnostics

We explore some technical aspects of the observed gene expression data in this section to ensure that we are correctly modeling our observations. While this analysis does not provide us with any scientific results, they serve to assess the technical quality of the statistical models and confirm underlying assumptions.

4.6.1.1 Modeling batch effects

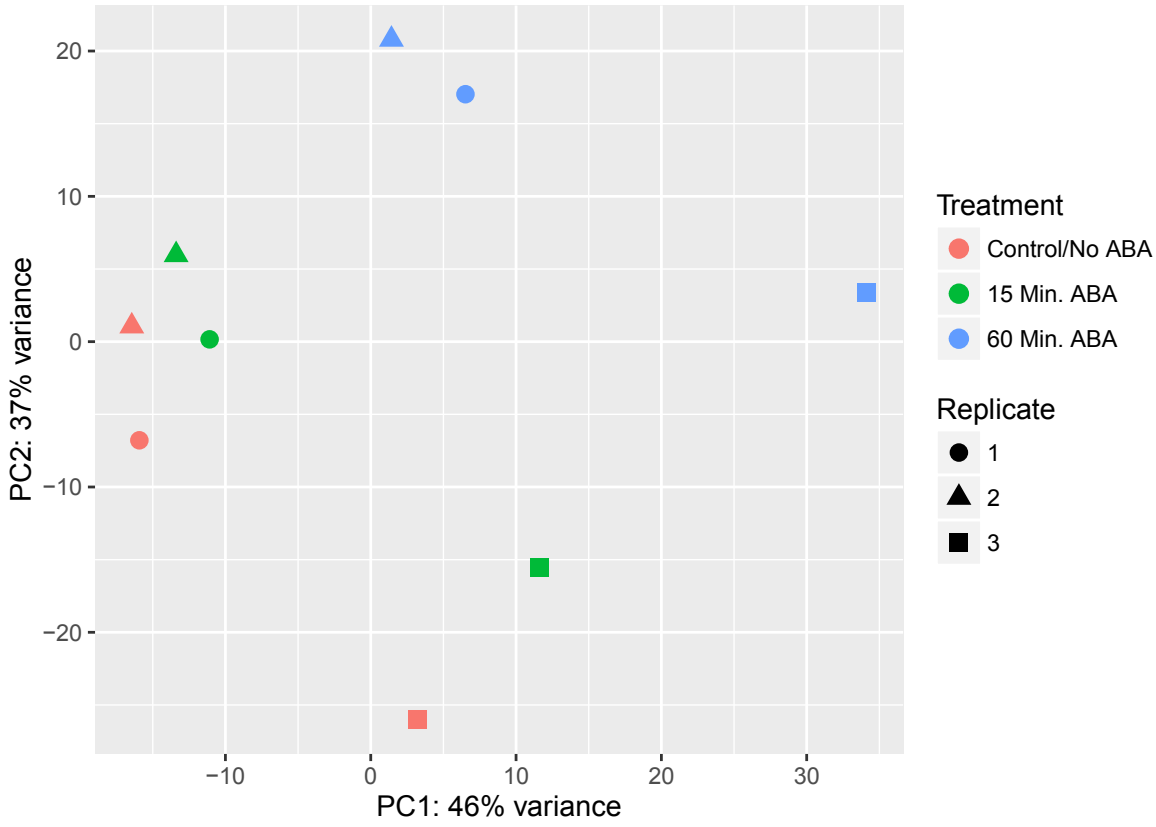

Figure 4-11: Principal component analysis of the variance in the log read counts of the genes in the 3 replicates at each of the 3 conditions. The first two principal components are plotted for all the 9 samples.

We collected mRNA from 3 replicates with 3 conditions each. The replicates help distinguish the relevant differential expression due to the effect of ABA from the background noise. It is possible that the 3 replicates are very similar to each other, in which case simply averaging over the replicates for each condition should reduce the noise. An alternative, however, is that the 3 replicates are very

different from each other, in which case the observed gene expression is meaningfully affected by both the replicate condition and ABA exposure. Principal component analysis provides an exploratory visualization of the sources of variation in high dimensional data. Since we have 9 samples, the read counts mapped to the genes from all the samples can be represented as $9 \times M$ matrix, where M is the number of genes. The covariance matrix for the observations can be calculated as a 9×9 matrix. Singular value decomposition of the covariance matrix provides to the principal components of our observed read counts.

Using the top 500 genes with the greatest variance as a proxy for the variation from biological sources as opposed to sequencing noise, we calculated the principal components. Each sample was mapped to the first (largest) two principal components in Figure 4-11, with the color and shape of the points denoting the sample condition (ABA exposure) and replicate. Both batch effects (in the direction PC1-PC2) ABA exposure effects (approximately in the PC1+PC2 direction) are evident.

When analysing the RNA-seq reads, this batch effect was explicitly modeled into the design matrix, as explained in the Methods (Section 4.5.4).

4.6.1.2 MA plot shows smaller differential expression at 15 minures

The mean of the log read counts versus the fold change between the control and treatment for every gene is shown in Figure 4-12. This visualization, also known as the MA plot or the Bland-Altman plot (Altman and Bland, 1983; Bland and Altman, 1999), confirms that the mean log fold change is zero for the bulk of the genes with both low and high basal gene expression. This is expected since the majority of the genes should not be either up-regulated or down-regulated. We see that there are more up-regulated than down-regulated genes, and there are slightly more up-regulated genes with higher mean expression than low expression. This might be because many of the genes involved in stomatal closure are basally expressed at higher rates in guard cells to maintain turgidity or to be able to respond to water deficits quickly. On the other hand, genes that are not required for guard cell functioning are not likely to be present in large quantities in the guard cell, and these are also unlikely to be regulated by ABA signaling.

Secondly, by comparing the graphs for 15 and 60 minutes, it is clear that, in general, larger fold changes are observed at 60 minutes compared to 15 minutes of ABA treatment.

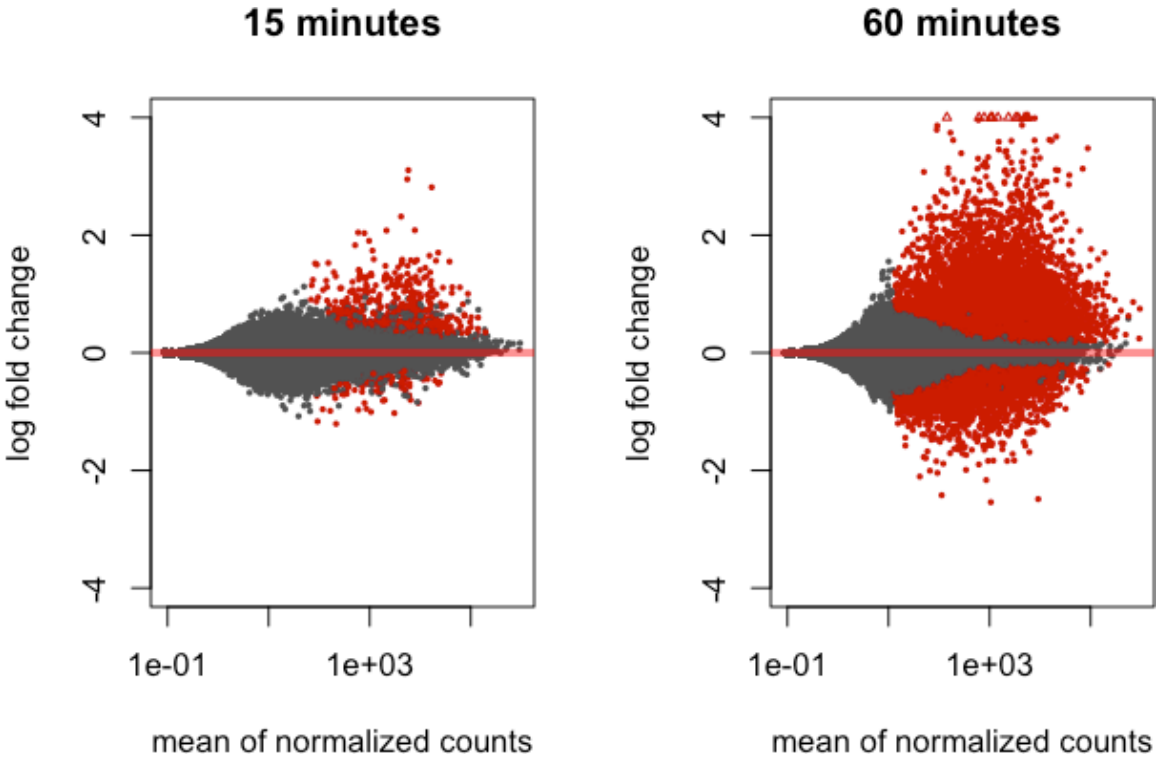

Figure 4-12: MA plot showing the log ratio (M) versus the average read count (A) for 15 minutes and 60 minutes. The red colored dots denote genes identified as significantly differentially expressed. The black colored dots represent genes that are not identified as significantly differentially expressed.

Chapter 5

Conclusion

This dissertation demonstrates the importance of mathematical modeling and statistical analysis for understanding aspects of cell signaling and pathway organization at increasing scales of length and organization: biological physics of molecular transport; interactions of cellular components in network and pathways; and functions of entire cells. We have made contributions in methods for biological physics, computational biology and network science, and understanding specific manifestations of cell signaling.

In terms of computational tools, we have developed and applied a range of methodologies including differential equation modeling (protein transfer in tunneling nanotubes), statistical testing (gene expression, evolution, and data integration), and statistical physics of disordered and stochastic systems (random walks, spin systems, and phase transitions). All of these tools can be employed for answering questions of real biological importance.

Our model of tunneling nanotubes (TNs) predicts the dependence of transfer through TNs on the size and localization (cytosolic vs. membrane bound) of the transferred biomolecules and the dynamics of TN growth. It relies on simplified assumptions about TN formation and passive transport. Experiments that test our predictions will be able to confirm or improve these features of the model. As the role of TNs in tissue and organ formation and cancer pathophysiology potentially gains prominence, more detailed mathematical models will be proposed and tested.

Graph diffusion kernels are a well known method of prioritizing candidate genes based on network data and known genes of interest. We relate these models to Ising model of spin interactions, which

assigns a distribution over the binary ($+1$ or -1) spins associated for each node in the graph as a Boltzmann distribtuion parametrized by temperature. The graph diffusion kernel and the linear estimate of the spin correlation between two nodes of a graph are related: the two are identical in the case of regular graphs (all nodes have the same degree), and of a similar form for general graphs of the kind encountered in network science in biological and non-biological domains. We show that the linear estimate of the spin-spin correlations out-performs than the graph diffusion kernel for predicting missing links in protein protein interaction networks. Secondly, we show that this best performance is when the spin distributions are calculated at a temperature just higher than the Curie point phase transition into ferromagnetism. Thus, there is a narrow parameter range where the spin-spin correlation function "understands" the correct network structure just before it starts to assigns all the nodes into a single cluster of aligned spins in accord a ferromagnetic phase transition, and the linear approximating function breaks down catastrophically. We also derive two novel non-linear approximations to the spin-spin correlation whose regime of valid temperature extends wider. The performance of our functions degrades gracefully below the Curie temperature.

In the context of analyzing interactome networks, this has two consequences. First, our method requires less stringent parameter tuning for good performance. Secondly, in certain networks with wildly differing edge densities and topology between regions of the same graph, the optimal temperatures for different regions may differ. In these cases, our method should show better performance than linear approximations even after an exhaustive parameter search.

We demonstrate and compare the performance of all these functions on two experimentally obtained protein-protein interaction networks. We also briefly discuss methods developed for the related problem of graph clustering; we describe how super-paramagnetic clustering and Markov clustering are related to our methods. Besides providing an intellectually satisfying unified view and a method for link prediction in networks, our method can be applied to candidate gene prioritization. Previously, candidate gene prioritization algorithms have been built on similar models of graph diffusion (Nitsch et al., 2010), random walks (Köhler et al., 2008), and electrical resistance (Suthram et al., 2008). Methods for candidate gene prioritization and related problems often integrate as much of the given knowledge about all the genes and interactions as possible, and attempt to rank the genes optimally based on these. The network analysis component of such a system can utlize our approximations of the spin-spin correlation as a measure of gene association.

Finally, we study the case of drought signaling in the guard cells of *Brassica napus* leading to

stomatal closure and limiting evaporative loss from the leaves. We analyse RNAseq reads from protoplasts treated with abscisic acid (ABA) to quantify its differential expression, and integrate this with cross-species expression, interactome, and regulatory sequence data. We are able to find the *Brassica napus* genes, interactions, and pathways underlying the transcriptomic and metabolic changes underlying the drought response. We find that the ABA response in guard cells is also related to a host of processes other than stomatal closure, most of which are biochemical changes to afford protection to the intracellular machinery during stress. Correlating sequence changes in cis-regulatory elements with differential expression, we uncover some evidence of the continuing evolution of drought response in *Brassica*. We hope that these insights can provide scientific clarity towards breeding and genetic engineering efforts for drought tolerant crops.

We can apply the network analysis developed in this work to identify additional *Brassica napus* genes involved in specific pathways in the drought response. We expect genes involved in a particular pathway to be closely associated with other genes of the same pathway in the interaction network. Since the *Brassica napus* interaction networks have not been experimentally oberved, we can translate the *Arabidopsis thaliana* protein-protein interaction (PPI) network to a purported *Brassica napus* PPI network using ortholog data. Assuming the association kernel is a matrix $\mathbf{G}$, the network association score of any gene i from a set of query genes Q is $\sum_{q \in Q} G_{iq}$, where the kernel can be the graph diffusion kernel, or in our case, the spin-spin correlation matrix. Using known members of a pathway (such as the ABA signaling network or a part thereof) as the query gene set Q, all the other genes can be ranked according to their network scores. The appropriate network model parameter (i.e., temperature for the spin-spin correlation) can either be tuned in a supervised learning fashion using cross-validation with known pathway members if we have sufficient data. As we have shown in this work, we can also use link prediction to learn the optimal temperature, which would be especially useful if we do not have sufficient confidence in the known true members of the pathway of interest for parameter fitting. We can also integrate other sources of data, such as differential expression, by adding a differential expression score as the fold change, p-value, or an equivalent z-score computed from it. In addition, multiple network scores could theoretically be computed if there were additional network data (such as genetic or regulatory interactions, multiple systematic studies, or homology based translation from multiple species). Obviously increasing the number of such features and their associated parameters and weights can lead to potentially more powerful prediction systems, but they require larger amounts of training data to learn all the parameters. With the support of additional experimental data, we would like to continue this work to explore

these ideas further and develop them into concrete algorithms and prediction systems.

Bibliography

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2002). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) Function as Transcriptional Activators in Abscisic Acid Signaling. *The Plant cell*, 15(1):63–78.

Abou-Khalil, R., Le Grand, F., Pallafacchina, G., Valable, S., Authier, F.-J., Rudnicki, M. A., Gherardi, R. K., Germain, S., Chretien, F., Sotiropoulos, A., Lafuste, P., Montarras, D., and Chazaud, B. (2009). Autocrine and Paracrine Angiopoietin 1/Tie-2 Signaling Promotes Muscle Satellite Cell Self-Renewal. *Cell Stem Cell*, 5(3):298–309.

Agarwal, P. K. and Jha, B. (2010). Transcription factors in plants and ABA dependent and independent abiotic stress signalling. *Biologia Plantarum*, 54(2):201–212.

Agnati, L. F., Guidolin, D., Leo, G., Guescini, M., Pizzi, M., Stocchi, V., Spano, P. F., Ghidoni, R., Ciruela, F., Genedani, S., and Fuxe, K. (2011). Possible new targets for GPCR modulation: allosteric interactions, plasma membrane domains, intercellular transfer and epigenetic mechanisms. *Journal of Receptors and Signal Transduction*, 31(5):315–331.

Ahmed, H. R. and Glasgow, J. I. (2014). Pattern discovery in protein networks reveals high-confidence predictions of novel interactions.

Ahmed, K. A. and Xiang, J. (2011). Mechanisms of cellular communication through intercellular protein transfer. *Journal of Cellular and Molecular Medicine*, 15(7):1458–1473.

Al-Nedawi, K., Meehan, B., Micallef, J., Lhotak, V., May, L., Guha, A., and Rak, J. (2008). Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. *Nature Cell Biology*, 10(5):619–624.

Alcazar, R., Bitrián, M., Bartels, D., Koncz, C., Altabella, T., and Tiburcio, A. F. (2011). Polyamine

metabolic canalization in response to drought stress in Arabidopsis and the resurrection plant Craterostigma plantagineum. *Plant Signaling & Behavior*, 6(2):243–250.

Allender, C. J. and King, G. J. (2010). Origins of the amphiploid species Brassica napus L. investigated by chloroplast and nuclear molecular markers. *BMC Plant Biology*, 10(1):54.

Altman, D. and Bland, J. (1983). Measurement in medicine: the analysis of method comparison studies. *The statistician*.

Ambudkar, S. V., Sauna, Z. E., Gottesman, M. M., and Szakacs, G. (2005). A novel way to spread drug resistance in tumor cells: functional intercellular transfer of P-glycoprotein (ABCB1). *Trends in Pharmacological Sciences*, 26(8):385–387.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2):166–9.

Anderson, J. P., Badruzsaufari, E., Schenk, P. M., Manners, J. M., Desmond, O. J., Ehlert, C., Maclean, D. J., Ebert, P. R., and Kazan, K. (2004). Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *The Plant cell*, 16(12):3460–79.

Annila, A. and Annila, E. (2008). Why did life emerge? *International Journal of Astrobiology*, 7(3-4):293.

Assmann, S., Snyder, J., and Lee, Y. (2000). ABA-deficient (aba1) and ABA-insensitive (abi1-1, abi2-1) mutants of Arabidopsis have a wild-type stomatal response to humidity. *Plant, Cell & Environment*.

Baba, E., Takahashi, Y., Lichtenfeld, J., Tanaka, R., Yoshida, A., Sugamura, K., Yamamoto, N., and Tanaka, Y. (2001). Functional CD4 T Cells after Intercellular Molecular Transfer of OX40 Ligand. *The Journal of Immunology*, 167(2):875–883.

Bader, G. D., Hogue, C. W., Bader, G. D., and Hogue, C. W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, 20(10):991–997.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85.

Bak, G., Lee, E.-J., Lee, Y., Kato, M., Segami, S., Sze, H., Maeshima, M., Hwang, J.-U., and Lee, Y. (2013). Rapid structural changes and acidification of guard cell vacuoles during stomatal closure require phosphatidylinositol 3,5-bisphosphate. *The Plant cell*, 25(6):2202–16.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Barnett, J. and Adger, W. N. (2007). Climate change, human security and violent conflict. *Political Geography*, 26(6):639–655.

Barzel, B. and Barabási, A.-L. (2013). Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, 31(8):720–725.

Bauer, H., Ache, P., Lautner, S., Fromm, J., Hartung, W., Al-Rasheid, K. A., Sonnewald, S., Sonnewald, U., Kneitz, S., Lachmann, N., Mendel, R. R., Bittner, F., Hetherington, A. M., and Hedrich, R. (2013). The Stomatal Response to Reduced Relative Humidity Requires Guard Cell-Autonomous ABA Synthesis.

Bechtold, U., Albihlal, W. S., Lawson, T., Fryer, M. J., Sparrow, P. A. C., Richard, F., Persad, R., Bowden, L., Hickman, R., Martin, C., Beynon, J. L., Buchanan-Wollaston, V., Baker, N. R., Morison, J. I. L., Schoffl, F., Ott, S., and Mullineaux, P. M. (2013). Arabidopsis HEAT SHOCK TRANSCRIPTION FACTORA1b overexpression enhances water productivity, resistance to drought, and infection. *Journal of Experimental Botany*, 64(11):3467–3481.

Begg, J. E. (1980). Morphological adaptations of leaves to water stress. In Turner, N. and Kramer, P., editors, *Adaptation of Plants to Water and High Temperature Stress*, pages 33–42. Wiley Interscience: New York.

Behnke, B. J., Armstrong, R. B., and Delp, M. D. (2011). Adrenergic control of vascular resistance varies in muscles composed of different fiber types: influence of the vascular endothelium. *AJP: Regulatory, Integrative and Comparative Physiology*, 301(3):R783–R790.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289 – 300.

Bethe, H. A. (1935). Statistical Theory of Superlattices. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 150(871):552–575.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., and West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17):7301–6.

Bininda-Emonds, O. (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, 6(1):156.

Blackman, P. and Davies, W. (1985). Root to shoot communication in maize plants of the effects of soil drying. *Journal of Experimental Botany*.

Blanc, G., Hokamp, K., and Wolfe, K. H. (2003). A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome. *Genome Research*, 13(2):137–144.

Bland, J. and Altman, D. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*.

Blatt, M., Wiseman, S., and Domany, E. (1996). Superparamagnetic Clustering of Data. *Physical Review Letters*, 76(18):3251–3254.

Blum, A. and Ebercon, A. (1976). Genotypic Responses in Sorghum to Drought Stress. III. Free Proline Accumulation and Drought Resistance1. *Crop Science*, 16(3):428.

Bosenburg, M. and Massague, J. (1993). Juxtacrine cell signaling molecules. *Current Opinion in Cell Biology*, 5(5):832–838.

Brin, S. and Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1/7):107–17.

Bukoreshtliev, N. V., Wang, X., Hodneland, E., Gurke, S., Barroso, J. F., and Gerdes, H.-H. (2009). Selective block of tunneling nanotube (TNT) formation inhibits intercellular organelle transfer between PC12 cells. *FEBS Letters*, 583(9):1481–1488.

Burgess, M., Adar, E., and Cafarella, M. (2016). Link-Prediction Enhanced Consensus Clustering for Complex Networks. *PLOS ONE*, 11(5):e0153384.

Camussi, G., Deregibus, M. C., Bruno, S., Cantaluppi, V., and Biancone, L. (2010). Exosomes/microvesicles as a mechanism of cell-to-cell communication. *Kidney International*, 78(9):838–848.

Carlin, L. M., Eleme, K., McCann, F. E., and Davis, D. M. (2001). Intercellular Transfer and Supramolecular Organization of Human Leukocyte Antigen C at Inhibitory Natural Killer Cell Immune Synapses. *The Journal of Experimental Medicine*, 194(10):1507–1517.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., and Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 44(D1):D471–80.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. a. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Correa, M., Da Silva, C., Just, J., Falentin, C., Koh, C. S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., Labadie, K., Alberti, A., Charles, M., Arnaud, D., Guo, H., Daviaud, C., Alamery, S., Jabbari, K., Zhao, M., Edger, P. P., Chelaifa, H., Tack, D., Lassalle, G., Mestiri, I., Schnel, N., Le Paslier, M.-C., Fan, G., Renault, V., Bayer, P. E., Golicz, a. a., Manoli, S., Lee, T.-H., Thi, V. H. D., Chalabi, S., Hu, Q., Fan, C., Tollenaere, R., Lu, Y., Battail, C., Shen, J., Sidebottom, C. H. D., Canaguier, A., Chauveau, A., Berard, A., Deniot, G., Guan, M., Liu, Z., Sun, F., Lim, Y. P., Lyons, E., Town, C. D., Bancroft, I., Meng, J., Ma, J., Pires, J. C., King, G. J., Brunel, D., Delourme, R., Renard, M., Aury, J.-M., Adams, K. L., Batley, J., Snowdon, R. J., Tost, J., Edwards, D., Zhou, Y., Hua, W., Sharpe, a. G., Paterson, a. H., Guan, C., and Wincker, P. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, 345(6199):950–953.

Charif, D. and Lobry, J. R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. pages 207–232. Springer Berlin Heidelberg.

Chen, F., Mackey, A. J., Stoeckert, C. J., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic acids research*, 34(Database issue):D363–8.

Chen, J., Hsu, W., Lee, M. L., and Ng, S.-K. (2005). Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in Medicine*, 35(1):37–47.

Cheng, F., Wu, J., Fang, L., and Wang, X. (2012). Syntenic gene analysis between Brassica rapa and other Brassicaceae species. *Frontiers in Plant Science*, 3:198.

Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of Brassica plants. *Horticulture research*, 1:14024.

Cheng, M.-C., Liao, P.-M., Kuo, W.-W., and Lin, T.-P. (2013). The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *Plant physiology*, 162(3):1566–82.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic acids research*, 26(1):73–9.

Chinnery, H. R., Pearlman, E., and McMenamin, P. G. (2008). Cutting Edge: Membrane Nanotubes In Vivo: A Feature of MHC Class II+ Cells in the Mouse Cornea. *The Journal of Immunology*, 180(9):5779–5783.

Choi, H.-i., Hong, J.-h., Ha, J.-o., Kang, J.-y., and Kim, S. Y. (2000). ABFs, a Family of ABA-responsive Element Binding Factors. *Journal of Biological Chemistry*, 275(3):1723–1730.

Christmann, A., Weiler, E. W., Steudle, E., and Grill, E. (2007). A hydraulic signal in root-to-shoot signalling of water shortage. *The Plant Journal*, 52(1):167–174.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics (Oxford, England)*, 22(13):1623–30.

Chung, F. (2007). The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A. D., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R. (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058):529–533.

Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

Comstock, J. and Mencuccini, M. (1998). Control of stomatal conductance by leaf water potential in Hymenoclea salsola (T. & G.), a desert subshrub. *Plant, Cell and Environment*, 21(10):1029–1038.

Cook, E. R., Woodhouse, C. A., Eakin, C. M., Meko, D. M., and Stahle, D. W. (2004). Long-term aridity changes in the western United States. *Science (New York, N.Y.)*, 306(5698):1015–8.

Cornic, G. (2000). Drought stress inhibits photosynthesis by decreasing stomatal aperture-not by affecting ATP synthesis. *Trends in plant science*, 5(5):187–188.

Costanzo, M., Abounit, S., Marzo, L., Danckaert, A., Chamoun, Z., Roux, P., and Zurzolo, C. (2013). Transfer of polyglutamine aggregates in neuronal cells occurs in tunneling nanotubes. *Journal of Cell Science*, 126(16):3678–3685.

Couttenier, M. and Soubeyran, R. (2014). Drought and Civil War In Sub-Saharan Africa. *The Economic Journal*, 124(575):201–244.

Cramer, H. (1947). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726.

Dai, A. (2012). Increasing drought under global warming in observations and models. *Nature Climate Change*, 3(1):52–58.

Davis, D. M. (2007). Intercellular transfer of cell-surface proteins is common and can affect many stages of an immune response. *Nature Reviews Immunology*, 7(3):238–243.

Davuluri, R., Sun, H., and Palaniswamy, S. (2003). AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC*.

Deane, C. M., Salwiński, Ł., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & cellular proteomics : MCP*, 1(5):349–56.

Denzer, K., Kleijmeer, M. J., Heijnen, H. F., Stoorvogel, W., and Geuze, H. J. (2000). Exosome: from internal vesicle of the multivesicular body to intercellular signaling device. *Journal of Cell Science*, 113(19):3365–3374.

Dickson, E., Thayanithy, V., Wong, P., Teoh, D., Argenta, P., Isaksson Vogel, R., Steer, C., Geller, M., Subramanian, S., and Lou, E. (2014). Novel mechanisms of chemoresistance in ovarian cancer: The role of tunneling nanotubes. *Gynecologic Oncology*, 133:109–110.

Ding, Y., Liu, N., Virlouvet, L., Riethoven, J.-J., Fromm, M., and Avramova, Z. (2013). Four distinct types of dehydration stress memory genes in Arabidopsis thaliana. *BMC plant biology*, 13(1):229.

Dreisen, R., Dispersyn, G., Verheyen, F., Vandeneijde, S., Hofstra, L., Thone, F., Dijsktra, P., Debie, W., Borgers, M., and Ramakers, F. (2005). Partial cell fusion: A newly recognized type of communication between dedifferentiating cardiomyocytes and fibroblasts. *Cardiovascular Research*, 68(1):37–46.

Eddelbuettel, D. and Fran, R. (2011). Rcpp : Seamless R and C++ Integration. *Journal Of Statistical Software*, 40:1–18.

Elliott, J., Deryng, D., Müller, C., Frieler, K., Konzmann, M., Gerten, D., Glotter, M., Flörke, M., Wada, Y., Best, N., Eisner, S., Fekete, B. M., Folberth, C., Foster, I., Gosling, S. N., Haddeland, I., Khabarov, N., Ludwig, F., Masaki, Y., Olin, S., Rosenzweig, C., Ruane, A. C., Satoh, Y., Schmid, E., Stacke, T., Tang, Q., and Wisser, D. (2014). Constraints and potentials of future irrigation water availability on agricultural production under climate change. *Proceedings of the National Academy of Sciences*, 111(9):3239–3244.

England, J. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12).

England, J. L. (2015). Dissipative adaptation in driven self-assembly. *Nature Nanotechnology*, 10(11):919–923.

Ermolaeva, M., Wu, M., Eisen, J., and Salzberg, S. (2003). The age of the Arabidopsis thaliana genome duplication. *Plant molecular biology*.

Eugenin, E., Gaskill, P., and Berman, J. (2009). Tunneling nanotubes (TNT) are induced by HIV-infection of macrophages: A potential mechanism for intercellular HIV trafficking. *Cellular Immunology*, 254(2):142–148.

Fang, Y. and Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cellular and Molecular Life Sciences*, 72(4):673–689.

Farach-Colton, M., Huang, Y., and Woolford, J. L. L. (2004). Discovering temporal relations in molecular pathways using protein-protein interactions. In *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04*, pages 150–156, New York, New York, USA. ACM Press.

Feng, C.-Z., Chen, Y., Wang, C., Kong, Y.-H., Wu, W.-H., and Chen, Y.-F. (2014). Arabidopsis RAV1 transcription factor, phosphorylated by SnRK2 kinases, regulates the expression of ABI3, ABI4, and ABI5 during seed germination and early seedling development. *The Plant journal : for cell and molecular biology*, 80(4):654–68.

Fields, S. and Song, O.-k. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Fields, S., Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

Fire, M., Tenenboim-Chekina, L., Puzis, R., Lesser, O., Rokach, L., and Elovici, Y. (2013). Computationally efficient link prediction in a variety of social networks. *ACM Transactions on Intelligent Systems and Technology*, 5(1):1–25.

Fisher, R. A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87.

Fjelde, H. (2015). Farming or Fighting? Agricultural Price Shocks and Civil War in Africa. *World Development*, 67:525–534.

Fraiman, D., Balenzuela, P., Foss, J., and Chialvo, D. R. (2009). Ising-like dynamics in large-scale functional brain networks. *Physical Review E*, 79(6):061922.

Fuchs, E. and Livingston, N. (1996). Hydraulic control of stomatal conductance in Douglas fir seedlings. *Plant, Cell & Environment*.

Gallavotti, G. and Cohen, E. G. D. (1995). Dynamical Ensembles in Nonequilibrium Statistical Mechanics. *Physical Review Letters*, 74(14):2694–2697.

Gaure, S. (2013a). lfe: Linear group fixed effects. *The R Journal*.

Gaure, S. (2013b). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis*, 66:8–18.

Georges, a. and Yedidia, J. S. (1991). How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173–2192.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A Protein Interaction Map of Drosophila melanogaster. *Science*, 302(5651):1727–1736.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–6.

Gleick, P. H. and Gleick, P. H. (2014). Water, Drought, Climate Change, and Conflict in Syria. *Weather, Climate, and Society*, 6(3):331–340.

Goldberg, D. S. and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8):4372–6.

Gowing, D. and Davies, W. (1990). A positive root-sourced signal as an indicator of soil drying in apple, Malus x domestica Borkh. *Journal of Experimental*.

Grabowski, A. and Kosiński, R. (2006). Ising-based model of opinion formation in a complex network of interpersonal interactions. *Physica A: Statistical Mechanics and its Applications*, 361(2):651–664.

Gregor, T., Bialek, W., van Steveninck, R. R. d. R., Tank, D. W., and Wieschaus, E. F. (2005). Diffusion and scaling during early embryonic pattern formation. *Proceedings of the National Academy of Sciences*, 102(51):18403–18407.

Groebe, K., Erz, S., and Mueller-Klieser, W. (1994). Glucose Diffusion Coefficients Determined from Concentration Profiles in Emt6 Tumor Spheroids Incubated in Radioactively Labeled L-Glucose. pages 619–625.

Gruber, C. and Griffiths, R. B. (1986). Phase transition in a ferromagnetic fluid. *Physica A: Statistical Mechanics and its Applications*, 138(1):220–230.

Guescini, M., Leo, G., Genedani, S., Carone, C., Pederzoli, F., Ciruela, F., Guidolin, D., Stocchi, V., Mantuano, M., Borroto-Escuela, D., Fuxe, K., and Agnati, L. (2012). Microvesicle and tunneling nanotube mediated intercellular transfer of g-protein coupled receptors in cell cultures. *Experimental Cell Research*, 318(5):603–613.

Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52):22073–8.

Gurke, S., Barroso, J. F., Hodneland, E., Bukoreshtliev, N. V., Schlicker, O., and Gerdes, H.-H. (2008). Tunneling nanotube (TNT)-like structures facilitate a constitutive, actomyosin-dependent exchange of endocytic organelles between normal rat kidney cells. *Experimental Cell Research*, 314(20):3669–3683.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). BioMart Central Portal–unified access to biological data. *Nucleic acids research*, 37(Web Server issue):W23–7.

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844.

Hanson, A. D., Nelsen, C. E., Pedersen, A. R., and Everson, E. H. (1979). Capacity for Proline Accumulation During Water Stress in Barley and Its Implications for Breeding for Drought Resistance. *Crop Science*, 19(4):489.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(supp):C47–C52.

Hauser, F., Waadt, R., and Schroeder, J. I. (2011). Evolution of abscisic acid synthesis and signaling mechanisms.

Heikkila, J. J., Papp, J. E. T., Schultz, G. A., and Bewley, J. D. (1984). Induction of Heat Shock Protein Messenger RNA in Maize Mesocotyls by Water Stress, Abscisic Acid, and Wounding. *PLANT PHYSIOLOGY*, 76(1):270–274.

Hoekstra, A. Y. and Chapagain, A. K. (2006). Water footprints of nations: Water use by people as a function of their consumption pattern. *Water Resources Management*, 21(1):35–48.

Hoey, J. (2012). The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test. *ArXiv e-prints*.

Hose, E., Steudle, E., and Hartung, W. (2000). Abscisic acid and hydraulic conductivity of maize roots: a study using cell- and root-pressure probes. *Planta*, 211(6):874–82.

Hoth, S., Morgante, M., Sanchez, J.-P., Hanafey, M. K., Tingey, S. V., and Chua, N.-H. (2002). Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. *Journal of cell science*, 115(Pt 24):4891–4900.

Huang, D., Wu, W., Abrams, S. R., and Cutler, A. J. (2008). The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. *Journal of experimental botany*, 59(11):2991–3007.

Huang, H., Jedynak, B. M., Bader, J. S., Hall, D., and Casamayor, A. (2007). Where Have All the Interactions Gone? Estimating the Coverage of Two-Hybrid Protein Interaction Maps. *PLoS Computational Biology*, 3(11):e214.

Huang, Z., Li, X., and Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05*, page 141, New York, New York, USA. ACM Press.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–74.

Jeh, G. and Widom, J. (2002). SimRank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 538, New York, New York, USA. ACM Press.

Jones, R. and Mansfield, T. (1970). Suppression of stomatal opening in leaves treated with abscisic acid. *Journal of Experimental Botany*.

Jonnson, R. (2009). Breeding for improved oil and meal quality in rape (Brassica napus L.) and turnip rape (Brassica campestris L.). *Hereditas*, 87(2):205–218.

Kalamaki, M. S., Alexandrou, D., Lazari, D., Merkouropoulos, G., Fotopoulos, V., Pateraki, I., Aggelis, A., Carrillo-López, A., Rubio-Cabetas, M. J., and Kanellis, A. K. (2009). Over-expression of a tomato N-acetyl-L-glutamate synthase gene (SlNAGS1) in Arabidopsis thaliana results in high ornithine levels and increased tolerance in salt and drought stresses. *Journal of experimental botany*, 60(6):1859–71.

Kappen, H. and Rodriguez, F. (1998). Boltzmann machine learning using mean field theory and linear response correction. In *Advances in neural information information processing systems*, pages 280–286.

Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–8.

Kastrin, A., Rindflesch, T. C., and Hristovski, D. (2016). Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. *Methods of Information in Medicine*, 55(4):340–346.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kelley, C. P., Mohtadi, S., Cane, M. A., Seager, R., and Kushnir, Y. (2015). Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proceedings of the National Academy of Sciences*, 112(11):3241–3246.

Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F. C. (2002). Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data. *Molecular Cell*, 9(5):1133–1143.

Khalil, K. S., Sagastegui, A., Li, Y., Tahir, M. A., Socolar, J. E. S., Wiley, B. J., and Yellen, B. B. (2012). Binary colloidal structures assembled through Ising interactions. *Nature Communications*, 3:794.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36.

Kimura, Y., Aoki, S., Ando, E., Kitatsuji, A., Watanabe, A., Ohnishi, M., Takahashi, K., Inoue, S.-i., Nakamichi, N., Tamada, Y., and Kinoshita, T. (2015). A Flowering Integrator, SOC1, Affects Stomatal Opening in Arabidopsis thaliana. *Plant and Cell Physiology*, 56(4):640–649.

Kishor, P., Hong, Z., Miao, G., and Hu, C. (1995). Overexpression of delta-1-pyrroline-5-carboxylate synthetase increases proline production and confers osmotolerance in transgenic plants. *Plant*.

Klein, D. J. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82(4):949–958.

Kondor, R. and Lafferty, J. D. (2002). Diffusion Kernels on Graphs and Other Discrete Input Spaces. In Sammut, C. and Hoffmann, A. G., editors, *Proceedings of the Nineteenth International Conference on Machine Learning ICML 2002*, pages 315–322, Sydney, Australia. Morgan Kaufmann.

Kondra, Z. P. and Stefansson, B. R. (1965). INHERITANCE OF ERUCIC AND EICOSENOIC ACID CONTENT OF RAPESEED OIL (BRASSICA NAPUS). *Canadian Journal of Genetics and Cytology*, 7(3):505–510.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

Kurchan, J. (1998). Fluctuation theorem for stochastic dynamics. *Journal of Physics A: Mathematical and General*, 31(16):3719–3729.

Lachmann, A. and Ma'ayan, A. (2010). Lists2Networks: Integrated analysis of gene/protein lists. *BMC Bioinformatics*, 11(1):87.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins,

D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–8.

Larkindale, J. and Knight, M. R. (2002). Protection against Heat Stress-Induced Oxidative Damage in Arabidopsis Involves Calcium, Abscisic Acid, Ethylene, and Salicylic Acid. *PLANT PHYSIOLOGY*, 128(2):682–695.

Leakey, A. D., Uribelarrea, M., Ainsworth, E. A., Naidu, S. L., Rogers, A., Ort, D. R., and Long, S. P. (2006). Photosynthesis, Productivity, and Yield of Maize Are Not Affected by Open-Air Elevation of CO2 Concentration in the Absence of Drought. *PLANT PHYSIOLOGY*, 140(2):779–790.

Leicht, E. A., Holme, P., and Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E*, 73(2):026120.

Leonhardt, N., Kwak, J. M., Robert, N., Waner, D., Leonhardt, G., and Schroeder, J. I. (2004). Microarray Expression Analyses of Arabidopsis Guard Cells and Isolation of a Recessive Abscisic Acid Hypersensitive Protein Phosphatase 2C Mutant. *THE PLANT CELL ONLINE*, 16(3):596–615.

Levchenko, A., Mehta, B. M., Niu, X., Kang, G., Villafania, L., Way, D., Polycarpe, D., Sadelain, M., and Larson, S. M. (2005). Intercellular transfer of P-glycoprotein mediates acquired multidrug resistance in tumor cells. *Proceedings of the National Academy of Sciences*, 102(6):1933–1938.

Li, M., Aliotta, J. M., Asara, J. M., Wu, Q., Dooner, M. S., Tucker, L. D., Wells, A., Quesenberry, P. J., and Ramratnam, B. (2010). Intercellular Transfer of Proteins as Identified by Stable Isotope Labeling of Amino Acids in Cell Culture. *Journal of Biological Chemistry*, 285(9):6285–6297.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution.

Li, X. and Chen, H. (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems*, 54(2):880–890.

Lichtenberger, B. M., Tan, P. K., Niederleithner, H., Ferrara, N., Petzelbauer, P., and Sibilia, M. (2010). Autocrine VEGF Signaling Synergizes with EGFR in Tumor Cells to Promote Epithelial Cancer Development. *Cell*, 140(2):268–279.

Liu, J.-S. and Ning, K.-C. (2011). Applying Link Prediction to Ranking Candidates for High-Level Government Post. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 145–152. IEEE.

Lo Gullo, M. A., Nardini, A., Salleo, S., and Tyree, M. T. (1998). Changes in root hydraulic conductance (KR) of Olea oleaster seedlings following drought stress and irrigation. *New Phytologist*, 140(1):25–31.

Lou, E., Fujisawa, S., Morozov, A., Barlas, A., Romin, Y., Dogan, Y., Gholami, S., Moreira, A. L., Manova-Todorova, K., and Moore, M. A. S. (2012). Tunneling Nanotubes Provide a Unique Conduit for Intercellular Transfer of Cellular Contents in Human Malignant Pleural Mesothelioma. *PLoS ONE*, 7(3):e33093.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Technical report.

Luby-Phelps, K. (1986). Probing the structure of cytoplasm. *The Journal of Cell Biology*, 102(6):2015–2022.

MacDonald, G. M. (2010). Water, climate change, and sustainability in the southwest. *Proceedings of the National Academy of Sciences*, 107(50):21256–21262.

Mack, M., Kleinschmidt, A., Brühl, H., Klier, C., Nelson, P. J., Cihak, J., Plachý, J., Stangassinger, M., Erfle, V., and Schlöndorff, D. (2000). Transfer of the chemokine receptor CCR5 between cells by membrane-derived microparticles: a mechanism for cellular human immunodeficiency virus 1 infection. *Nature Medicine*, 6(7):769–775.

Marzo, L., Gousset, K., and Zurzolo, C. (2012). Multifaceted Roles of Tunneling Nanotubes in Intercellular Communication. *Frontiers in Physiology*, 3.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome research*, 11(12):2120–6.

Medrano, H., Eecalona, J. M., Bota, J., Gulias, J., and Flexas, J. (2002). Regulation of Photosynthesis of C3 Plants in Response to Progressive Drought: Stomatal Conductance as a Reference Parameter. *Annals of Botany*, 89(7):895–905.

Meyer, D., Zeileis, A., Hornik, K., Gerber, F., Friendly, M., and Meyer, M. D. (2016). Package 'vcd'.

Michaelian, K. (2009). Thermodynamic dissipation theory for the origin of life. *Earth System Dynamics, Volume 2, Issue 1, 2011, pp.37-51*, 2:37–51.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is There a Bias in Proteome Research? *Genome Research*, 11(12):1971–1973.

Nakai, Y., Nakahira, Y., Sumida, H., Takebayashi, K., Nagasawa, Y., Yamasaki, K., Akiyama, M., Ohme-Takagi, M., Fujiwara, S., Shiina, T., Mitsuda, N., Fukusaki, E., Kubo, Y., and Sato, M. H. (2013). Vascular plant one-zinc-finger protein 1/2 transcription factors regulate abiotic and biotic stress responses in Arabidopsis. *The Plant journal : for cell and molecular biology*, 73(5):761–75.

Nakanishi, K. and Takayama, H. (1997). Mean-field theory for a spin-glass model of neural networks: TAP free energy and the paramagnetic to spin-glass transition. *Journal of Physics A: Mathematical and General*, 30(23):8085–8094.

Nakashima, K., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2014). The transcriptional regulatory network in the drought response and its crosstalk in abiotic stress responses including drought, cold, and heat. *Frontiers in plant science*, 5:170.

Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2003). Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. In *Pacific Symposium on Biocomputing (PSB03)*, pages 336–347.

Narusaka, Y., Nakashima, K., Shinwari, Z. K., Sakuma, Y., Furihata, T., Abe, H., Narusaka, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. *The Plant Journal*, 34(2):137–148.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102.

Nicholson, C. and Tao, L. (1993). Hindered diffusion of high molecular weight compounds in brain extracellular microenvironment measured with integrative optical imaging. *Biophysical Journal*, 65(6):2277–2290.

Nishizawa, A., Yabuta, Y., and Shigeoka, S. (2008). Galactinol and Raffinose Constitute a Novel Function to Protect Plants from Oxidative Damage. *PLANT PHYSIOLOGY*, 147(3):1251–1263.

Nishizawa-Yokoi, A., Nosaka, R., Hayashi, H., Tainaka, H., Maruta, T., Tamoi, M., Ikeda, M., Ohme-Takagi, M., Yoshimura, K., Yabuta, Y., and Shigeoka, S. (2011). HsfA1d and HsfA1e Involved in the Transcriptional Regulation of HsfA2 Function as Key Regulators for the Hsf Signaling Network in Response to Environmental Stress. *Plant and Cell Physiology*, 52(5):933–945.

Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B., and Moreau, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460.

Niu, X., Gupta, K., Yang, J. T., Shamblott, M. J., and Levchenko, A. (2009). Physical transfer of membrane and cytoplasmic components as a general mechanism of cell-cell communication. *Journal of Cell Science*, 122(5):600–610.

North, G. B. and Nobel, P. S. (1992). Drought-induced changes in hydraulic conductivity and structure in roots of Ferocactus acanthodes and Opuntia ficus-indica. *New Phytologist*, 120(1):9–19.

Novillo, F., Alonso, J. M., Ecker, J. R., and Salinas, J. (2004). CBF2/DREB1C is a negative regulator of CBF1/DREB1B and CBF3/DREB1A expression and plays a central role in stress tolerance in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3985–90.

Okamoto, M., Kuwahara, A., Seo, M., Kushiro, T., Asami, T., Hirai, N., Kamiya, Y., Koshiba, T., and Nambara, E. (2006). CYP707A1 and CYP707A2, which encode abscisic acid 8'-hydroxylases, are indispensable for proper control of seed dormancy and germination in Arabidopsis. *Plant physiology*, 141(1):97–107.

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, 43(8):691–8.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, 21(6):832–4.

Palm, R. and Korenivski, V. (2009). A ferrofluid based neural network: design of an analogue associative memory. *New Journal of Physics*, 11.

Pan, L., Zhou, T., Lü, L., and Hu, C.-K. (2016). Predicting missing links and identifying spurious links via likelihood analysis. *Scientific reports*, 6:22955.

Panek, J. A. and Goldstein, A. H. (2001). Response of stomatal conductance to drought in ponderosa pine: implications for carbon and ozone uptake. *Tree Physiology*, 21(5):337–344.

Panikulangara, T. J., Eggers-Schumacher, G., Wunderlich, M., Stransky, H., and Schöffl, F. (2004). Galactinol synthase1. A novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in Arabidopsis. *Plant physiology*, 136(2):3148–58.

Pap, E., Pállinger, É., Pásztói, M., and Falus, A. (2009). Highlights of a new type of intercellular communication: microvesicle-based information transfer. *Inflammation Research*, 58(1):1–8.

Parkin, I. A. P., Sharpe, A. G., Keith, D. J., and Lydiate, D. J. (1995). Identification of the A and C genomes of amphidiploid Brassica napus (oilseed rape). *Genome*, 38(6):1122–1131.

Pataki, D. E., Oren, R., and Tissue, D. T. (1998). Elevated carbon dioxide does not affect average canopy stomatal conductance of Pinus taeda L. *Oecologia*, 117(1-2):47–52.

Phillips, G. N. (1997). Structure and dynamics of green fluorescent protein. *Current Opinion in Structural Biology*, 7(6):821–827.

Piotrowski, M., Janowitz, T., and Kneifel, H. (2003). Plant C-N hydrolases and the identification of a plant N-carbamoylputrescine amidohydrolase involved in polyamine biosynthesis. *The Journal of biological chemistry*, 278(3):1708–12.

Plefka, T. (1982). Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971–1978.

Poorter, L. and Markesteijn, L. (2008). Seedling Traits Determine Drought Tolerance of Tropical Tree Species. *Biotropica*, 40(3):321–331.

Port, M., Tripp, J., Zielinski, D., Weber, C., Heerklotz, D., Winkelhaus, S., Bublak, D., and Scharf, K.-D. (2004). Role of Hsp17.4-CII as coregulator and cytoplasmic retention factor of tomato heat stress transcription factor HsfA2. *Plant physiology*, 135(3):1457–70.

Prochiantz, A. (2011). Homeoprotein Intercellular Transfer, the Hidden Face of Cell-Penetrating Peptides. pages 249–257.

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics (Oxford, England)*, 20(18):3508–15.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S. J. (2007). Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *PROTEOMICS*, 7(6):944–960.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. *Methods*, 24(3):218–229.

Qi, Y., Suhail, Y., Lin, Y.-y., Boeke, J. D., and Bader, J. S. (2008). Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome research*, 18(12):1991–2004.

Qin, L., Bromberg-White, J. L., and Qian, C.-N. (2012). Opportunities and Challenges in Tumor Angiogenesis Research. pages 191–239.

Qin, X. and Zeevaart, J. A. D. (1999). The 9-cis-epoxycarotenoid cleavage reaction is the key regulatory step of abscisic acid biosynthesis in water-stressed bean. *Proceedings of the National Academy of Sciences*, 96(26):15354–15361.

Qiu, D., Morgan, C., Shi, J., Long, Y., Liu, J., Li, R., Zhuang, X., Wang, Y., Tan, X., Dietrich, E., Weihmann, T., Everett, C., Vanstraelen, S., Beckett, P., Fraser, F., Trick, M., Barnes, S., Wilmer, J., Schmidt, R., Li, J., Li, D., Meng, J., and Bancroft, I. (2006). A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. *Theoretical and Applied Genetics*, 114(1):67–80.

Quah, B. J. C., Barlow, V. P., McPhun, V., Matthaei, K. I., Hulett, M. D., and Parish, C. R. (2008). Bystander B cells rapidly acquire antigen receptors from activated B cells by membrane transfer. *Proceedings of the National Academy of Sciences*, 105(11):4259–4264.

Rabbani, M. A., Maruyama, K., Abe, H., Khan, M. A., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. *Plant physiology*, 133(4):1755–67.

Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S. B., Phanse, S., Ceol, A., Häuser, R., Siszler, G., Wuchty, S., Emili, A., Babu, M., Aloy, P., Pieper, R., and Uetz, P. (2014). The binary protein-protein interaction landscape of Escherichia coli. *Nature Biotechnology*, 32(3):285–290.

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586).

Reinhart-King, C. A., Dembo, M., and Hammer, D. A. (2008). Cell-Cell Mechanical Communication through Compliant Substrates. *Biophysical Journal*, 95(12):6044–6051.

Riboni, M., Galbiati, M., Tonelli, C., and Conti, L. (2013). GIGANTEA enables drought escape response via abscisic acid-dependent activation of the florigens and SUPPRESSOR OF OVER-EXPRESSION OF CONSTANS. *Plant physiology*, 162(3):1706–19.

Rizhsky, L., Liang, H., Shuman, J., Shulaev, V., Davletova, S., and Mittler, R. (2004). When Defense Pathways Collide. The Response of Arabidopsis to a Combination of Drought and Heat Stress. *PLANT PHYSIOLOGY*, 134(4):1683–1696.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A.-L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., and Vidal, M. (2014). A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159(5):1212–1226.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178.

Rustom, A. (2004). Nanotubular Highways for Intercellular Organelle Transport. *Science*, 303(5660):1007–1010.

Rustom, A. (2016). The missing link: does tunnelling nanotube-based supercellularity provide a new understanding of chronic and lifestyle diseases? *Open Biology*, 6(6):160057.

Ryan, A. C., Hewitt, C. N., Possell, M., Vickers, C. E., Purnell, A., Mullineaux, P. M., Davies, W. J., and Dodd, I. C. (2014). Isoprene emission protects photosynthesis but reduces plant productivity during drought in transgenic tobacco (Nicotiana tabacum) plants. *New Phytologist*, 201(1):205–216.

Saito, R., Suzuki, H., and Hayashizaki, Y. (2003). Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics (Oxford, England)*, 19(6):756–63.

Saliendra, N., Sperry, J., and Comstock, J. (1995). Influence of leaf water status on stomatal response to humidity, hydraulic conductance, and soil drought in Betula occidentalis. *Planta*, 196(2):357–366.

Samaras, Y., Bressan, R. A., Csonka, L. N., Garcia-Rios, M. G., Paino, D., and Rhodes, D. (1995). Proline accumulation during drought and salinity. *Environment and Plant Metabolism, Bios Scientific Publishers, Oxford*, pages 161–187.

Schrodinger, E. (1992). *What is life? with Mind and matter and Autobiographical sketches*. Cambridge University Press, Cambridge.

Schulze, E. and Hall, A. (1982). Stomatal responses, water loss and CO2 assimilation rates of plants in contrasting environments. *Physiological plant ecology II*.

Schwartz, P. and Randall, D. (2003). An abrupt climate change scenario and its implications for United States national security. Technical report, Jet Propulsion Laboratory Pasadena, CA, Pasadena, CA.

Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., Huang, H.-P., Harnik, N., Leetmaa, A., Lau, N.-C., Li, C., Velez, J., and Naik, N. (2007). Model projections of an imminent transition to a more arid climate in southwestern North America. *Science (New York, N.Y.)*, 316(5828):1181–4.

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Taji, T., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y., and Shinozaki, K. (2002). Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *The Plant Journal*, 31(3):279–292.

Seo, P. J., Lee, S. B., Suh, M. C., Park, M.-J., Go, Y. S., and Park, C.-M. (2011). The MYB96 transcription factor regulates cuticular wax biosynthesis under drought conditions in Arabidopsis. *The Plant cell*, 23(3):1138–52.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology*, 3(1):88.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (1996). Molecular responses to drought and cold stress. *Current opinion in biotechnology*, 7(2):161–7.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (2006). Gene networks involved in drought stress response and tolerance. *Journal of Experimental Botany*, 58(2):221–227.

Singh, A. B. and Harris, R. C. (2005). Autocrine, paracrine and juxtacrine signaling by EGFR ligands. *Cellular Signalling*, 17(10):1183–1193.

Sofo, A., Dichio, B., Xiloyannis, C., and Masia, A. (2004). Lipoxygenase activity and proline accumulation in leaves and roots of olive trees in response to drought stress. *Physiologia Plantarum*, 121(1):58–65.

Son, S.-W., Jeong, H., and Noh, J. D. (2006). Random field Ising model and community structure in complex networks. *The European Physical Journal B*, 50(3):431–437.

Song, K. and Osborn, T. C. (1992). Polyphyletic origins of Brassica napus : new evidence based on organelle and nuclear RFLP analyses. *Genome*, 35(6):992–1001.

Soundarajan, S. and Hopcroft, J. (2012). Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 607, New York, New York, USA. ACM Press.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–9.

Stark, M. J. R. (2001). The awesome power of yeast genetics - practical approaches and recipes for success. *Journal of Cell Science*, 114(14):2551 – 2552.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach,

H., and Wanker, E. E. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome.

Still, S., Sivak, D. A., Bell, A. J., and Crooks, G. E. (2012). Thermodynamics of Prediction. *Physical Review Letters*, 109(12):120604.

Stine, M. J., Wang, C. J., Moriarty, W. F., Ryu, B., Cheong, R., Westra, W. H., Levchenko, A., and Alani, R. M. (2011). Integration of Genotypic and Phenotypic Screening Reveals Molecular Mediators of Melanoma-Stromal Interaction. *Cancer Research*, 71(7):2433–2444.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964.

Su, C., Peregrin-Alvarez, J. M., Butland, G., Phanse, S., Fong, V., Emili, A., and Parkinson, J. (2008). Bacteriome.org–an integrated protein interaction database for E. coli. *Nucleic acids research*, 36(Database issue):D632–6.

Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology*, 4(1):162.

Tanaka, T. (1998). Mean-field theory of Boltzmann machine learning. *Physical Review E*, 58(2):2302–2310.

Thayanithy, V., Dickson, E. L., Steer, C., Subramanian, S., and Lou, E. (2014). Tumor-stromal cross talk: direct cell-to-cell transfer of oncogenic microRNAs via tunneling nanotubes. *Translational Research*, 164(5):359–365.

Trenberth, K. E., Dai, A., van der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J. (2013). Global warming and changes in drought. *Nature Climate Change*, 4(1):17–22.

Twiss, J. L. and Fainzilber, M. (2009). Ribosomes in axons - scrounging from the neighbors? *Trends in Cell Biology*, 19(5):236–243.

Ulbricht, A., Eppler, F. J., Tapia, V. E., van der Ven, P. F., Hampe, N., Hersch, N., Vakeel, P., Stadel, D., Haas, A., Saftig, P., Behrends, C., Fürst, D. O., Volkmer, R., Hoffmann, B., Kolanus, W., and Höhfeld, J. (2013). Cellular Mechanotransduction Relies on Tension-Induced and Chaperone-Assisted Autophagy. *Current Biology*, 23(5):430–435.

Urao, T., Yamaguchi-Shinozaki, K., Urao, S., and Shinozaki, K. (1993). An Arabidopsis myb homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence. *The Plant cell*, 5(11):1529–39.

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., and Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, 9(6):654–659.

Vallabhaneni, K. C., Haller, H., and Dumler, I. (2012). Vascular Smooth Muscle Cells Initiate Proliferation of Mesenchymal Stem Cells by Mitochondrial Transfer via Tunneling Nanotubes. *Stem Cells and Development*, 21(17):3104–3113.

Valverde-Rebaza, J. and de Andrade Lopes, A. (2013). Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, 3(4):1063–1074.

Van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141.

Vikne, H., Gundersen, K., Liestøl, K., Maelen, J., and Vøllestad, N. (2012). Intermuscular relationship of human muscle fiber type proportions: Slow leg muscles predict slow neck muscles. *Muscle & Nerve*, 45(4):527–535.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*.

Wang, J., Bras, R. L., Lerdau, M., and Salvucci, G. D. (2007). A maximum hypothesis of transpiration. *Journal of Geophysical Research: Biogeosciences*, 112(G3):n/a–n/a.

Wang, R.-S., Pandey, S., Li, S., Gookin, T. E., Zhao, Z., Albert, R., and Assmann, S. M. (2011). Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC genomics*, 12(1):216.

Wang, X., Bukoreshtliev, N. V., Gerdes, H.-H., Kettenmann, H., and Klinkert, W. (2012). Developing Neurons Form Transient Nanotubes Facilitating Electrical Coupling and Calcium Signaling with Distant Astrocytes. *PLoS ONE*, 7(10):e47429.

Wani, A. S., Ahmad, A., Hayat, S., and Tahir, I. (2016). Is foliar spray of proline sufficient for mitigation of salt stress in Brassica juncea cultivars? *Environmental science and pollution research international*, 23(13):13413–23.

Weis, S. M. and Cheresh, D. A. (2011). Tumor angiogenesis: molecular pathways and therapeutic targets. *Nature Medicine*, 17(11):1359–1370.

West, G. B., Brown, J. H., and Enquist, B. J. (1999). The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science (New York, N.Y.)*, 284(5420):1677–9.

Wheeler, T. and von Braun, J. (2013). Climate change impacts on global food security. *Science (New York, N.Y.)*, 341(6145):508–13.

Wittig, D., Wang, X., Walter, C., Gerdes, H.-H., Funk, R. H. W., and Roehlecke, C. (2012). Multi-Level Communication of Human Retinal Pigment Epithelial Cells via Tunneling Nanotubes. *PLoS ONE*, 7(3):e33195.

Xin, Z., Zhao, Y., and Zheng, Z.-L. (2005). Transcriptome analysis reveals specific modulation of abscisic acid signaling by ROP10 small GTPase in Arabidopsis. *Plant physiology*, 139(3):1350–65.

Yamada, M., Morishita, H., Urano, K., Shiozaki, N., Yamaguchi-Shinozaki, K., Shinozaki, K., and Yoshiba, Y. (2005). Effects of free proline accumulation in petunias under drought stress. *Journal of Experimental Botany*, 56(417):1975–1981.

Yang, Y., Lai, K., Tai, P., and Li, W. (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *Journal of Molecular Evolution*.

Yang, Z., Liu, J., Tischer, S. V., Christmann, A., Windisch, W., Schnyder, H., and Grill, E. (2016). Leveraging abscisic acid receptors for efficient water use in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24):6791–6.

Yasui, Y. and Kohchi, T. (2014). VASCULAR PLANT ONE-ZINC FINGER1 and VOZ2 repress the FLOWERING LOCUS C clade members to control flowering time in Arabidopsis. *Bioscience, biotechnology, and biochemistry*, 78(11):1850–5.

Yedidia, J. (2001). An idiosyncratic journey beyond mean field theory. *Advanced mean field methods: Theory and practice*.

Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic acids research*, 39(Database issue):D1118–22.

Yoshiba, Y., Kiyosue, T., Katagiri, T., and Ueda, H. (1995). Correlation between the induction of a gene for delta-1-pyrroline-5-carboxylate synthetase and the accumulation of proline in Arabidopsis thaliana under osmotic stress. *The Plant*.

Yoshida, T., Fujita, Y., Maruyama, K., Mogami, J., Todaka, D., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2015). Four A rabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. *Plant, Cell & Environment*, 38(1):35–49.

Yoshida, T., Fujita, Y., Sayama, H., Kidokoro, S., Maruyama, K., Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2010). AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. *The Plant Journal*, 61(4):672–685.

Yung, Y. and Russell, M. (2010). The search for life on Mars. *Journal of Cosmology*, 5:1121–1130.

Zheleva, E., Getoor, L., Golbeck, J., and Kuter, U. (2008). Using Friendship Ties and Family Circles for Link Prediction. In *2nd ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD)*.

Zhou, T., Lü, L., and Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630.

Vita

Yasir Suhail received his B.Tech. degree in Electical Engineering from the Indian Insitute of Technology, Delhi, India and his M.S. degree in Electrical and Computer Engineering from Michigan State University. His initial interests in signal processing and neuronal computations led him to working with neural electrode array recordings. At Johns Hopkins, he started out working on voltage gated Calcium channels and then moved on to bioinformatics, network analysis, short read seqeuncing data analysis. His research interests are in using disparate data sources to predict testable hypotheses and candidate genes via probabilistic and statistical models. He can be reached at yasir.suhail@gmail.com.