STRUCTURES AND BOOLEAN DYNAMICS IN GENE REGULATORY NETWORKS

By

Anthony Szedlak

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics – Doctor of Philosophy

2017

ABSTRACT

STRUCTURES AND BOOLEAN DYNAMICS IN
GENE REGULATORY NETWORKS

By

Anthony Szedlak

Cancer is a complex family of diseases primarily characterized by the accumulation of mutations in cells within multicellular organisms, leading to accelerated proliferation and aggressive competition over limited space and resources. This is coupled with a loss of DNA replication quality control mechanisms, creating genetically heterogeneous clonal lines within patients and even within individual tumors. Each of these clonal lines may respond differently to the same perturbations, making the system naturally resistant to drug therapies. Great progress has been made in identifying and deciphering the roles of driver mutations in individual genes, but the mechanisms of oncogenesis can only be truly understood in the context of the misregulation of dynamical processes in cells' underlying gene regulatory networks (GRNs).

This dissertation discusses the topological and dynamical properties of GRNs in cancer, and is divided into four main chapters. First, the basic tools of modern complex network theory are introduced. These traditional tools as well as those developed by myself (set efficiency, interset efficiency, and nested communities) are crucial for understanding the intricate topological properties of GRNs, and later chapters recall these concepts. Second, the biology of gene regulation is discussed, and a method for disease-specific GRN reconstruction developed by our collaboration is presented. This complements the traditional exhaustive experimental approach of building GRNs edge-by-edge by quickly inferring the existence of as of yet undiscovered edges using correlations across sets of gene expression data. This method also provides insight into the distribution of common mutations across GRNs. Third, I demonstrate that the structures present in these reconstructed networks are strongly related to the evolutionary histories of their constituent genes. Investigation of how the forces of evolution shaped the topology of GRNs in multicellular organisms by growing outward from a core of ancient, conserved genes can shed light upon the "reverse evolution" of

normal cells into unicellular-like cancer states. Next, I simulate the dynamics of the GRNs of cancer cells using the Hopfield model, an infinite range spin-glass model designed with the ability to encode Boolean data as attractor states. This attractor-driven approach facilitates the integration of gene expression data into predictive mathematical models. Perturbations representing therapeutic interventions are applied to sets of genes, and the resulting deviations from their attractor states are recorded, suggesting new potential drug targets for experimentation. Finally, I extend the Hopfield model to modular networks, cyclic attractors, and complex attractors, and apply these concepts to simulations of the cell cycle process. Futher development of these and other theoretical and computational tools is necessary to analyze the deluge of experimental data produced by modern and future biological high throughput methods.

This dissertation is dedicated to my family and my teachers

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Carlo Piermarocchi, supporting me throughout my dissertation work, and for always encouraging me to tilt at my own ideas, be they giants or windmills.

I would also like to thank our past and present collaborators in San Diego: Giovanni Paternostro, Nicholas Smith, Yunyi Kang, Andrew Hodges, and Edison Ong. The interdisciplinary work presented herein is the fruit of their patience.

Importantly, I would like to thank the members of my dissertation committee: David Arnosti, Norman Birge, Lisa Lapidus, and George Mias. Years ago I leveraged their agreeable natures against them and convinced them to join my committee. This dissertation and my oral defense are the last vestiges of their momentary lapses of reason, and the only remaining impediments to their freedom.

Lastly, I would like to thank the National Science Foundation, National Institutes of Health, and the Department of Defense for sponsoring this quixotic misadventure called a doctoral dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

I must warn the reader that this chapter should be read with care, for I have not the skill to make myself clear to those who do not wish to concentrate their attention.

—Jean-Jacques Rousseau*, The Social Contract*

This dissertation covers a broad range of topics arranged in a bottom-up manner. I begin by discussing the fundamentals of network theory and complexity from a general, mathematical point of view. I then focus on the particular case of the gene regulatory networks (GRNs) that regulate the processes in cancer cells, examining both how these networks are inferred from a pool of disease-specific experimental data and the resulting topological traits that emerge. These topological traits are also shown to be strongly related to the evolutionary properties of individual genes and clusters of genes. I then apply a dynamical Boolean model, the Hopfield model, to simulate the effects of specific perturbations to cancer GRNs, suggesting potential therapeutic targets for future experiments. Finally, I introduce a new generalized version of the Hopfield model, and I apply a specific version to simulate perturbations to genes involved in cell cycle.

Because it is likely that most readers will be knowledgeable of only some of the topics discussed herein, this dissertation was written under the assumption that all readers are both highly intelligent and oblivious to all of these things. Unlike Rousseau, I introduce the material in what I believe to be an approachable, edifying, and hopefully entertaining manner. Because of the diversity of topics, it was deemed more natural to include introductions at the opening of each chapter rather than kludging together a single long and meandering introduction. This has the added benefit of creating compartmentalized chapters, allowing readers who are familiar with any introductory material to easily skip to sections containing unfamiliar material. Some of the gory details have been pushed to the appendices, including a mathematically dense proof and two lengthy deriva-

tions. Any reader interested in further details can find them in the original publications referenced in the text. The remaining chapters are organized as follows.

- Chapter 2, *Complex Networks*: an introduction to network theory, including many terms and concepts used in later chapters
- Chapter 3, *Gene Regulatory Networks*: an overview of proteins, genes, and gene regulatory networks, as well as network reconstruction algorithms used in modern bioinformatics
- Chapter 4, *Evolution of Gene Regulatory Networks*: a description of the relationship between the topology of gene regulatory networks and the evolutionary histories of their constituent genes
- Chapter 5, *The Hopfield Model*: a synopsis of the basic properties of one kind of neural network and its application to simulate the dynamics of cancerous gene regulatory networks
- Chapter 6, *Generalized Hopfield Models*: an introduction to three generalizations of the Hopfield model (modular networks, cyclic attractors, and complex attractors) and applications to cell cycle
- Chapter 7, *Conclusions*: a summary of the material covered in this dissertation and a list of work to be done in the future

CHAPTER 2

COMPLEX NETWORKS

> *A culture's icons are a window onto its soul. Few would disagree that, in the culture of molecular biology that dominated much of the life sciences for the last third of the 20th century, the dominant icon was the double helix. In the present, post-modern, 'systems biology' era, however, it is, arguably, the hairball.*
>
> —Arthur Lander, *The edges of understanding*

The following chapter reviews the basics of complex network theory as well as my own contributions, and is organized as follows.

- Section 2.1, *The Königsberg bridge problem*: a synopsis of the founding problem and solution of network theory
- Section 2.2, *Basic terminology*: a review of the basic terminology and notation used in modern network theory
- Section 2.3, *The problem with pictures*: explanation of network layouts and motivation for development of non-Euclidean network measures
- Section 2.4, *Basic network measures*: a review of basic network measures such as degree
- Section 2.5, *Advanced network measures*: a review of advanced network measures such as betweenness centrality and global efficiency
- Section 2.6, *Set and interset efficiency*: definitions of two related measures I developed, the set efficiency and interset efficiency
- Section 2.7, *Communities*: definition of network communities and modularity
- Section 2.8, *Nested communities*: an unpublished generalization of communities which I developed, nested communities

Sections 2.1-2.5 and 2.7 may be safely ignored by readers with basic knowledge of the listed topics.

For detailed definitions and explanations of concepts in canonical network theory, see *Networks: An Introduction* by Mark Newman [110].

2.1 The Königsberg bridge problem

Like much of mathematics, the modern field of graph theory was borne from a simple puzzle. Fig. 2.1 shows a detailed 18^{th} century map from the Prussian city of Königsberg, now named Kaliningrad. The goal was to find a route through the city that crossed each one of its seven bridges exactly once. In 1736, Leonhard Euler issued the first proof in graph theory, which demonstrated that no such route exists.

Euler recognized that most of the information shown in such maps was extraneous. Neither the details of the city (buildings, streets, shape of the shoreline) nor the details of the bridges (location, length, construction materials) were relevant to solving the problem; all that mattered was the manner in which land masses were connected to one another. By reducing each landmass to a point (a *node* or a *vertex*) and each bridge to a line (an *edge* or a *link*) connecting points, he was left with the simplest possible model that could still address the problem. The resulting *graph* is shown in Fig. 2.2 with the nodes labeled A, B, C, and D. He concluded that if such a path (a so called *Eulerian path*) were to exist, there must be exactly 0 or exactly 2 nodes with an odd number of edges (or *degree*). Since all four nodes have odd degree, no Eulerian path exists. See Appendix A for an informal proof.

2.2 Basic terminology

Networks, or *graphs*,[1] provide a basic mathematical architecture for organizing systems with interacting components, and are especially useful for representing large, heterogeneous systems. Nodes typically represent entities in a system, e.g. movie actors, and edges typically represent interactions or relationships between entities, e.g. two actors performing in the same movie.

[1] Some literature draws a distiction between the terms *network* and *graph*, but they will be used interchangeably here.

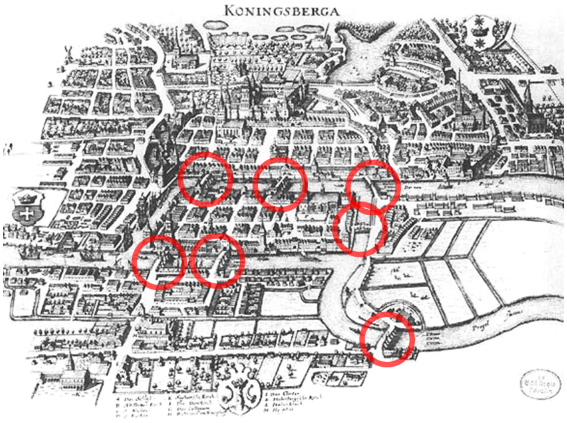


Figure 2.1 Detailed map of Königsberg. For clarity, the bridges have been circled in red.

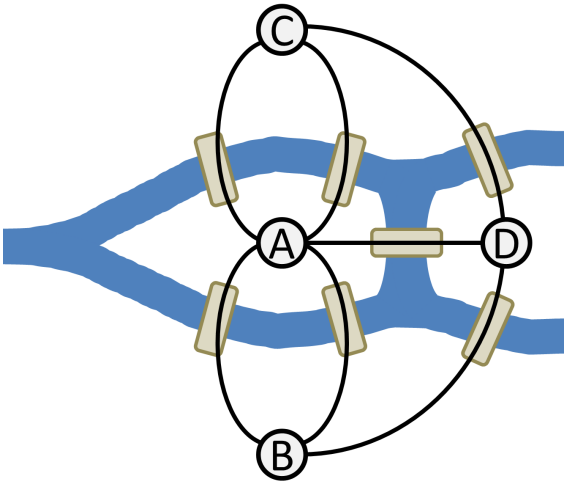


Figure 2.2 Bridges of Königsberg, overlayed with Euler's graphical representation.

There are many ways to represent networks mathematically. Euler used a diagram to solve the Königsberg bridge problem, but this was practical only because the network was relatively small with simple topology. Analyzing large complex networks requires the assistance of a computer, which requires a symbolic rather than geometric representation. The most compact and convenient notation for the purposes of this dissertation is using an *adjacency matrix* A, where

$$A_{ij} = \begin{cases} 1 & \text{if } j \to i \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

and $j \to i$ denotes a directed edge from node j to node i.[2] Note that for undirected networks, $A_{ij} = A_{ji}$. In addition, many real-world networks have *weighted edges*, signifying the strength of the connection. Weighted adjacency matrices are sometimes denoted by W, where $W_{ij} \geq 0$ is the weight of the connection from j to i. For example, the weighted, undirected network shown in Fig. 2.3A has

$$W = \begin{bmatrix} 0 & 2 & 5 \\ 2 & 0 & 1.5 \\ 5 & 1.5 & 0 \end{bmatrix}, \tag{2.2}$$

[2]Some literature and software packages use the convention $i \to j$, but all equations can be translated between the two conventions using appropriate matrix transposes.

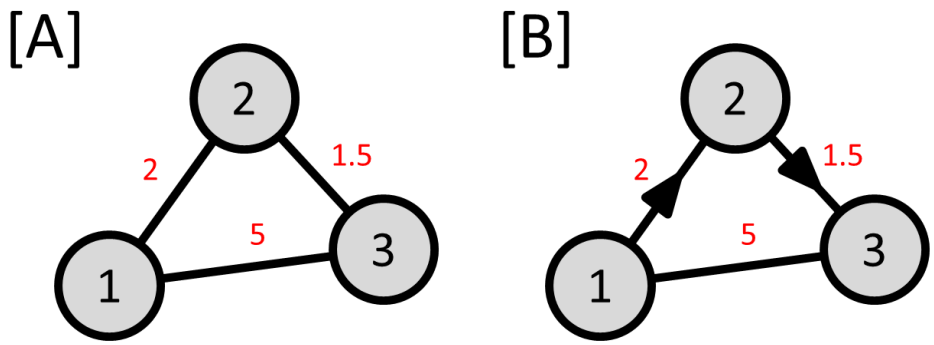


Figure 2.3 [A] A weighted network with undirected edges, and [B] a weighted network with two directed edges and one undirected edge. Node labels are shown in black, and edge weights are shown in red.

and the weighted, directed network shown in Fig. 2.3B has

$$W = \begin{bmatrix} 0 & 0 & 5 \\ 2 & 0 & 0 \\ 5 & 1.5 & 0 \end{bmatrix}. \tag{2.3}$$

Note that in all network diagrams in this dissertation, edges without arrowheads are assumed to be undirected edges, and networks without edge weight labels are assumed to be unweighted networks.

Although a network can be simplistically described as a set of pairwise relationships, information may need to flow between nodes lacking a direct connection, e.g. airline passengers traveling long distances or between small cities may need to make connecting stops, and a request from a personal computer for information from a remote server may need to connect through several intermediaries. This sequence of transfers through a network is called a *path*. Traversing a path usually comes at a cost such as time or money, and intelligent companies and consumers attempt to find paths of minimum cost. In both Fig. 2.3A and B, the shortest path, or *geodesic*, from node 1 to node 3 is the sequence $1 \to 2 \to 3$, and the total distance is $d_{31} = W_{21} + W_{32} = 3.5$. However, the geodesic from 3 to 1 in Fig. 2.3A is $3 \to 2 \to 1$ with a distance of $d_{13} = W_{23} + W_{12} = 3.5$, whereas in Fig. 2.3B the geodesic is $3 \to 1$ with a distance of $d_{13} = W_{13} = 5$.

2.3 The problem with pictures

It can be useful to examine networks diagrammatically, especially for Euclidean networks such as the Königsberg bridge network or a network of airports linked by connecting flights, since the network can be drawn to scale such that pairwise distances between nodes are preserved. As Euler realized, this simplifies analysis by removing unnecessary details contained in a normal map. However, many complex networks such as movie costar networks do not represent spatial relationships and therefore do not have a single "best" spatial layout.

Projections provide a way to reduce high dimensional data to a low dimensional form, and are typically used to help humans visualize the same complex data from multiple simplified perspectives. Perhaps the most familiar examples are projections of the Earth's surface onto a 2D plane, some of which are shown in Fig. 2.4. Maps like these allow the Earth to be comfortably stored on the pages of a book or on the walls of the Pentagon, but while a good projection preserves *some* properties of the original data (like the surface area of countries or the distance between parallels), all projections necessarily introduce distortions and so should be used with caution [1].[3] Only a globe can capture the true topology of Earth's surface, but certain properties can be gleaned from consideration of multiple kinds of projections.

Similarly, *network projections*, also known as *network layouts*, aim to arrange nodes in two or three dimensions such that the spatial distances between nodes is related to a property of the network, often but not always the network distance between nodes. For example, a *force-directed layout* treats all nodes as positively charged masses confined to a plane and treats all edges as springs, and then attempts to find a minimum energy configuration as defined by Coulomb's law and Hooke's law. The Coulomb force pushes all particles apart, whereas the spring forces keep densely connected groups of nodes close. Because there are many energetic minima in such systems, the layout determined by the algorithm is not unique. Other examples include *spectral*, *orthogonal*, *tree*, and *dominance layouts*; and the choice of layout depends on the specific network

[3] thetruesize.com is an interactive tool that allows the user to move silhouettes of countries on a Mercator map, demonstrating how country shapes and sizes morph as they are translated northward and southward.

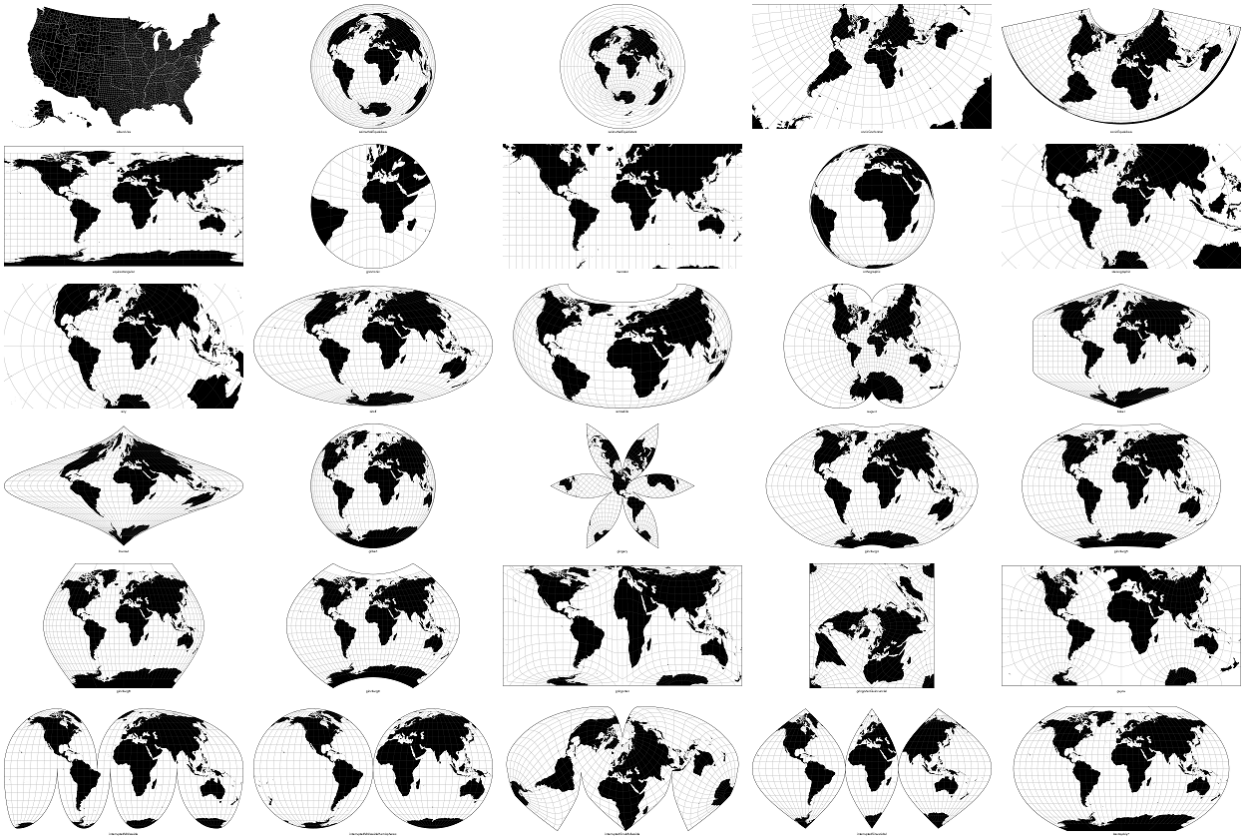


Figure 2.4 Some of the countless available map projections, including both common and exotic projections. Image adapted from here.

and the specific application. Some example layouts are shown in Figs. 2.5–2.8.

By virtue of most networks being both high dimensional and non-Euclidean, however, network layouts tend to suffer greater distortion than maps of the Earth, which can produce misleading results. Even the normal notion of metrics breaks down. In Euclidean space, distances are symmetric, i.e. $d(\vec{x},\vec{y}) = d(\vec{y},\vec{x})$; but this is not necessarily true for distances between nodes in directed networks. It is important to remember that network layouts provide primarily qualitative information. In order to understand the structure of complex networks quantiatively, new sorts of non-Euclidean measures must be defined.

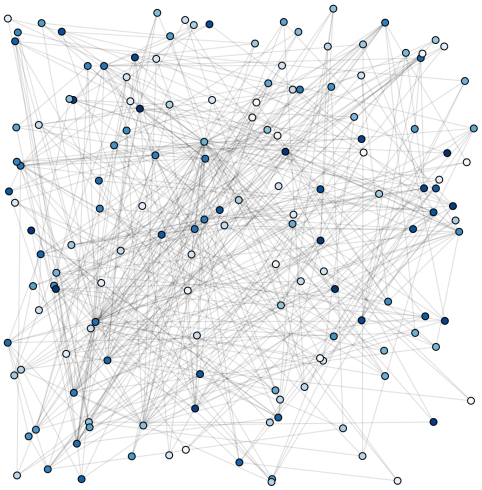


Figure 2.5 A random layout of a randomly generated scale-free network. Nodes are colored from white (low degree) to dark blue (high degree).

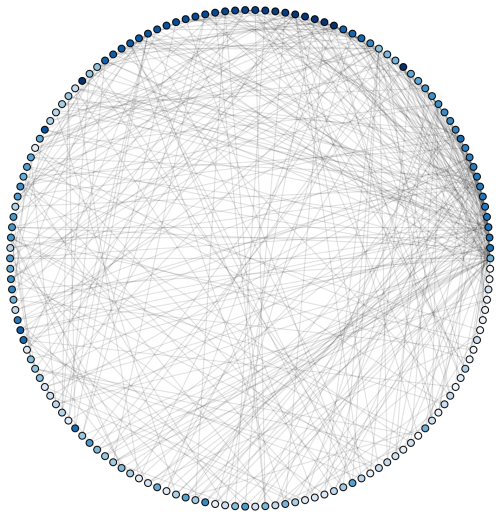


Figure 2.6 A circular layout of the same network shown in Fig. 2.5.

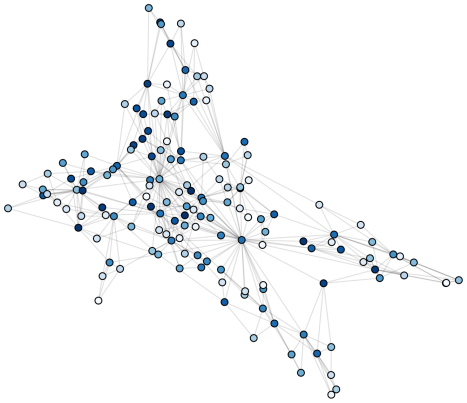


Figure 2.7 A force-directed layout of the same network shown in Fig. 2.5.

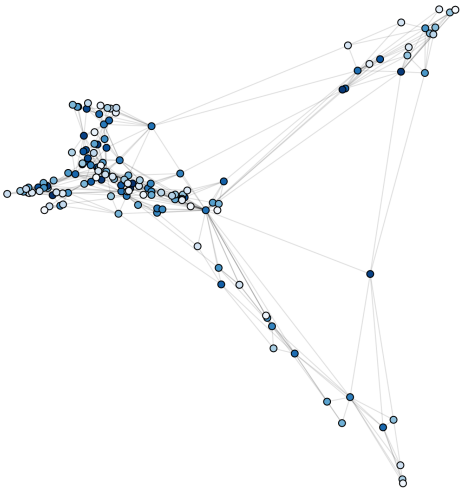


Figure 2.8 A spectral layout of the same network shown in Fig. 2.5.

2.4 Basic network measures

A natural question to ask is, how does a particular node compare with another node, or how does one node compare with all other nodes? What makes a given node unique, if anything?

One of the most basic measures is the *degree* of a node i, which is simply the number of edges connected to i,

$$k_i = \sum_j A_{ij} \tag{2.4}$$

Because nodes in directed networks have both incoming and outgoing edges, the measure splits into *indegree*

$$k_i^{\text{in}} = \sum_j A_{ij} \tag{2.5}$$

and *outdegree*

$$k_i^{\text{out}} = \sum_j A_{ji} = \sum_j \left(A^{\text{T}}\right)_{ij} \tag{2.6}$$

for the number of upstream and downstream neighbors, respectively. Note that nodes with zero indegree are termed *sources*, and nodes with zero outdegree are termed *sinks*. For weighted networks, the role of the degree is played by the *strength*, which is the sum of the weights of the edges connected to a node,

$$s_i = \sum_j W_{ij} \tag{2.7}$$

For weighted, directed networks, the *instrength* and *outstrength* are given by

$$s_i^{\text{in}} = \sum_j W_{ij} \tag{2.8}$$

and

$$s_i^{\text{out}} = \sum_j W_{ji} = \sum_j \left(W^{\text{T}}\right)_{ij} \tag{2.9}$$

respectively. The degree or strength of one node relative to another gives some indiciation of the nodes' relative importance. For example, an airport-to-airport network could be constructed by

$W_{ij} =$ number of people in one year who departed from airport j and arrived at airport i

This construction means the total number of people who left airport i in one year is given by the outstrength s_i^{out}. By most reasonable definitions of "importance," the Harsfield-Jackson Atlanta International Airport, which served over 49 million departing passengers in 2015, is more important than the Houghton County Memorial Airport, which served less than 26,000 [45].

Similarly, one may be interested not in comparing the properties of two nodes within the same network, but one network with another. A quantitative comparison between the topologies of two complex networks, particularly large networks, can be quite difficult. Even discerning whether two networks are *isomorphic* (i.e. whether two networks are identical under a relabeling of one of the network's nodes), the so-called *graph isomorphism problem*, is an active area of research.

Although each network has its own unique topology, distilling its complex microscopic details into a handful of coarse-grained measures that capture its basic properties can be enlightening. The simplest way to do this is to examine the network's *degree distribution*. A convenient model network is the *Erdős-Rényi* network, which is constructed by including an edge $j \to i$ with probability p for each node pair (i, j). This is a simple type of *random network model* that is both interesting in its own right and serves as a control for studying the properties of real networks. Because the presence or absence of each edge is an independent trial with a fixed probability of success p, Erdős-Rényi networks have binomial degree distributions,

$$P_k = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2.10}$$

as shown in Fig. 2.9. This means that *hubs*, nodes with much higher degree than the mean degree, are rare because P_k decreases exponentially for $k \gg \langle k \rangle = np$.

However, real networks rarely have binomial degree distributions. One of the most prevalent distributions is the *power law degree distribution* which has $P_k \sim k^{-\gamma}$ for some $\gamma > 1$, and typically $2 \lesssim \gamma \lesssim 3$ for real networks. Although power law distributions also fall to zero as $k \to \infty$, they do so much more slowly than binomial distributions, resulting in many more hubs than in Erdős-Rényi networks. Fig. 2.10 shows the degree distribution for the Enron email network in which an edge $j \to i$ means that employee j sent at least one email to employee i. Plotting the histogram in log-log scale reveals that the Enron email network has a power law degree distribution with $\gamma \approx 1.6$.

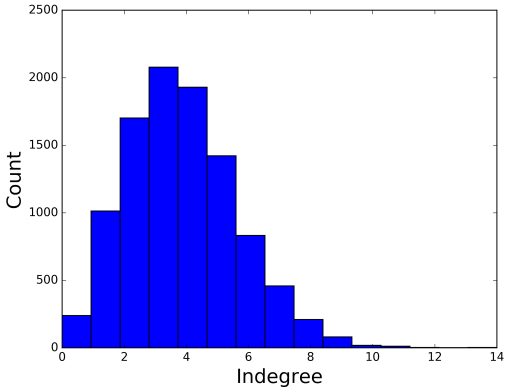


Figure 2.9 Indegree distribution for an Erdős-Rényi network with $\langle k^{\text{in}} \rangle \approx 3.69$.

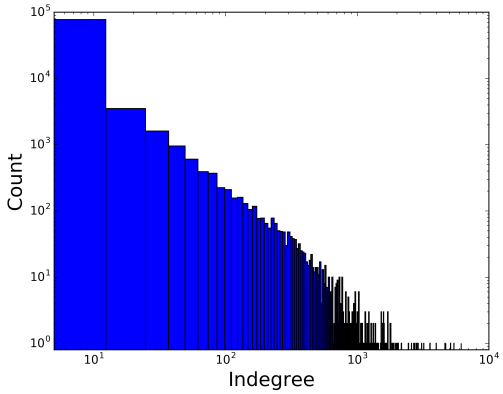


Figure 2.10 Indegree distribution for Enron's email network with $\langle k^{\text{in}} \rangle \approx 3.69$ and $\gamma \approx 1.6$. Note that both axes are logarithmic.

There are many other local network measures besides degree. The *assortativity coefficient* [111] is the Pearson correlation between the degree of neighboring nodes across the network, the value of which determines whether the network is assortative (hubs tend to connect to other hubs), disassortative (hubs tend to connect to non-hubs), or mixed (no significant relationship between the degree of neighboring nodes). The *clustering coefficient* [157] counts the number of "triangles" (fully connected node triplets) in the network out of the total number of possible triangles.

2.5 Advanced network measures

All of the previously discussed measures are entirely local. They provide important information about microscopic properties, but typically cannot identify high level structural properties. For example, the degree distribution, assortativity coefficient, and clustering coefficient all fail to differentiate between the networks in Fig. 2.11A and 2.11B, even though there is an obvious, critical difference: while there is a path from every node to every other node in the network in 2.11A, the network in Fig. 2.11B is composed of two *components*.

Many intermediate- and long-range measures have been developed to understand important structures in networks. One common class of measures, *centrality measures*, assigns a score to each node in the network and identifies nodes that are in some sense "important" or in the "middle"

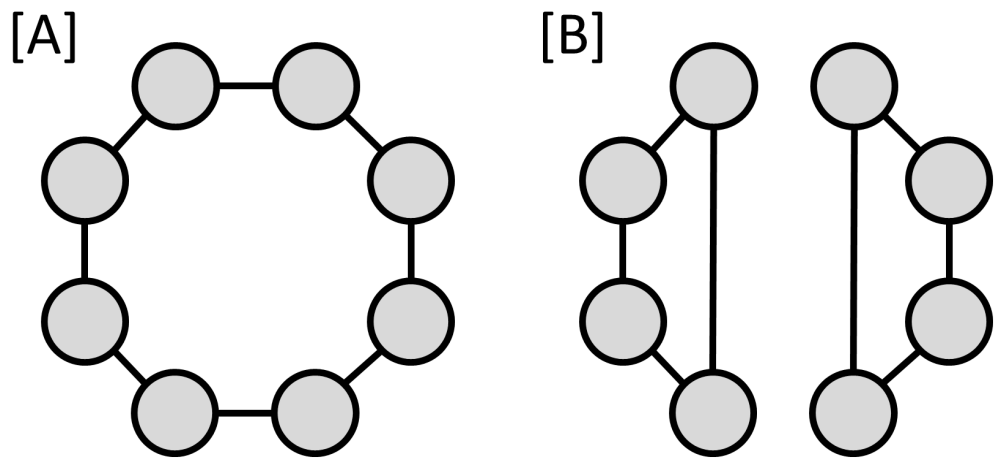


Figure 2.11 (A) A network with one component, and (B) a network with two components. Both networks have identical degree distributions since all nodes have degree 2.

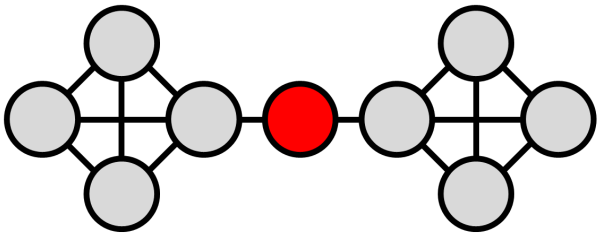


Figure 2.12 An example network. The red node has high betweenness centrality, since all paths between the left and right sides must pass through the red node.

of the network. The *betweenness centrality* [54], for example, counts the fraction of geodesics that pass through a given node. It is given by

$$b_i = \frac{1}{(N-1)(N-2)} \sum_{\substack{j=1..N \\ h=1..N \\ i\neq j, i\neq h, j\neq h}} \frac{\rho_{hj}(i)}{\rho_{hj}} \tag{2.11}$$

where ρ_{hj} is the number of geodesics from node j to node h, and $\rho_{hj}(i)$ is the number of geodesics from j to h that pass through node i. The motivation behind the definition of betweenness centrality is that removal of a node with high betweenness centrality disrupts important avenues of communication between remote nodes. The red node in Fig. 2.12, for example, has high betweenness centrality because all paths between the left and right sides must pass through it.

PageRank [118] is one of the most famous and widely used centrality measures. It was created by Larry Page in 1999 in an effort to "bring order to the web," and served as the foundation of Google's search engine. Like other search engines, Google identifies a large set of pages containing text that matches all or part of the user's query. However, Google's power comes from integrating text matching information with network topology to intelligently sort the search results. The World Wide Web is composed of pages (nodes) and hyperlinks that point from one page to another (edges), and so can be represented as a directed network. The basic assumption behind PageRank is that a page's reputation is a function of its upstream neighbors' reputations, a sort of "reputation by consensus." In other words, if many trustworthy sources of information point to a page i, then i is likely trustworthy as well. This effect should be tempered by the outdegree of those pages, however. Wikipedia's domain hosts a large number of links to other domains, so while Wikipedia may have high PageRank, the algorithm should be designed such that the score it passes downstream is inversely proportional to the number of pages to which it links. Formally, node i's PageRank x_i is obtained by self-consistently solving

$$x_i = \alpha \sum_{j=1..N} A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta \tag{2.12}$$

for $\vec{x}$ given constants α and β. Under this definition, the PageRank algorithm assigns a score to each node based on the scores of its upstream neighbors divided equally amongst all outgoing links, plus a small free score so that source nodes contribute a nonzero score to their targets. Since the primary goal of centrality measures is to rank nodes from most to least central, overall multiplicative factors are irrelevant, so dividing both sides of Eq. 2.12 by β and renaming $x_i/\beta \to x_i$ gives a form of PageRank with only one free parameter α,

$$x_i = 1 + \alpha \sum_{j=1..N} A_{ij} \frac{x_j}{k_j^{\text{out}}} \tag{2.13}$$

Conventionally, α is set equal to 0.85, and was likely empirically chosen by Google simply because it produces the best results [110]. There is no guarantee that $\alpha = 0.85$ is the best choice for all applications, however. Google no doubt has improved the performance of its search engine since the company was founded, but PageRank remains an integral part of the algorithm. There are

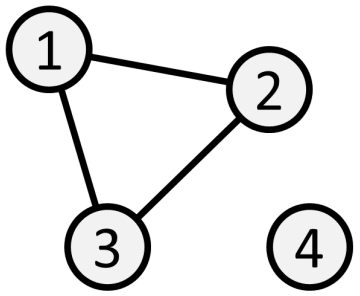


Figure 2.13 An undirected, unweighted, disconnected network with two components.

many variations of PageRank [149], but a simple weighted version of Eq. 2.13 can be obtained by replacing A_{ij} with W_{ij} and the outdegree k_i^{out} with the outstrength s_i^{out},

$$x_i = 1 + \alpha \sum_{j=1..N} W_{ij} \frac{x_j}{s_j^{\text{out}}} \tag{2.14}$$

Since there can only be a nonnegative integer number of hyperlinks from a page j to a page i, the World Wide Web has $W_{ij} \in \{0, 1, 2, \dots\}$; but in general, Eq. 2.14 can be computed and simply interpreted as long as W_{ij} is nonnegative and real. As with all network measures, the "correct" form of PageRank depends on the particular application.

One relatively simple yet potentially informative measure is the mean distance between all pairs of nodes in a network,

$$\bar{d} = \frac{1}{N(N-1)} \sum_{\substack{i=1..N \\ j=1..N \\ i \neq j}} d_{ij} \tag{2.15}$$

where d_{ij} is the geodesic distance from node j to node i, and $d_{ij} \equiv \infty$ if no path exists from j to i. The limiting cases produce sensible results: unweighted complete networks have $\bar{d} = 1$, since all nodes are adjacent to all other nodes, and edgeless networks have $\bar{d} = \infty$. But real networks are rarely this regular. Consider the simple network in Fig. 2.13. For the subnetwork composed of nodes 1, 2, and 3, $\bar{d} = 1$; but including node 4 means $\bar{d} = \infty$. This network is clearly more connected than an edgeless network, however, and a different measure is required to capture this fact.

The *global efficiency* of a network is defined as

$$E_{\text{global}} = \frac{1}{N(N-1)} \sum_{\substack{i=1..N \\ j=1..N \\ i \neq j}} \frac{1}{d_{ij}}. \tag{2.16}$$

This is the mean inverse distance between all distinct ordered pairs of nodes in the network. The inverse of the global efficiency gives a measure similar to the mean distance, and it is identical in the extreme cases: for unweighted complete networks, $E_{\text{global}}^{-1} = \bar{d} = 1$, and for edgeless networks, $E_{\text{global}}^{-1} = \bar{d} = \infty$. However, the global efficiency is less sensitive to disconnected and weakly connected pairs of nodes than the mean distance. For the network shown in Fig. 2.13, $E_{\text{global}}^{-1} = 2$, whereas $\bar{d} = \infty$. Any network with at least one (directed or undirected) edge with nonzero weight has $E_{\text{global}}^{-1} < \infty$.

2.6 Set and interset efficiency

Most commonly used network measures concern either global properties (like the assortativity coefficient, clustering coefficient, and global efficiency) or single-node properties (like degree, betweenness centrality, and PageRank). However, one may be interested in the properties of a set of nodes, for example how a set of nodes is distributed across a network. In [116], I defined the *set efficiency*[4] of a set of nodes I as

$$E_I = \frac{1}{|I|(|I|-1)} \sum_{\substack{i,j \in I \\ i \neq j}} \frac{1}{d_{ij}} \tag{2.17}$$

where $|I|$ is the number of nodes in I. Note that if I is the set of all nodes in the network, $E_I = E_{\text{global}}$. While the summation over i and j is restricted to a subset of the nodes in the network, the geodesic connecting j to i may pass through nodes not in I. The set efficiency measures the proximity of a set of nodes embedded in a networks, where large E_I implies short distances between nodes in set I. The red nodes in Fig. 2.14 are scattered across the network and so have

[4]The *set* efficiency was originally termed the *intraset* efficiency in [116], but its name was changed to the *set* efficiency in [142] to avoid confusion with the *interset* efficiency.

low set efficiency, whereas the red nodes in Fig. 2.15 are separated by only one or two edges and so have high set efficiency.

In [142], I defined the *interset efficiency* from a set of nodes J to a set of nodes I as

$$E_{IJ} = \frac{1}{|I||J| - |I \cap J|} \sum_{\substack{i \in I \\ j \in J \\ i \neq j}} \frac{1}{d_{ij}} \tag{2.18}$$

where $I \cap J$ is the intersection between I and J. As with the set efficiency, geodesics may pass through any nodes in the network; the only constraint is that the geodesic begins in J and ends in I. For example, the interset efficiency from the red nodes to the blue nodes is low in Fig. 2.16 and high in Fig. 2.17. The interset efficiency is a generalization of the set efficiency, since $E_{II} = E_I$. See Appendix B for the derivation of the interset efficiency's normalization. Applications of the set and interset efficiency will be discussed in Sections 3.4 and 4.3.

It is worth emphasizing that the set efficiency is a property of a set of nodes, and the interset efficiency is a property of a set of two sets of nodes. In other words, they are both collective properties of the elements of a set, not merely an average of single-node properties. One possible set-to-set analogue of the betweenness centrality could be the *interset betweenness centrality*,[5]

$$b_i^{HJ} = \frac{1}{\text{Normalization}} \sum_{\substack{j \in J, h \in H, \\ i \neq j, i \neq h, j \neq h}} \frac{\rho_{hj}(i)}{\rho_{hj}} \tag{2.19}$$

This is the fraction of all geodesics that start in node set J and end in node set H that pass through node i. As with the set and interset efficiency, the only difference between Eq. 2.11 and Eq. 2.19 is the limits of the summation. Indeed, many single-node and global network measures can be adapted to answer questions about sets. This set-oriented approach may be more appropriate when asking questions about "robust" networks, where redundant connections emerge (as with metabolic, gene regulatory, and ecological networks) or are intelligently designed (as with transportation, power, and information networks) such that the failure of a single node does not cause

[5]The author leaves the derivation of the normalization and the task of thinking of a better name for Eq. 2.19 as an exercise for Carlo's next graduate student.

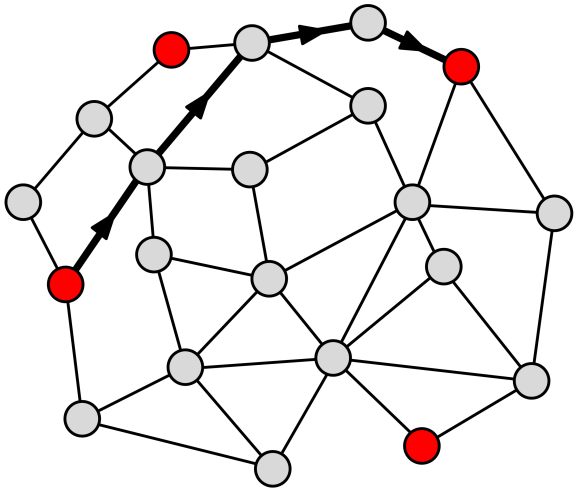


Figure 2.14 The set efficiency of the set of red nodes is relatively low. One geodesic for one (i, j) pair is shown with thick edges.

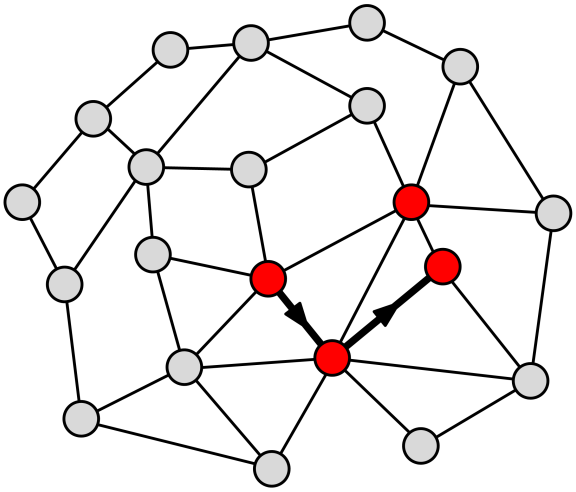


Figure 2.15 The set efficiency of the set of red nodes is relatively high. One geodesic for one (i, j) pair is shown with thick edges.

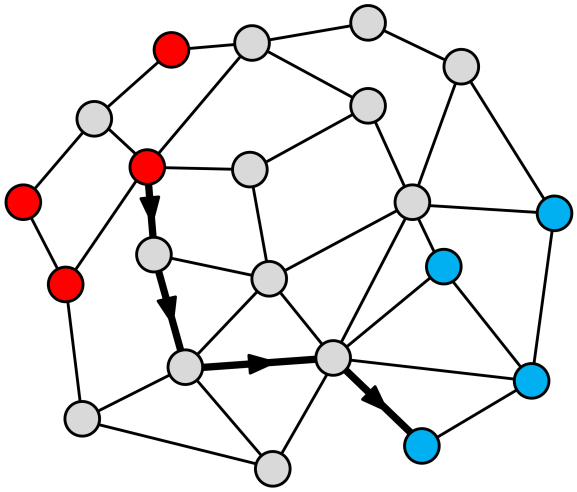


Figure 2.16 The interset efficiency from the set of red nodes to the set of blue nodes is relatively low. One geodesic for one (i, j) pair is shown with thick edges.

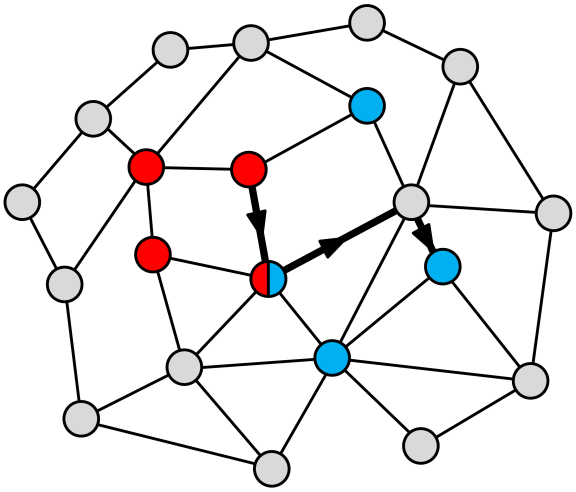


Figure 2.17 The interset efficiency from the set of red nodes to the set of blue nodes is relatively high. The half-blue-half-red node belongs to both sets. One geodesic for one (i, j) pair is shown with thick edges.

dramatic changes to the system (failure/removal of the red node in Fig. 2.12, for example, engenders such a dramatic change, since the network separates into two components). It may be more useful to inquire about the behavior of "regions" of the network than individual nodes.

2.7 Communities

The previously discussed set measures are motivated by questions of the form, "what properties does a given set of nodes have?" *Clustering* in networks, by contrast, is motivated by questions of the form, "can a set of nodes which satisfies a given property be identified?" Answers to the second question are usually far more difficult to compute because the combinatorics of the problem makes exhaustive searches exponentially harder for every node added to the network. Most clustering algorithms employ efficient heuristic techniques to find good solutions, but truly optimal solutions are rarely obtained. Like maps of Earth and network layouts, clustering is a dimensionality-reduction method that gives researchers a bird's eye view of how a network is structured without the need to understand the properties of every individual node and edge.

One common feature of real world networks is the organization of nodes into *communities*, also known as *modules*. The quantitative definition of what constitutes a community varies, but most are designed to segregate the network into relatively densely connected subnetworks. The most common method for community detection is maximizing an objective function $Q(\vec{c})$ called the *modularity* [112],

$$Q = \sum_{ij} \left(A_{ij} - \frac{k_i^{\text{in}} k_j^{\text{out}}}{m} \right) \delta_{c_i, c_j} \tag{2.20}$$

where $m = \sum_i k_i^{\text{in}} = \sum_i k_i^{\text{out}}$ is the total number of edges in the network, c_i is the index of the community to which node i belongs, and $\delta_{x,y}$ is the Kronecker delta function. The first term in the sum is the adjacency matrix, A_{ij}, which is the actual number of edges from j to i (either 0 or 1). The second term, $k_i^{\text{in}} k_j^{\text{out}}/m$, is the expected number of edges from j to i given their out- and indegrees, respectively, under a *degree-preserving randomization*. Finally, the Kronecker delta function ensures that a given (i, j) term contributes to the summation if and only if i and j are

assigned to the same community. Maximizing Q means searching for a $\vec{c}$ that includes as many positive terms in the summation as possible, i.e. searching for groups of nodes that share more edges within each group than expected at random. A weighted version of Eq. 2.20 can be obtained by replacing A_{ij} with W_{ij} and the in/outdegree with the in/outstrength. Applications of modularity will be discussed in Chapter 4.

2.8 Nested communities

The following section contains unpublished findings.

Hierarchical organization is a ubiquitous feature of large complex systems. *Phylogenetic trees*, for example, represent the difference between the genomes of organisms that share a common ancestor, and are usually represented with *dendrograms* as shown on the left side of Fig. 2.18. The difference between genomes is computed using a metric such as the total number of mismatched base pairs (i.e. the number of *single nucleotide polymorphisms*, or SNPs) when the DNA sequences of two organisms are compared. In the phylogenetic tree in Fig. 2.18, this metric has been mapped to the estimated point in time at which two distinct genomes emerged from a common ancestor, which is reflected by the x-coordinate of the forking point between any two *clades* (subnetworks of a phylogenetic tree formed from a chosen forking point plus all its decendants), with more distant forks (very different genomes) placed toward the left. The causes of the divergence in human (and Neanderthal) genomes are many and varied. Geographical divides between prehistoric populations caused humans to evolve in a compartmentalized fashion, with great distances, bodies of water, mountains, and deserts separating major clades, e.g. Sub-Saharan Africans and Aboriginal Australians, and more modest barriers between genetically similar subgroups, e.g. English and French. The evolution of languages is similarly hierarchical. Fig. 2.19 shows an inferred dendrogram of 103 Indo-European languages. In addition to the best estimates of the fork locations, this figure also includes 95% confidence intervals (marked with light blue lines), and the gray distribution on the left corresponds to the computed probability density of the location of the root node (the most recent common ancestor language of all 103 languages shown). Social systems also

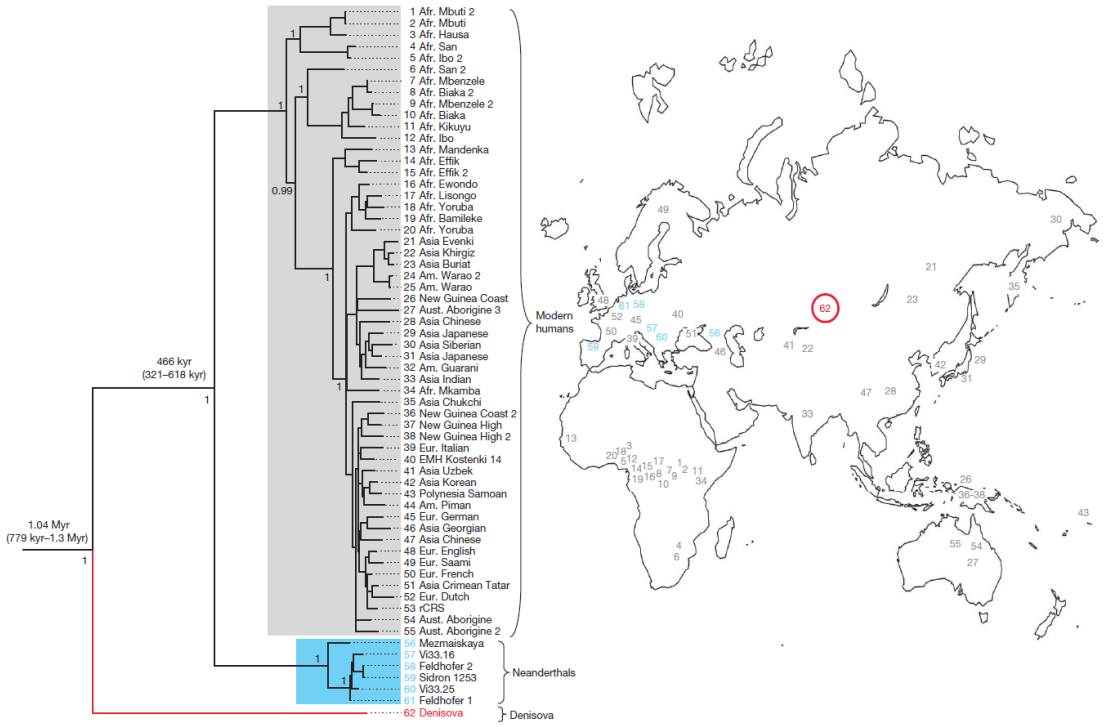


Figure 2.18 An example phylogenetic tree of human ancestry derived from mitochondrial DNA. Original image from [80].

tend to be organized hierarchically into city, county, state, and federal governments, as well as a multitude of international organizations like the United Nations and the World Trade Organization. Although there is great complexity across all scales of such hierarchies, their clades are not necessarily self-similar, as is the case with fractals. Each level of the hierarchy may have its own distinct topology.

Likewise, numerous real world networks exhibit hierarchical structures [91]. Many of the multitude of existing network measures (geodesic distances, for example) may be directly used or adapted to build dendrograms, and for some networks, this may prove illuminating. However, dendrograms are typically constructed using continuous variables for the locations ("heights") of the forking points, and for large networks, the resulting dendrograms are highly complex. Futhermore, assigning nodes to discrete clusters requires a user-defined threshold height that collapses all subsets below this threshold (i.e. clades which are similar enough) to a single point, making this

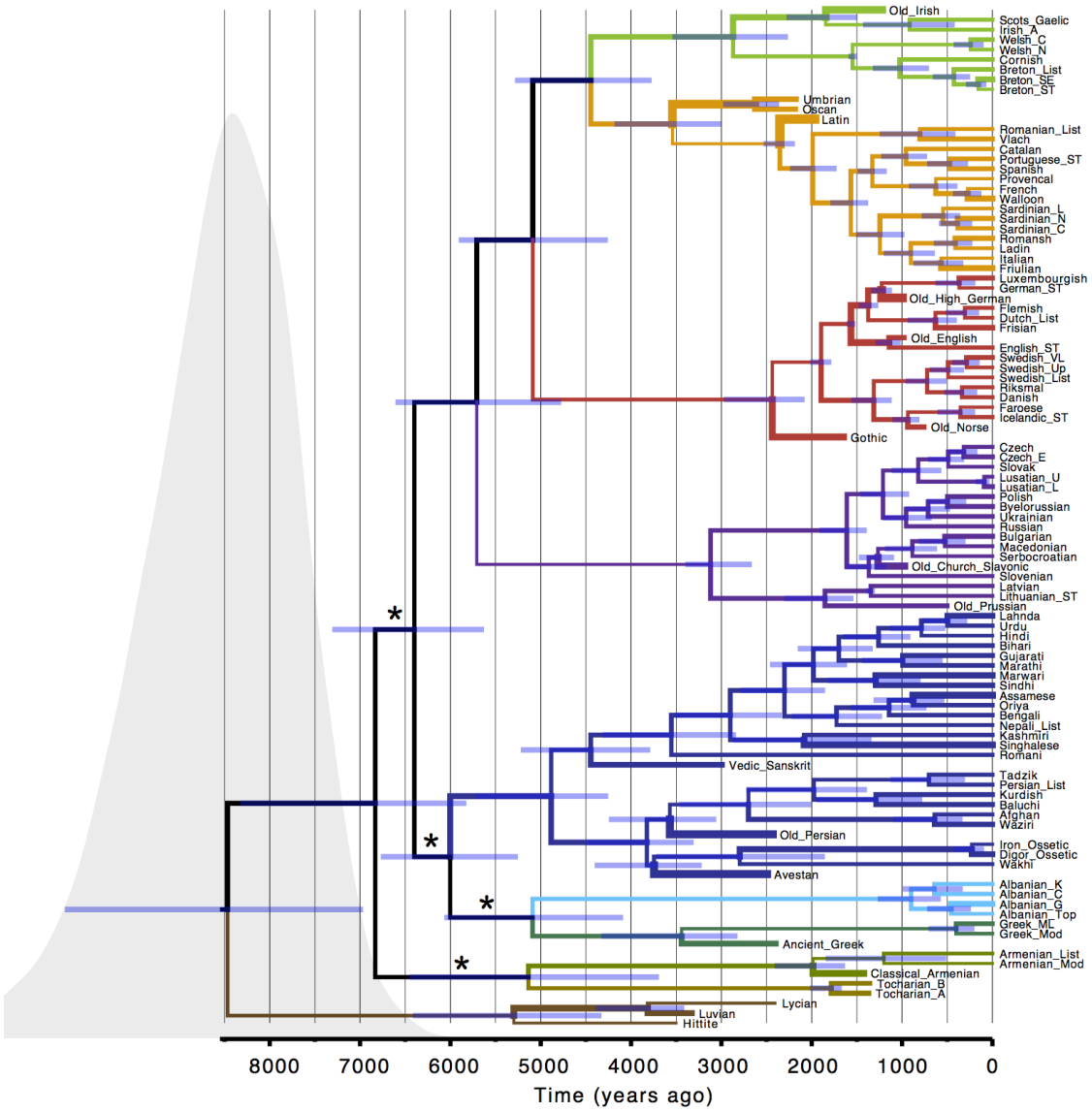


Figure 2.19 Inferred dendrogram of 103 Indo-European languages. Line thicknesses correspond to languages' rates of "diffusion" across land masses. 95% confidence intervals are marked with light blue lines and are included at all forking points. The gray distribution on the left corresponds to the computed probability density of the location of the root node (the most recent common ancestor language of all 103 languages shown). Original image from [19].

a parametric clustering method.

My simple but powerful nonparametric generalization of the concept of communities is *nested communities*, which as its name suggests involves searching for communities inside of communities in a recursive manner. The first step of the algorithm applies the traditional community finding algorithm to the full network, breaking it into standard communities. The nested community algorithm then independently applies the same traditional community finding algorithm to the subnetworks composed of the nodes from each community from the previous step. This method is iterated over each output set of communities until some condition is met, terminating that particular branch of the tree. The algorithm must terminate at the single-node level, but depending on the application, additional criteria can be used to decide when to break from the algorithm. These could include reaching a depth where the optimal modulary becomes nonpositive; reaching a depth where the optimal modularity is no longer statistically signficant compared to a distribution of random subdivisions using a chosen p-value threshold; or using external, non-topological information (such as gene enrichment, discussed in Section 3.2).

In more mathematical notation, applying the traditional community search algorithm f to C (the set of all nodes in the network) results in n communities, $f(C) = \{C_1, C_2, \ldots, C_n\}$. Note that $\bigcup_i C_i = C$ and $C_i \cap C_{i'} = \{\varnothing\}$ for all $i \neq i'$. The next level of depth is obtained by applying f to each of the n communities detected. For community i, for example, its n' subcommunities $\{C_{ij}\}$ are found via $f(C_i) = \{C_{i1}, C_{i2}, \ldots, C_{in'}\}$, and the same union and intersection properties from the first output set apply to these output sets. This is applied iteratively until no further communities are detected. Note that a community's depth is equal to its number of indicies. Fig. 2.20 shows an example network with two labelled levels of community structure.

Figs. C.1-C.4 show *spy plots* of the adjacency matrix for Google's internal web page network [119]. Each node (row and column indicies) is a web page owned by Google, and each edge (a black dot located at (i, j)) is a hyperlink from page j to page i, both of which are managed by Google (google.com, mail.google.com, calendar.google.com, etc.). Nodes were assigned a random index in Fig. C.1, so no clear patterns emerge. Fig. C.2 shows the same network, but with

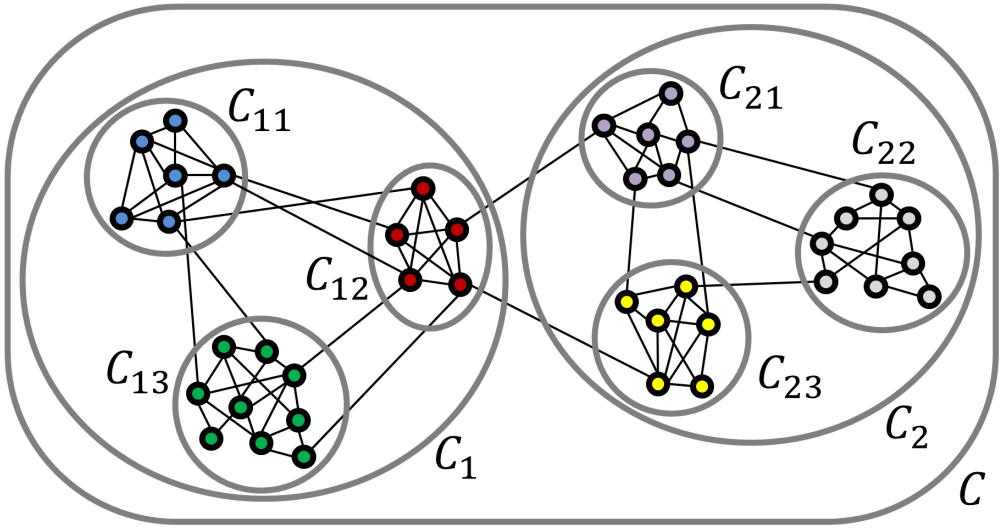


Figure 2.20 An example network with two levels of community structure. The largest oval encloses all nodes, C; the two midsized ovals enclose nodes in level-1 communities, $\{C_i\}$; and the six small ovals enclose nodes in level-2 communities, $\{C_{ij}\}$.

its nodes sorted into standard level-1 communities (boxed in red). Note that this ordering tends to drive edges toward the diagonal, leading to clear block stucture. Figs. C.3 and C.4 show level-2 and level-3 communities respectively, the result of further iterations of the nested community algorithm. This reveals that a significant amount of substructure is overlooked by standard community searches. Figs. C.5-C.8 are the same as Figs. C.1-C.4, but zoomed to show only the first $5,000$ nodes to emphasize the patterns detected by the nested community algorithm.

To further demonstrate the power of the nested community algorithm, the same analysis was performed for an arXiv theoretical high energy physics citation network [90], shown in Figs. C.9-C.16. In this network, each node is an author of at least one high energy physics publication posted on arXiv, and an edge from author j to author i means that at least one of j's articles on arXiv cited at least one of i's articles on arXiv. These plots show more intercommunal edges than the Google network, with dense off-diagonal blocks indicating strongly coupled but still distinct subfields.

Finally, gene regulatory networks such as AML 2.3 and HumanNet (discussed in detail in Chapters 3 and 4) show 7 and 8 layers of nested community structure, respectively, as shown in Figs. C.17-C.20. The communities in these networks – especially HumanNet – show greater

heterogeneity than the Google and citation networks, perhaps because these networks are inferred rather than directly and precisely measured. However, a distribution of communities with high and low connectivities interconnected in more diverse ways may be an important property for cells. Some of the more fundamental cellular functions may need to be tightly regulated to keep the cell alive, whereas others may need to be more flexible to quickly respond to changing stresses. This could be addressed in the future using enrichment tools such as DAVID (discussed in Section 3.2) to determine if there is any relationship between the connectivity of communities in these networks and the kinds of pathways they host.

It is perhaps unsurprising that the Google, citation, and gene regulatory networks show heterogeneous community structures: the rules that govern how these networks are designed (in the case of Google's pages) or emerge (in the case of citations and gene regulation) are very different from one another. However, the nested community algorithm, a rather simple extension of the standard algorithm, successfully identifies deeper underlying trends with little additional effort.

CHAPTER 3

GENE REGULATORY NETWORKS

Quis custodiet ipsos custodes?

(Who will guard the guards themselves?)

—Juvenal, *Satire VI*

The following chapter briefly reviews the fundamental concepts of biology underpinning my research as well as my own contributions, and is organized as follows.

- Section 3.1, *Basic biology*: a basic description of proteins and genes
- Section 3.2, *Gene regulation*: a discussion of how gene expression is regulated
- Section 3.3, *Cancer*: brief overview of gene regulatory networks and oncogenesis
- Section 3.4, *GRN reconstruction*: a network reconstruction algorithm that our collaboration developed and published, and my primary contribution to the publication concerning an analysis of the topology and distribution of cancer mutations in a reconstructed acute myeloid leukemia network

Sections 3.1-3.3 may be safely ignored by readers with basic knowledge of the listed topics. For further details concerning mechanisms in cellular biology and genetics, see introductory texts such as [24] and [77].

3.1 Basic biology

Many cellular functions involve *proteins*, macromolecules dubbed the "workhorses of life." Proteins are produced by the cell in a wide array of shapes and sizes (one of the smallest known natural proteins, spoVM, has a mass less than 3 kDa [33], and the largest, titin, has a mass over 3 MDa [117]) and perform a huge variety of vital tasks. Some of the broad categories of proteins include [24]

- *structural proteins*, which provide the materials for structures like histones (the protein complexes which package DNA), cytoskeletons (the scaffolding within cells that maintains their shape and provides "highways" for the movement of material), and large-scale structures such as hair, spider silk, and connective tissues;
- *transport proteins*, which transport material within a cell, between the cell and its environment, or from one part of the body to another (such as hemoglobin, which moves oxygen throughout the body);
- *hormonal proteins*, which transmit information throughout the body and regulate many processes;
- *receptor proteins*, which detect chemicals in the environment surrounding cells;
- *contractile proteins*, which are used in both the motion of single-celled organisms (as in the case of cilia and flagella) as well as the motion of multicellular organisms (such as actin and myosin in muscles);
- *defensive proteins*, which combat bacteria and viruses;
- and *enzymatic proteins*, which alter the rates of chemical reactions within cells.

The information for the structure of each protein is encoded in one or multiple *genes*. New proteins are created within the cell via *protein biosynthesis*, which follows these basic steps in eukaryotes (e.g. human cells):

1. The gene's contents are read by the protein complex RNA polymerase II (RNAP II) and the information is copied to pre-messenger RNA (pre-mRNA) in a process called *transcription*.
2. The pre-mRNA is processed into messenger RNA (mRNA), which is transferred from the cell nucleus to the cytoplasm.
3. The mRNA is processed by a complex of proteins and RNA called a ribosome, which builds a protein in a process called *translation*.
4. The protein may undergo further post-translational processing to produce the mature, active protein.

There may be processes before, after, and between these steps as well, depending on the organism,

the particular gene, and the current state of the cell.

In this dissertation, *gene expression* refers to the concentration of a given gene's mRNA in a given cell, i.e. highly expressed genes have high mRNA concentrations. It would be preferable to use the more traditional definition of expression, meaning the concentration of the final gene product (either a protein or functional RNA), since there are regulatory steps that occur after mRNA synthesis. However, it is very difficult to simultaneously measure the concentrations of all final gene products in a single cell or cell population. Paired with incomplete knowledge about the existence, functions, chemical states, and interactions between the numerous genes, RNAs, and proteins in any given organism, compiling a comprehensive snapshot of the internal state of a cell requires more advanced technologies than are available today. Consequently, many of the findings contained in this dissertation are based on the simplifying assumption that mRNA concentration is a good proxy for final gene product concentration. The most plentiful and comprehensive sources of gene expression data available today come from *microarray* and *mRNA sequencing* (RNA-seq) measurements. See Appendix D for a brief overview of microarray and RNA-seq technologies.

3.2 Gene regulation

The demand for cellular processes varies over time. Cells such as red blood cells die naturally in the human body and are regularly replenished by stem cells, but unrestricted cell division is one of the primary hallmarks of cancer. Humans of course require an active immune system to guard against pathogens, but a hyperactive immune system can lead to a plethora of diseases in which the immune system attacks the host. The nonlinear, nonequilibrium nature of biological systems requires each cell to carefully regulate its rate of protein production in response to its immediate needs and its environment.

Regulation of the amount of final gene product can happen at any point in the biosynthetic process. Histones and DNA can be chemically altered to change whether a gene is exposed for transcription; mRNA and proteins can be *degraded*, or disassembled into their basic components, by other proteins; but one of the most important and well studied forms of regulation comes from

the action of *transcription factors* (TFs). Transcription factors are proteins that bind to specific segments of DNA, increasing (*activators*) or decreasing (*repressors*) the likelihood that RNAP II successfully binds to and transcribes a given gene. In many cases, transcription of a given gene cannot begin until its specific transcription factors bind to their respective sites on the DNA.

Of course, gene regulation is useless unless the regulators themselves are regulated. Individual transcription factors do not make autonomous decisions about whether to up- or down-regulate their target genes; rather they respond to their own regulatory inputs from upstream genes, which in turn depend on their upstream regulators, *ad infinitum*. This complex web of regulatory interactions and feedback loops, called a *gene regulatory network* (GRN), provides each cell with a sort of "brain" that allows it to make decentralized decisions [20] about how to modulate the production of the cell's various proteins at any given time.[1]

As shown in Section 2.8, GRNs are typically modular, where genes share many interactions within communities and fewer interactions between communities. These communities host *pathways*, or subnetworks of a cell's GRN responsible for constructing particular cellular components or executing particular cellular functions. This modular structure allows pathways within the same community to synchronize and cooperate, while operating more or less independently of pathways in other communities. Tools such as DAVID [38] have been created to detect which pathways are *enriched*, or present to a statistically significantly degree, in a given set of genes drawn from a given population of genes. The p-value of the enrichment of a given pathway in a given set of genes is usually estimated using the cumulative distribution function of a hypergeometric distribution,

$$p(k_0) = \sum_{k=k_0}^{K} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{K}} \tag{3.1}$$

where N is the number of genes in the population, n is the number of genes in the input set, K is the number of genes in the pathway, and k_0 is the number of genes that are present in both the input set and the pathway. Applications of gene enrichment in the context of evolution will be discussed in Chapter 4.

[1] This emergent decision-making ability motivated us to simulate the dynamics of gene regulatory networks using a model developed in neuroscience, the Hopfield model, and will be described Chapter 5.

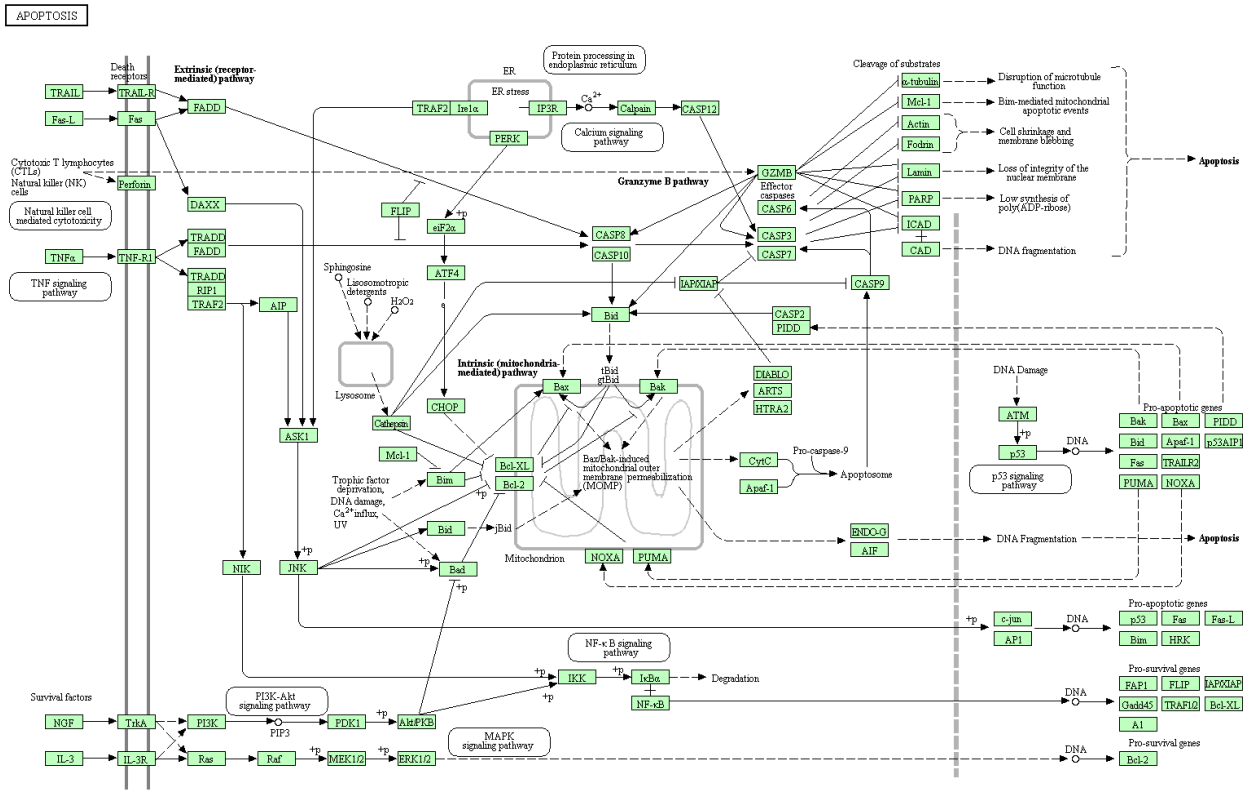


Figure 3.1 Schematic of some known apoptosis components and signaling pathways in humans. Figure adapted from the interactive version available from KEGG [72].

3.3 Cancer

GRNs make many decisions in a collective rather than centralized manner, which stabilizes against both external and internal perturbations [98] (however, some functions such as developmental processes may be governed by the action of only a handful of genes). These decisions are executed by proteins, which must fold into very precise shapes largely determined by the their encoding genes' sequences to function correctly, although the environment and post-transcriptional modifications play a role in folding, localization, and interactions as well. *Mutations*, or permanent changes to a gene's nucleotide sequence, are sometimes harmless, but certain mutations or collections of mutations can cause encoded proteins to misfold or exhibit altered function. Furthermore, *cross linking* and *double-stranded breaks* in DNA can also cause misregulation of important processes. Cells have machinery to attempt to repair damaged DNA, but if the damage is too extensive, the cell

may initiate *apoptosis*, or the process of "programmed cell death" in which the cell destroys itself for the sake of the multicellular organism. Apoptosis is one of the most important signaling pathways being investigated today. Fig. 3.1 shows a high level summary of some of the experimentally verified stimuli, genes, and reactions involved in the apoptosis signaling pathway in humans.

Elements of the apoptosis signaling pathway are also susceptible to mutation. One of the most commonly mutated genes in cancer, *TP53* (which encodes the protein *p53*), is integral to DNA maintenance and apoptosis initiation. Mutations in TP53 result in not only the loss of p53's original functionality, but can give rise to new, undesired interactions [107]. TP53 is also a kind of *tumor suppressor gene* because p53 is responsible for pausing cell cycle to allow the cell to check that the DNA can be replicated without errors. Without properly functioning p53s, cells undergo division regardless of the fidelity of the daughter cells' DNA [63], allowing for the introduction of yet more errors in the next generation's DNA. Another related category of genes, *oncogenes*, is responsible for up-regulating processes like cell cycle. The acquisition of mutations that disrupt the function of apoptosis, cause tumor suppressor genes to malfunction, and cause oncogenes to become hyperactive are necessary steps in *carcinogensis*, or the evolution of normal cells to cancer cells.[2] Recently, more of these hallmarks of cancer have been identified [61], and are shown in Fig. 3.2. For a review of different cancer types and their corresponding driver mutations, see [153].

Any viable theraputic intervention must kill or control cancer cells while doing little or no damage to normal cells, and because cancer cells are physically very similar to their normal counterparts, treating cancer clinically is quite difficult. Although broad classes of cancers have many distinguishing mutations and characteristics, each patient has their own sets of mutations that make their cancer unique. With the advent of single-cell RNA-seq technology, it is rapidly becoming apparent that the mutation and gene expression profiles of cells drawn from the same tumors are far from homogeneous [22, 120], a direct consequence of the misregulation of cellular processes.

[2] In fact, many of the mutations that commonly drive carcinogensis tend to be located in the parts of the genome responsible for multicellular cooperation (for example, regulation of proliferation). These kinds of mutations could be considered a form of "reverse evolution" [25] in which the mutated cell behaves like a unicellular organism, acting with no regard for the health of the host organism.

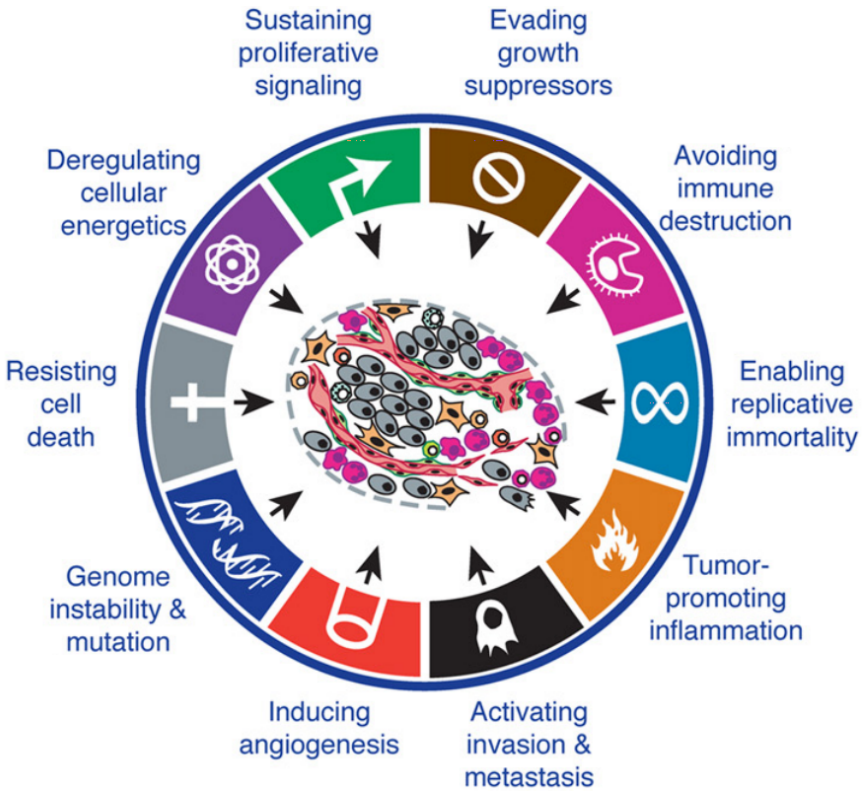


Figure 3.2 Ten of the "hallmarks of cancer." Acquiring each of these features drives cells into more cancer-like phenotypes. Figure adapted from [61].

As the *founder clone* (marked with N's in Fig. 3.3) becomes cancerous, it mutates and divides at a higher rate than normal cells. Each successive generation follows its own unique evolutionary path, driven by random mutations and natural selection. These mutations accumulate and, just as with the phylogenetic tree in Fig. 2.18, their genomes begin to diverge. Cells from the same or different lines may compete (because space and resources are limited) and/or cooperate (by releasing growth factors to further up-regulate growth and proliferation in other cells), increasing each successive generation's fitness compared with normal cells. Designing a drug regimen that effectively kills one of the clonal lines may simply allow more aggressive, drug-resistant lines to fill the vacuum.

The apoptosis diagram in Fig. 3.1 represents how at least part of a typical functioning signaling pathway works, but there are a large number of ways for that system to malfunction (removing or adding edges, changing binding strengths, etc.). Before any progress can be made in treating a

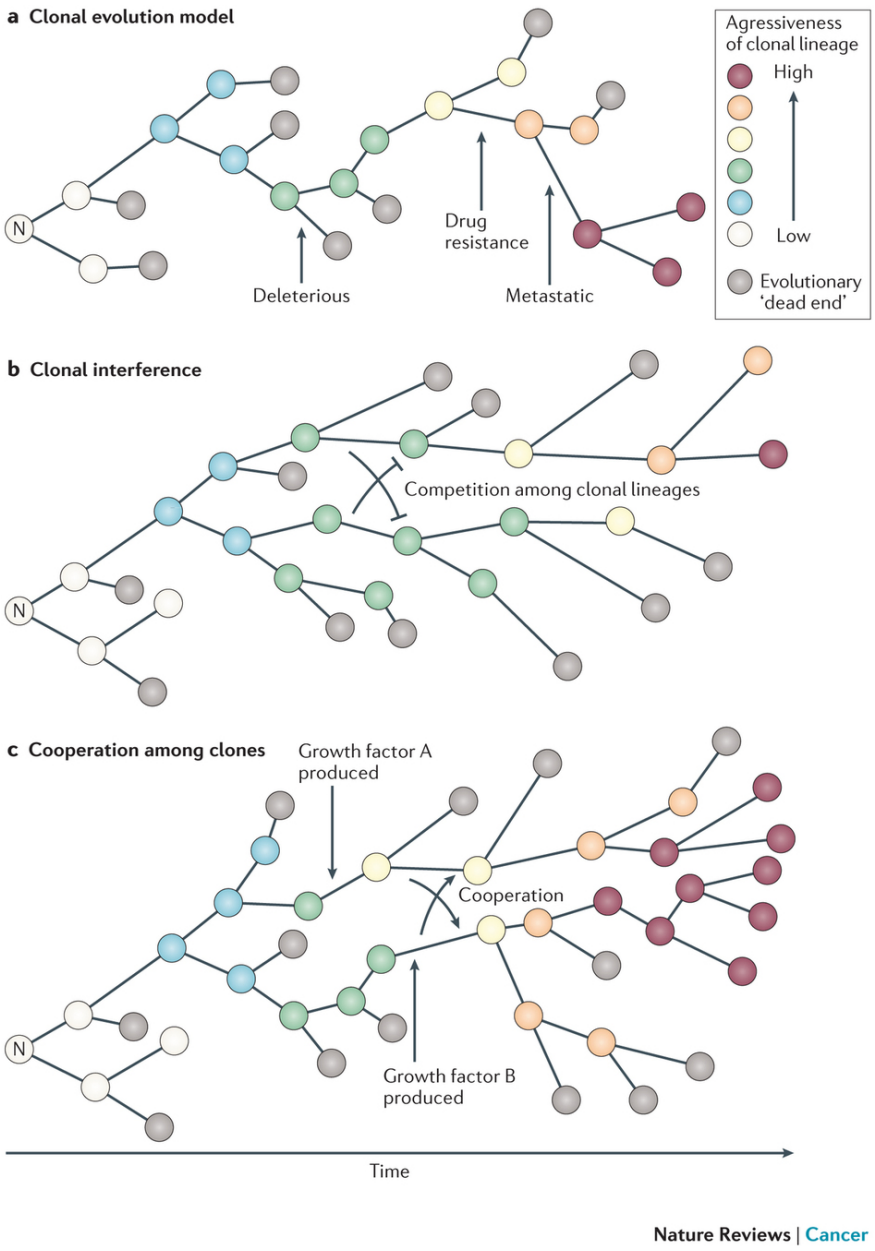


Figure 3.3 Schematic of clonal evolution within a single patient. As the founder clone (marked with N's) becomes cancerous, it mutates and divides at a higher rate than normal cells. Each successive generation follows its own unique evolutionary path, driven by random mutations and natural selection. These mutations accumulate and, just as with the phylogenetic tree in Fig. 2.18, their genomes begin to diverge. Cells from the same or different lines may compete (because space and resources are limited) and/or cooperate (by releasing growth factors to further up-regulate growth and proliferation in other cells), increasing each successive generation's fitness compared with normal cells. Figure taken from [79].

malfunctioning GRN, the structure of the GRN itself must be deduced.

3.4 GRN reconstruction

The human genome contains an estimated 23,000 genes [122] that together encode for an estimated 29,000 proteins [106] (some genes encode multiple proteins thanks to *alternative splicing*). Although nowhere near the complexity of the network of neurons in the human brain, these numbers pose a problem for understanding gene regulation: explicitly determining whether one gene regulates another or one protein interacts with another across all possible pairs is an enormous task. This is further complicated by the fact that many genes require the presence of multiple transcription factors before transcription initiates, effectively giving rise to 3+ body interactions.[3] Nevertheless, much effort has been devoted to assembling manually-curated databases of regulatory interactions. Some databases like KEGG [72] focus on organizing information about individual pathways. Others databases like TRANSFAC [101] gather data about transcription factor/gene interactions across many organisms and cell types. Curated databases such as these rely on manual or computer-assisted searches through abstracts and publications to identify and categorize known interactions.

However, assembling a list of known interactions alone cannot predict the existence of unknown interactions. The true underlying GRN in a given cell type likely has more edges than have been directly measured. Additionally, it may not be necessary to understand the precise details of every regulatory interaction to understand some of the statistical properties of GRNs. A practical alternative to fully combinatorial experimental detection is to employ *network reconstruction algorithms*, which infer the structure of the GRN based on some sort of correlation between the expression of pairs of genes across multiple gene expression profiles. A number of reconstruction algorithms have been developed recently, driven in large part by the vast quantity of gene

[3] *Hypergraphs* are graphs (networks) that are composed of nodes and hyperedges. A *hyperedge* is the n-body generalization of the pairwise relationship that an edge in a normal network represents. Hypergraphs are interesting mathematical constructs and are of potential utility in the study of gene regulation, but will not be covered here.

expression and other data produced in recent years. ARACNE [16] is a well known network reconstruction algorithm based on mutual information; TIGRESS [65] uses least angle regression to detect edges; and GENIE3 [71] ranks the significance of edges using a tree-based ensemble method.

In 2015 I coauthored *A scalable method for molecular network reconstruction identifies properties of targets and mutations in acute myeloid leukemia* [116], which reported on an efficient network reconstruction algorithm developed by our collaboration. Curiously, we neglected to name the algorithm in the original publication, so for the sake of this dissertation the algorithm will be called *Ong's method*, named for the primary author. At the most basic level, Ong's method builds a GRN from the Pearson correlation between the expression of pairs of genes across sets of microarray and/or RNA-seq data. The reconstructed network discussed in the original publication used five carefully selected, high quality gene expression experiments (two RNA-seq and three microarray) of *acute myeloid leukemia* (AML), a form of blood cancer from the myeloid line of cells.[4] Each sample in each data set represented a bulk gene expression profile of untreated AML cells from a unique patient, so that the full collection of data covered many independent, hetergeneous instances of AML. This disease-specific network was named *AML 2.1*. Because my chief contribution to the paper was an analysis of the resulting networks and designing the set efficiency measure covered in Section 2.6, some of the more esoteric details have been omitted. Full details can be found in the original publication.

The Data Processing step runs each of the five gene expression data sets z through common software packages[5] (which convert the raw data to a usable form and conduct statistical tests to control for experimental errors) to produce a "normalized" gene expression matrix X^z, where $0 \leq X^z_{i,\mu} < \infty$ is the expression of gene i in sample μ from data set z. Although the gene expression from two distinct samples μ and μ' taken from two distinct data sets z and z' cannot be directly compared (since the reported absolute expression is sensitive to experimental conditions and platforms), the

[4] This video provides an excellent summary of acute myeloid leukemia (AML) as well as a related but distinct form of cancer, acute lymphoblastic leukemia (ALL).

[5] Expression data was processed using the R package Bioconductor for microarrays and the "Tuxedo" suite for RNA-seq.

normalization procedure enables comparisons between $X^z_{i,\mu}$ and $X^z_{i,\mu'}$ for $\mu \neq \mu'$. This means the correlations can be sensibly computed for any two genes i and j across all samples μ within a single data set z.

Some high but coincidental correlations may be detected in which one or both of the genes exhibit only minor random fluctuations around a high mean value, such as *housekeeping genes* (genes which are consistently active across many or all cell types). The second step in Ong's method, Optimization and Method Selection, was designed as both a filter for such genes and a way to compare the four algorithms (ARACNE, TIGRESS, GENIE3, and Ong's method). A minimum threshold for the coefficient of variation (CV, the standard deviation divided by the mean) was introduced so that genes are included in data set z only if they show high variation across samples relative to their means. To determine the best value of the CV threshold, 90 TFs which showed high expression in the AML samples were randomly selected from the TRANSFAC database. These 90 TFs had 2486 known regulatory interactions (hereafter *true interactions* or TIs) spread across 1273 unique target genes. These 90+1273=1363 genes were selected from the full data set, and each of the four algorithms attempted to reconstruct this subnetwork using the five data sets. The CV threshold was varied until the number of TIs identified in the top 100 most significant interactions was maximized for each of the data sets and each algorithm, producing a total of $(5 \text{ data sets}) \times (4 \text{ algorithms}) = 20$ optimal CV cutoffs.

All four algorithms aim to reconstruct the true underlying GRN, but because each was designed using different assumptions, they produce somewhat different networks for the same set of input data. In order to judge which method produces the best results, the list of inferred interactions was compared with the TIs. Of course, TIs only represent a fraction of all existing interactions, and the coverage of various regulatory interactions in scientific literature is uneven: famous genes such as TP53 have been the focus of a multitude of experiments because of their importance in diseases or development, while other genes like C9orf37 and C16orf59 are currently *predicted genes* (predicted from sequence motifs) with no known functions. These may not be genes at all, or may play subtle but crucial roles in studied or novel biological functions. Even considering

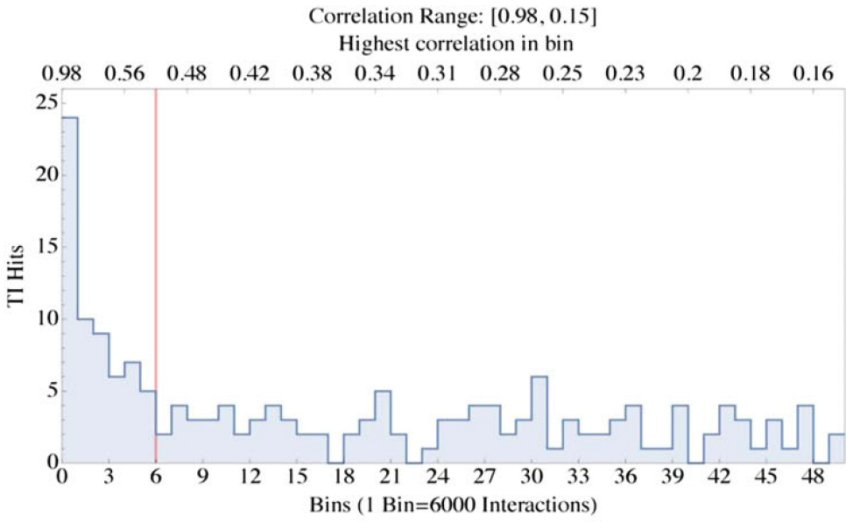


Figure 3.4 Binned Pearson correlation of inferred TF/gene interactions in one of five data sets using Ong's method, with maximum correlation (0.98) on the left and minimum correlation (0.15) on the right. The vertical red line identifies the computed Pearson correlation threshold, meaning all edges with Pearson correlation above the threshold are included in the final network. Figure taken from [116].

this potential bias, however, the ability to recreate a large number of known interactions is likely indicative of a meaningfully reconstructed network. The top 100 most strongly correlated pairs of genes identified by Ong's method contained as many or more TIs than the top 100 using ARACNE, TIGRESS, and GENIE3, both before and after CV optimization. Coupled with its relative speed and algorithmic simplicity, Pearson correlation was selected for Ong's method.

The TFG (transcription factor/gene) Subnetwork Reconstruction step proceeds with the five full data sets, excluding genes whose expression profiles failed the above CV threshold test. The Pearson correlation coefficient was computed between all pairs of genes (i, j) where at least one of i and j were known to be TFs according to the databases AnimalTFDB [162] and KEGG. The resulting list of interactions was ranked from largest to smallest Pearson correlation. Under the assumption that genes linked by true interactions should show higher correlation than two randomly selected genes, a Pearson correlation threshold was computed to identify which edges were likely real and which should be discarded.

To determine this threshold, the sorted Pearson correlations were divided into 50 bins, and the

number of TIs identified in each bin was tabulated, as shown in Fig. 3.4. (Note that, as expected, the bin containing edges with the highest correlations were most enriched for TIs, labelled as "TI hits" in the figure.) For a single randomly chosen interaction from this list, the probability that it is a TI is $q = N_T/N$, where N_T is the number of TIs and N is the number of inferred interactions. Thus, given N_B randomly chosen interactions in a given bin, the probability that at least k_0 of these interactions are TIs (i.e. the p-value of the number of TIs) follows a binomial CDF,

$$p(k_0) = \sum_{k=k_0}^{N_B} \binom{N_B}{k} q^k (1-q)^{N_B-k} \tag{3.2}$$

which, given $N_B \gg 1$ and $q \ll 1$, can be approximated as a Poisson distribution,

$$p(k_0) = \sum_{k=k_0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.3}$$

and $\lambda \approx 2$. The Pearson correlation threshold was chosen by moving through the bins in Fig. 3.4 from left to right and locating the first bin with $p(k_0) > 0.15$, or in other words, the first bin with $k_0 \leq 3$ (indicated with a red line). All edges in bins to the left of the cutoff were included in that data set's final list of edges.

Finally, the resulting TFG subnetwork was built by combining the edges inferred from each of the five data sets according to the above rules. An edge's *overlap category*, or the number of times it was detected across these five networks, is a measure of its reproducibility. To further avoid reporting spurious edges, only edges in overlap $\geq$2 (i.e. edges which were detected in at least two data sets) were included in the final network. Because Pearson correlation is a symmetric measure, the direction of inferred TFG edges were taken from TRANSFAC when available, and otherwise were left undirected. Although not directly addressed in the original publication, additional external information could aid in assigning directionality. Since only regulators can have nonzero outdegree, a list of which genes are regulators could define the direction of edges linking regulators and non-regulators.

The PPI (protein-protein interaction) Subnetwork Reconstruction step is a parallel and nearly identical step to the TFG step, except that it used the protein-protein interaction database HIPPIE [131] as its TI database. Because HIPPIE does not report any directionality, these inferred

interactions were left undirected. Combining the TFG and PPI subnetworks resulted in the completed (unweighted) network AML 2.1.

The network clustering algorithm *MCODE*[6] was used to identify sets of highly interacting genes, which were in turn provided to DAVID for enrichment analyses. Thirteen of these clusters are shown in Fig. 3.5, with some of the most significantly enriched pathways listed. The multicolored, dense core hosts pathways crucial for any cell from nearly any organism, such as DNA and RNA processes and cell cycle. In contrast, the more remote, homogeneously colored clusters host cell type-specific functions like leukocyte/lymphocyte activation, immune response, and heme biosynthetic process. The presence of these blood-specific pathways is to be expected given that the input data sets were from AML studies. The relationship between the topological structures of GRNs and the functions of their genes will be discussed more quantitatively in Chapter 4.

The process of building this AML-specific network reduced the number of nodes from over 20,000 genes in the raw microarray and RNA-seq data to 5,667 genes. Out of the 26 most common mutations found in AML cells [86], 21 were present in the 5,667 genes in AML 2.1. By Eq. 3.1, the p-value of this enrichment was 2.3×10^{-8}. To determine the functional context of these mutations, the set of the 21 mutations plus each mutation's immediate upstream and downstream neighbors (a total of 257 genes) was provided to DAVID to test for enrichment. Table 3.1 shows the resulting enrichment profile for all pathways with FDR-corrected (Benjamini) p-values below 0.01. As expected, nearly all identified pathways are related to one of the hallmarks of cancer, particularly processes involving DNA metabolism and maintenance (DNA metabolic process; chromosome organization; response to DNA damage stimulus; DNA repair; chromatin modification; DNA replication; histone modification; and negative regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process), proliferation (cell cycle), and resisting apoptosis (regulation of cell death).

[6]MCODE [10] is an older, parametric network clustering algorithm designed specifically for biological/molecular networks like GRNs, and can be easily implemented using network analysis software like Cytoscape [133]. However, modularity-based algorithms (which are typically non-parametric) are the most common network clustering algorithms used today. See this video for an animated description of the MCODE algorithm.

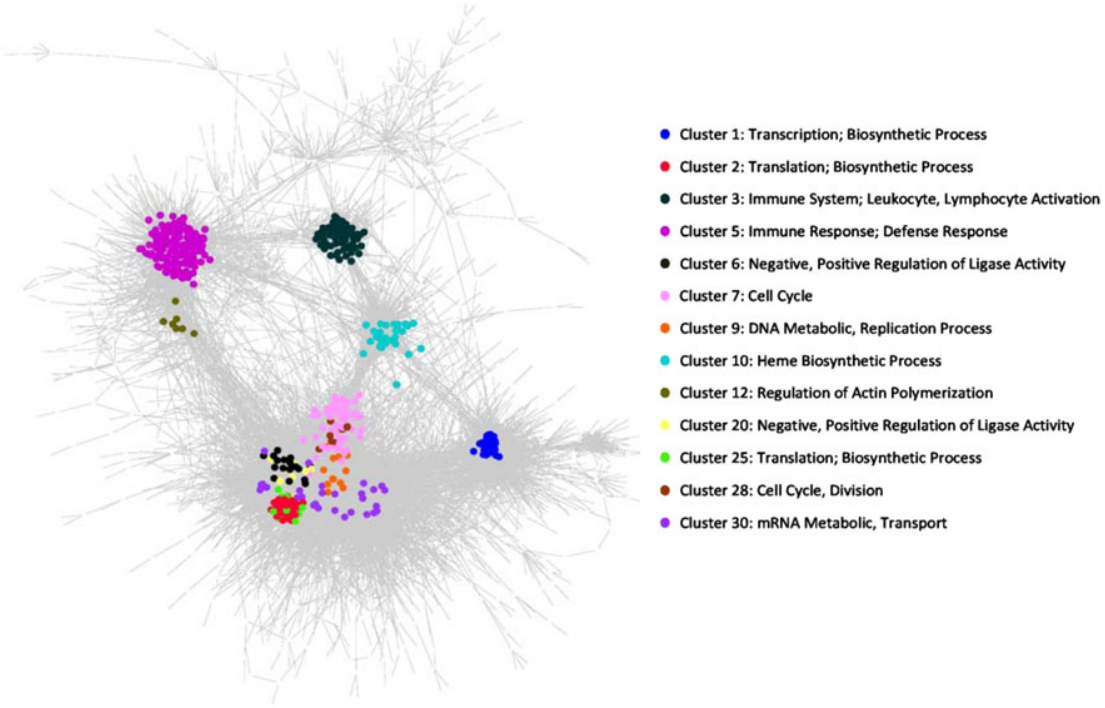


Figure 3.5 Force-directed layout of the largest connected component of AML 2.1. Because of the large number of nodes in the network, only the edges (gray lines) are rendered. Some of the most highly enriched GO pathways within clusters identified by MCODE are colored. Figure adapted from [116].

Pathway name	*Number of genes detected*	*p-value*	*Benjamini value*
DNA metabolic process	43	4.28E-16	7.55E-13
Chromosome organization	42	5.06E-16	9.44E-13
RNA processing	42	3.00E-14	5.08E-11
Response to DNA damage stimulus	33	7.17E-13	1.22E-9
Cell cycle	47	3.29E-12	5.58E-9
Cellular response to stress	39	8.55E-12	1.45E-8
DNA repair	27	2.79E-11	4.73E-8
Chromatin modification	23	1.16E-8	1.97E-5
DNA replication	18	1.20E-7	2.03E-4
Histone modification	14	5.17E-7	8.77E-4
Regulation of cell death	37	1.69E-6	2.90E-3
Cellular macromolecular complex subunit organization	22	4.39E-6	7.40E-3
Negative regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	27	4.67E-6	7.9E-3

Table 3.1 Pathway enrichment of 21 common AML mutations plus their first neighbors in AML 2.1. Nearly all pathways are related to the hallmarks of cancer from Fig. 3.2.

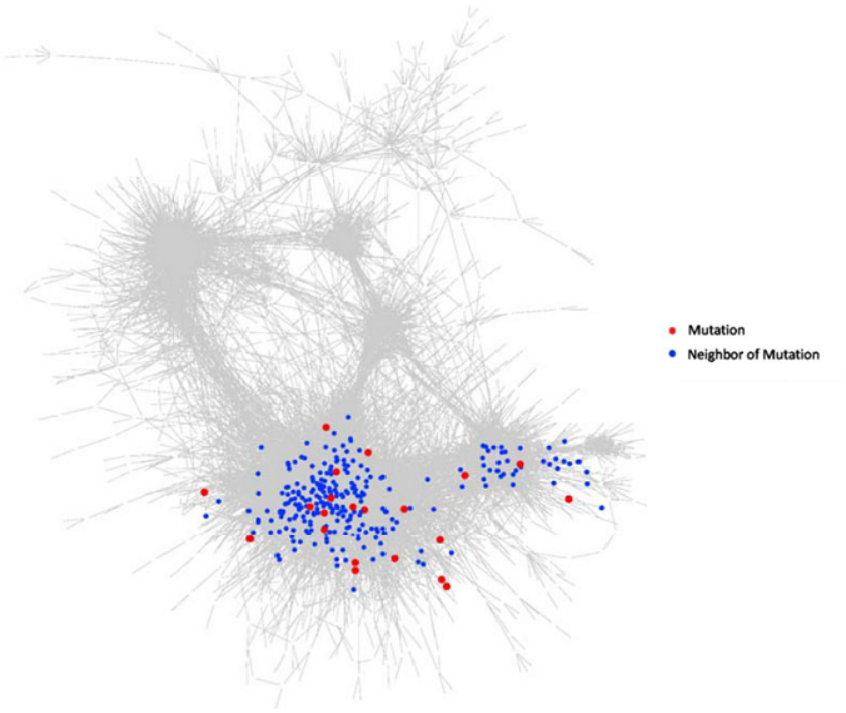


Figure 3.6 Force-directed layout of the largest connected component of AML 2.1. Mutations and neighbors of mutations are identified. Figure adapted from [116].

The close spatial proximity of the mutated nodes highlighted in red in Fig. 3.6 hints that the mutations are not randomly distributed across the network, but are concentrated in certain regions. As discussed in Section 2.3, however, a measure needed to be designed to assess the significance of the proximity of mutations in AML 2.1. This problem inspired the creation of the set efficiency from Section 2.6. To repeat, the set efficiency of a set of genes I is given by

$$E_I = \frac{1}{|I|(|I|-1)} \sum_{\substack{i,j \in I \\ i \neq j}} \frac{1}{d_{ij}} \tag{3.4}$$

where d_{ij} is the distance from node j to node i. Defining I to be the set of 21 genes commonly mutated in AML, the set efficiency was computed for I as well as for a control composed of 10 million randomly generated sets of 21 genes. E_I was found to be 0.2979, 1.8% larger than the largest set efficiency of the random sets observed and 145% larger than the mean. A skew normal probability distribution was fitted to a histogram of the randomized sets with $R^2 = 0.999982$, and an approximate right-tailed p-value of 7.3×10^{-8} was obtained for E_I (see Fig. 3.7). A more conservative

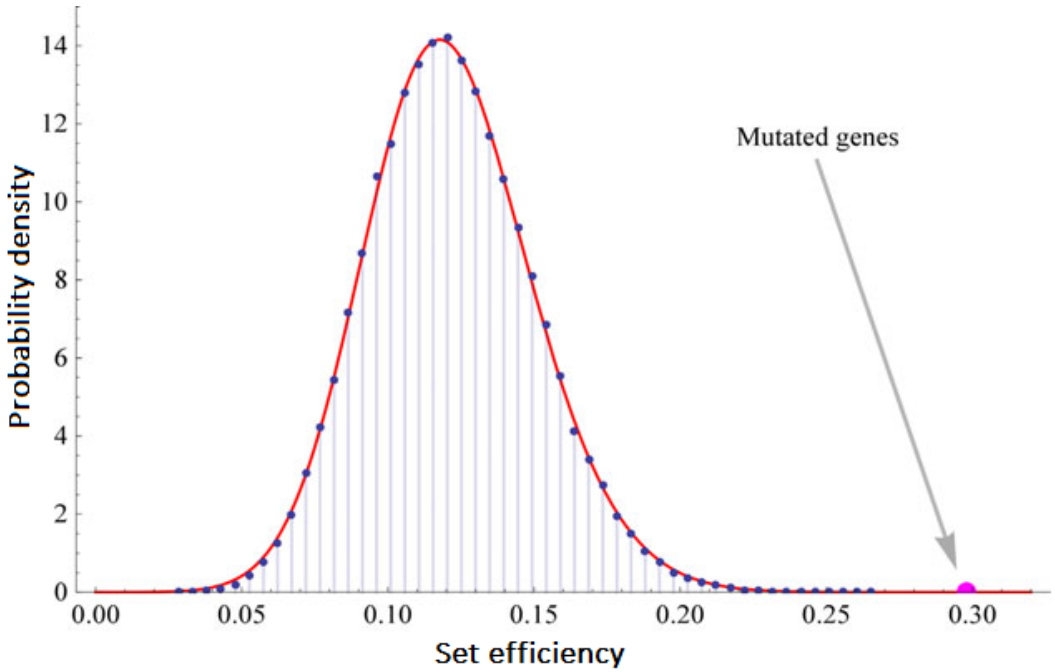


Figure 3.7 Set efficiency of 21 common AML mutations in AML 2.1, and the distribution of the set efficiency of 10 million randomly selected sets of 21 genes from the 5,667 genes in AML 2.1. The set efficiency of the mutated genes is far greater than expected at random. Figure adapted from [116].

estimate of the p-value was obtained by assigning a random direction to each protein-protein interaction whose true direction is unknown (rather than using an undirected edge). Adding edges to a network guarantees an increase in the global efficiency as well as the average set efficiency. The genes in I communicate with their neighbors predominantly through protein-protein interactions, and some of the PPI edges are listed as undirected in the AML 2.1 network because their true directions are unknown. To ensure that these undirected edges were not the sole cause of the statistical significance of E_I, a new network was constructed in which each undirected PPI edge was assigned a random direction. The significance of E_I was then calculated from the new network using 10 million random sets of 21 nodes as a control. Only 63 of these sets had set efficiencies greater than E_I, which gives an estimated p-value of $63/10^7 = 6.3 \times 10^{-6}$. Both estimates of the p-value clearly demonstrate the significance of the proximity of the mutations in AML 2.1.

It is well known in data science that even simple models based on a large amount of data

can outperform complex models based on a small amount of data [59, 66]. Efficient network reconstruction algorithms that can leverage the large and ever increasing amount of freely available gene expression (and perhaps other) data across many organisms and cell types could provide researchers with insights into the mechanisms governing properly functioning cells and suggest therapies for malfunctioning cells. Given its ability to identify expected characteristics such as the compartmentalization of pathways in communities and the functional location and close proximity of mutations, Ong's method appears to successfully reconstruct a reasonable AML GRN.

Our collaboration recently developed a weighted AML network, *AML 2.3*, built using a tweaked version of Ong's method with a total of twelve AML data sets. Ten of these data sets were used to build a network in precisely the same way as AML 2.1. The weight assigned to each of the x_n edges in overlap category n is given by the fraction of those x_n edges detected by performing Ong's method on the two remaining data sets. For example, if $x_n = 1{,}000$ edges were detected in overlap 3, and 600 of those same edges were detected in either of the two remaining data sets, then all edges in overlap 3 would be assigned a weight of $600/1000 = 0.6$. By construction, $0 \leq W_{ij} \leq 1$. This weight can be interpreted as the probability that an edge from a given overlap category is actually representative of a real edge based on its expected reproducibility. Chapter 4 analyzes the evolutionary properties of AML 2.3, further validating the reliability of Ong's method.

CHAPTER 4

EVOLUTION OF GENE REGULATORY NETWORKS

Evolution is a tinkerer.

—François Jacob*, Evolution and tinkering*

The following chapter is adapted from my 2016 publication, *Evolutionary and topological properties of genes and community structures in human gene regulatory networks* [142], and is organized as follows.

- Section 4.1, *Background*: an overview of previous work concerning the evolution of individual genes and GRNs
- Section 4.2, *Single-gene evolutionary measures*: a discussion of the connection between the two evolutionary measures of interest in this chapter, gene age and gene evolutionary rate
- Section 4.3, *Evolution and GRN topology*: an analysis of how the evolution of individual genes is linked to the topology of two GRNs, AML 2.3 and HumanNet
- Section 4.4, *Conclusion*: a summary of the chapter's findings

Supplementary information including extra plots and tables deemed too unwieldy for this dissertation can be found in the original publication.

4.1 Background

The evolution of a gene can be mapped in various ways. The absolute *evolutionary rate* (ER) of a gene, for example, can be computed from observed differences in *orthologs*[1] across species in the context of their phylogenetic relationships [82], whereas the *age* of a gene can be measured by

[1] Orthologs, also known as *orthologous genes*, are genes from separate species that can be traced back to a common ancestral gene, but whose sequences may have diverged due to accumulated mutations.

tracing when the gene first appeared in the organism's phylogenetic tree [25]. Quantities such as these allow researchers to chronicle the journey of individual genes across evolutionary history.

But genes do not exist, and therefore do not evolve, in isolation. Mutations in a transcription factor may affect the expression of the genes it regulates, since changes in a protein's amino acid sequence can cause it to lose compatibility with former binding partners, and gain compatibility with new partners. Accumulation of these alterations can lead to changes in fitness and, eventually, speciation. The evolution of individual genes is thus coupled with the evolution of the structure of the organism's GRN, and network properties should be related to the evolutionary properties of its constituent nodes and edges.

It has been proposed that GRNs grow and evolve incrementally via gene duplication followed by mutation and functional divergence [15, 115, 130, 146, 152], although changes may have occasionally arrived in bursts, as in whole-genome duplication [37]. This time-dependent network formation suggests that GRNs are composed of a core of ancient, conserved genes with fundamental functions, and younger, peripheral genes with species- or cell type-specific function, which mutate frequently until the functions of the newly created pathways are optimized. These mutations can alter GRNs by creating, removing, reassigning, or changing other properties of nodes and edges.

Fraser et al. demonstrated that interacting pairs of proteins have similar ERs [53]. This constraint is likely driven by the necessity of coevolution, since a change in one protein's sequence may require a corresponding change in its partner's sequence in order for the pair to remain compatible. Daub et al. showed that genes which are part of many biological pathways have lower ERs than genes which belong to few or no known pathways, further supporting the idea that related genes share similar evolutionary properties [34]. It has also been shown that ERs are weakly, but significantly, negatively correlated with degree, closeness centrality, and betweenness centrality, and that essential genes have high centrality and low ERs [58].

Originally in [142] and reproduced below, I establish quantitative relationships between the evolutionary history of genes and their topological properties in AML 2.3 [116] as well as for a

general human GRN, *HumanNet*[2] [88]. In contrast to the earlier studies mentioned above, this analysis moves beyond single-node centrality and pairwise measures by studying the connection between network topology and evolution, particularly from the point of view of network community structures. It will be shown that the ERs and ages of genes are not randomly distributed across the networks, but are naturally organized in communities with well-defined evolutionary characteristics: old genes cluster with old genes, and young cluster with young. Likewise, "cold genes" (genes with low ERs) cluster with cold genes, and "hot genes" (genes with high ERs) cluster with hot genes. This segregation also exists for enriched groups of genes identified by DAVID within the communities. In terms of network topology, genes and DAVID groups which are old and cold tend to be central, and those which are young and hot tend to be peripheral. This is demonstrated with traditional single-node centrality measures as well as with new network measures, the set efficiency and the interset efficiency (defined and discussed in Section 2.6). PageRank, a finite-range centrality measure, shows stronger biological significance than degree (a local measure) and betweenness centrality (a global measure), and the set efficiency and interset efficiency correlate strongly with the evolutionary histories of individual genes and DAVID groups.

4.2 Single-gene evolutionary measures

To compute the ER of a gene, the absolute ER for each amino acid position of the protein it encodes was computed using the method from Kumar et al. [82]. Given the multiple alignment at an amino acid position in 46 species [55], the ER of a given amino acid position equals the number of different residues divided by the total evolutionary time span, based on a known phylogenetic tree [82]. The ER of a gene is the average ER over all amino acid positions, in units of the number of substitutions per amino acid site per one billion years. The ER value ranges from 0.011 (most conserved) for LSM2 to 6.928 (least conserved) for CDRT15.

Ages, taken from Chen et al. [25], were estimated by comparing the human genome to the

[2]HumanNet is a GRN constructed from 21 different methods using diverse data types, including microarray co-expression, databases and mass spectrometry proteomics.

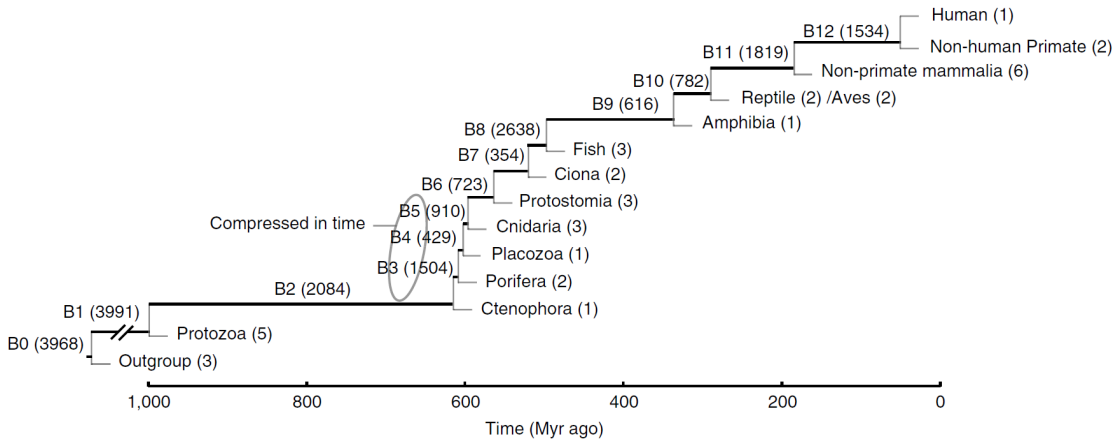


Figure 4.1 Phylogenetic tree used to compute gene ages. Image taken from [25].

genomes of species from 13 major clades which branched from the evolutionary path that resulted in humans, indexed 0 (oldest) through 12 (youngest). A gene's age (more technically, a lower bound on its age) was determined by searching for the earliest time at which an orthologous gene appears in an organism from one of these 13 clades (see Fig. 4.1).

While computed differently, a gene's ER and its age are related. It could be that young genes with novel functions need time to fine-tune their properties in order to optimize the fitness of the host organism, so young genes tend to be hot. Likewise, old genes with fundamental roles, such as protein translation, may have had enough time to sufficiently optimize their functions, and so should change very slowly. As expected, the ERs and ages of genes are strongly correlated ($R = 0.504$, $p < 10^{-300}$).

4.3 Evolution and GRN topology

Consistent with previous results [53], interacting genes tend to have similar ERs and ages. The distributions of differences in ERs and ages between genes linked by an edge in AML 2.3 are closer to zero than those of degree-preserving randomizations of the same network. The distribution of ER differences is shown in Fig. 4.2. To estimate the significance of this difference in widths, 20,000 degree-preserving randomization of AML 2.3 were conducted and the standard deviation

of each distribution was recorded. Comparing the true standard deviation to the randomized trials resulted in an approximate Z-score[3] of -96.8 for differences in ER and -72.0 for differences in age. This tendency for connected genes to have similar ERs and ages hints that there may be large-scale segregation between clusters of old, cold genes and young, hot genes. Indeed, this is reflected in the natural community structure present in AML 2.3, as well as in the DAVID groups present within these communities.

The main results of the community analysis are in Table 4.2 for AML 2.3 and Table 4.3 for HumanNet. These tables list the ER and age properties for the ten largest network communities, and for the three most significantly enriched DAVID groups found within each community. Communities and DAVID groups in Tables 4.2 and 4.3 labeled "cold" and "hot" have significantly lower and higher ERs than the network average, respectively. Likewise, groups of genes labeled "old" and "young" are significantly older and younger than the network's average age, respectively. A relatively strict one-tailed significance level of $p < 10^{-3}$ in the difference from the mean was chosen for both ER and age.

The ERs and ages for many of these DAVID groups reflect their biological functions. Zinc finger proteins, which are enriched in both AML 2.3 and HumanNet, are involved in a large number of heterogeneous cellular processes [28], so it may be that their genes need to adapt more often than genes with very specific singular functions. They also have a particularly high rate of duplication and loss, so while the family itself is old (found in animals, plants [28], and fungi [14]), individual genes in this family are young [95]. Genes involved in transcriptional regulation may also need to be flexible enough to tune the expression of target genes in response to environmental changes over time [145, 160]. The olfactory group is enriched in HumanNet, and it is significantly younger than average. A small number of olfactory genes were present in early chordates, but olfactory systems became far more complex and diverse in land-dwelling animals, particularly in mammals [113]. Conversely, the most fundamental DAVID groups have experienced few changes since early single-celled lifeforms. DAVID groups such as mRNA metabolic process [8] and trans-

[3] $Z = (x - \mu)/\sigma$ where x is the measurement, and μ and σ are the distribution's mean and standard deviation.

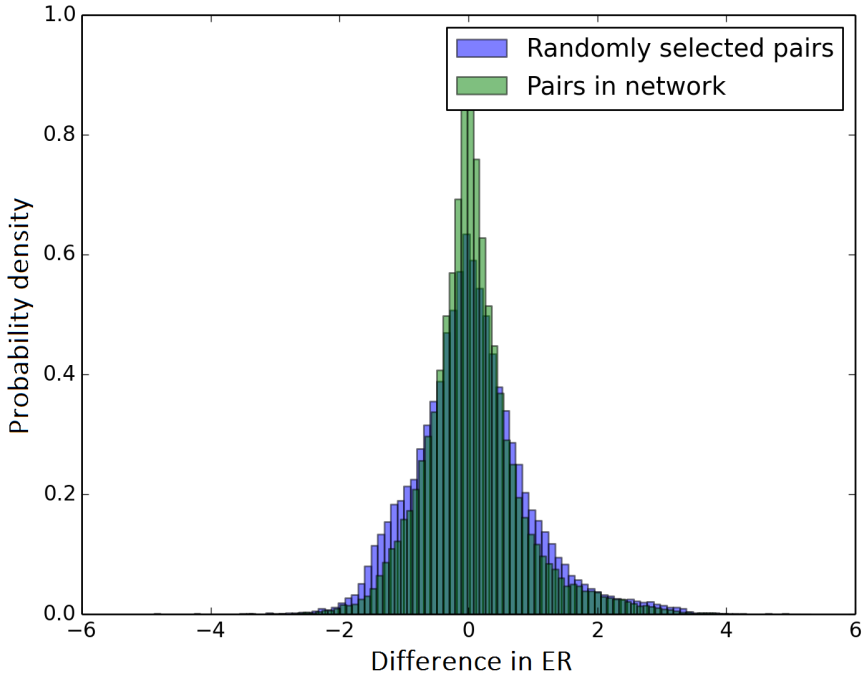


Figure 4.2 Difference in evolutionary rates (ER) between pairs of genes, $(ER_j - ER_i)$, for pairs connected by an edge $j \rightarrow i$ in AML 2.3 (green) and for pairs connected by an edge in one degree-preserving randomization of AML 2.3 (purple). Note that the distribution is asymmetric because AML 2.3 is a directed network. The width of the true distribution is significantly smaller than the mean of the standard deviation of 20,000 degree-preserving randomizations, resulting in an approximate Z-score of -96.8.

lational elongation [78] in AML 2.3 as well as ribosome [138] and protein kinase core [99] in HumanNet are old and stable, having nearly optimized their functions long ago.

As a control for the enriched DAVID groups, ten new communities were built by randomly shuffling the genes between communities from the network, while maintaining the size of each community. The resulting random communities were then analyzed using DAVID. This randomization procedure was followed for both AML 2.3 and HumanNet, and in both cases, the enrichment was far less significant than for the real communities. The enriched DAVID groups in Tables 4.2 and 4.3 are thus biologically meaningful, not merely coincidental.

Fig. 4.3 analyzes DAVID groups in AML 2.3 and their relationship with network communities and evolutionary properties. Fig. 4.3A shows that the ER distribution of translational elonga-

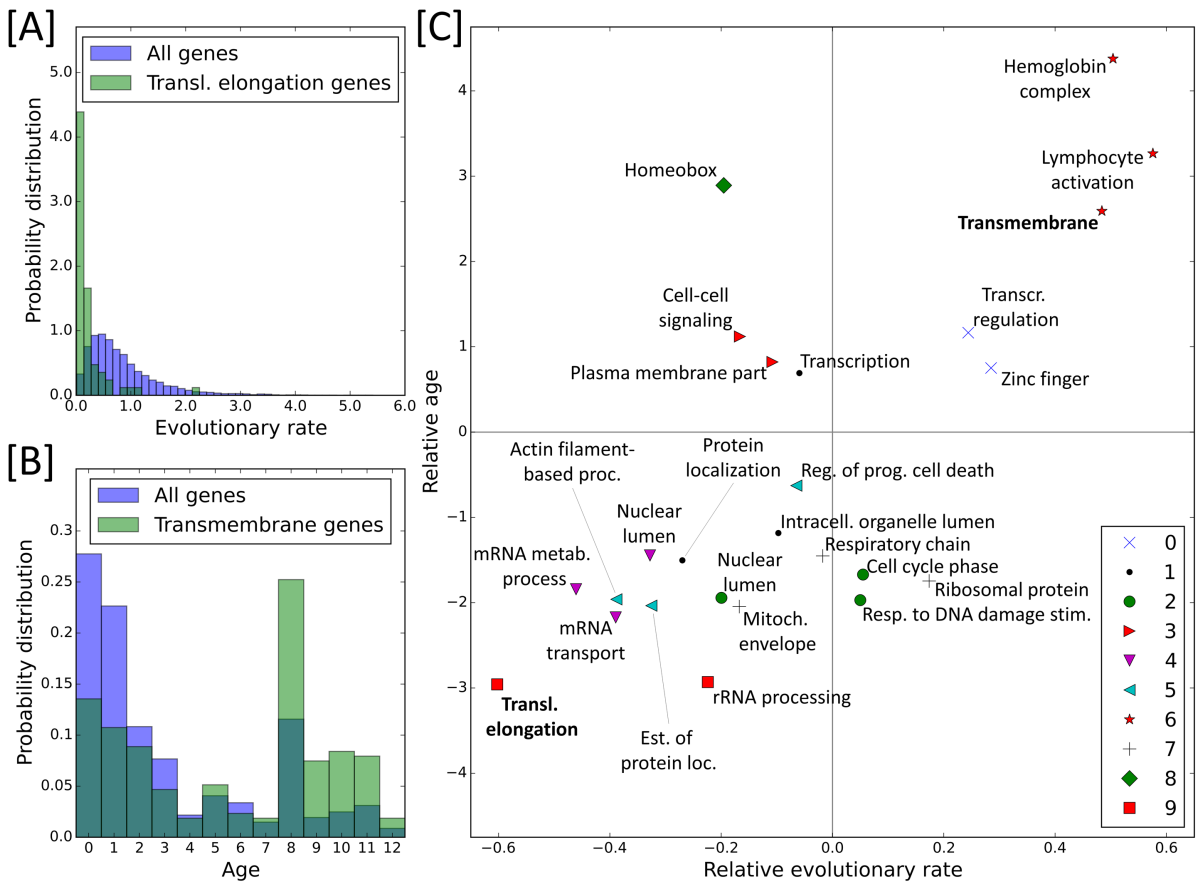


Figure 4.3 (A) Distribution of evolutionary rates (ERs), measured in units of the number of nonsynonymous substitutions per amino acid site per billion years, for all genes (purple) and for genes in the translational elongation DAVID group (green). This DAVID group has a very low ER compared to the background distribution. (B) Distribution of ages for all genes (purple) and genes in the transmembrane DAVID group (green), where age=0 is the oldest and age=12 is the youngest. Transmembrane genes are much younger than average. (C) Summary of mean ER and mean age for DAVID groups in Table 4.2. The relative ERs on the x-axis are computed from $ER_{\text{relative}} = ER_{\text{DAVID group mean}} - ER_{\text{network mean}}$, and likewise for relative age on the y-axis. The DAVID groups from (A) and (B) have bold labels in (C). Each marker type corresponds to one of communities 0 through 9. As expected, old DAVID groups tend to have a low average ER (i.e. are "cold"), and young DAVID groups tend to evolve frequently (i.e. are "hot"). Unabbreviated DAVID group names are listed in Table 4.2.

tion genes is noticeably left-shifted relative to the ERs of all genes, indicating that it hosts relatively slowly evolving genes. Transmembrane genes are much younger than average, as shown in Fig. 4.3B. Fig. 4.3C provides a comprehensive picture of the evolutionary properties of the ten largest network communities (symbols) with their main DAVID groups (as labeled).

	Degree centrality				*PageRank*				*Betweenness centrality*			
	R	***p-value***	ρ	***p-value***	***R***	***p-value***	ρ	***p-value***	***R***	***p-value***	ρ	***p-value***
Single-gene ER	-0.06	8.9E-10	-0.15	7.4E-51	-0.14	1.5E-43	-0.25	5.6E-141	-0.07	1.2E-11	-0.21	3.0E-103
Single-gene age	-0.06	7.3E-09	-0.18	2.5E-73	-0.12	4.1E-33	-0.22	1.7E-111	-0.04	5.5E-05	-0.14	2.0E-47
DAVID group ER	-0.13	5.3E-01	-0.04	8.4E-01	-0.58	2.3E-03	-0.57	2.8E-03	-0.25	2.3E-01	-0.18	3.9E-01
DAVID group age	-0.1	6.4E-01	-0.21	3.1E-01	-0.75	1.4E-05	-0.86	5.1E-08	-0.26	2.1E-01	-0.23	2.7E-01

Table 4.1 Correlations (Pearson's R and Spearman's ρ) between evolutionary measures and network measures in AML 2.3. The most significant correlation is between PageRank and ER for individual genes, and between PageRank and the mean age of DAVID groups (see Fig. 4.4).

Dividing genes into DAVID groups causes stronger relationships between network topology and evolutionary properties to emerge. Traditional single-node centrality measures such as degree, betweenness centrality, and PageRank show small but significant correlation with ERs and ages, with the oldest, coldest genes being the most central (see Table 4.1). Grouping genes by DAVID group leads to stronger correlations, the clearest of which is between the mean PageRank and mean age, shown in Fig 4.4 (Pearson's $R = -0.75$, $p = 1 \times 10^{-5}$; Spearman's $\rho = -0.86$, $p = 5 \times 10^{-8}$). These three centrality measures are related, but differ in their global reach. Degree is completely local, only dependent on the number of neighbors of a gene; betweenness centrality is global, requiring information from the entire network; but PageRank is between these extremes, influenced by all genes but with more weight granted to those genes which are nearby. The strong correlation between PageRank and evolutionary measures thus may be explained by the presence of communities in the GRN, since community structure itself is strongly correlated with ER and age, as shown in Tables 4.2 and 4.3.

Because of the strong correlation between a gene's history and that of its neighbors, genes are expected to evolve in groups rather than as individuals, which should be evident in the structure of the network. The set efficiency is shown in Fig. 4.5 for genes in AML 2.3 ranked from coldest to hottest, computed for the first 500 genes in the list, then the first 510 genes, etc. in steps of 10. As a control, the same "cumulative set efficiency" was computed for 100 sets of randomly ranked genes (but leaving the underlying network, AML 2.3, unchanged). Fig. 4.5 shows the mean of these 100 controls (solid green line) plus/minus one standard deviation (dashed green lines). Fig. 4.6 shows the same plot for genes ranked from oldest to youngest, with similar results. These plots indicates that the oldest, coldest genes tend to be close, separated by approximately four directed

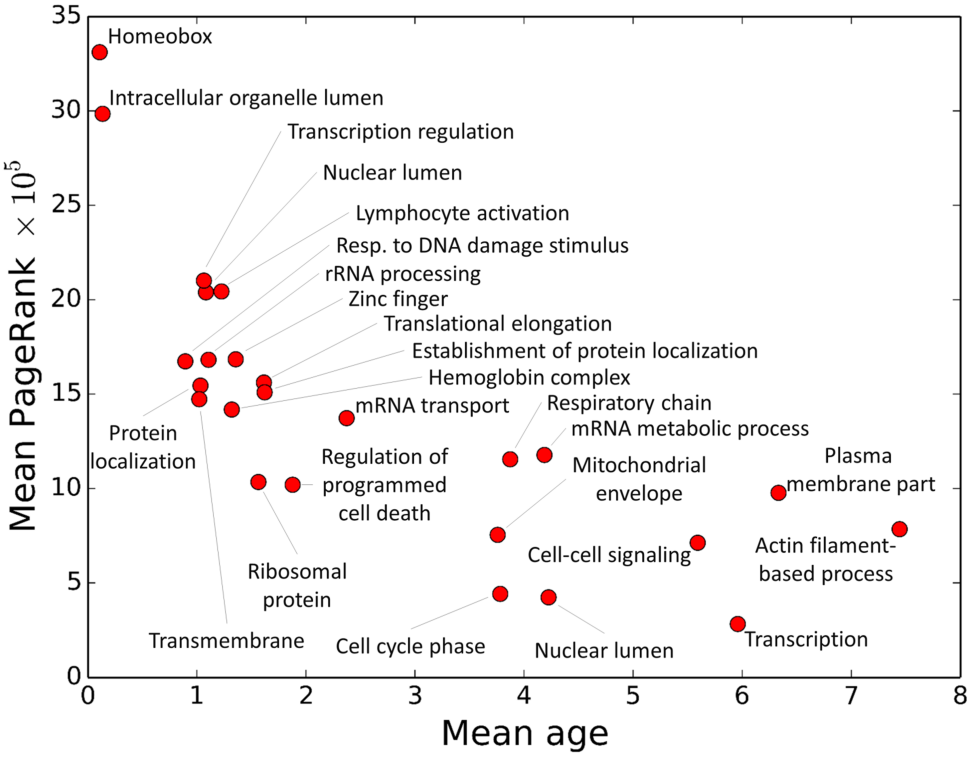


Figure 4.4 Mean PageRank versus mean age of each DAVID group from Table 4.2 (age=0 is the oldest and age=12 is the youngest). Old DAVID groups tend to have high PageRank. Unabbreviated DAVID group names are listed in Table 4.2.

edges, significantly smaller than the network average of approximately six. The set efficiency monotonically declines as hotter, younger genes are included. Note that the plateau below rank 2000 in Fig. 4.6 is a result of the discrete nature of the age data.

Furthermore, the oldest DAVID groups efficiently exchange information with each other, and the youngest DAVID groups are distant from the oldest DAVID groups as well as from each other. Fig. 4.7 shows the interset efficiency between all pairs of DAVID groups in AML 2.3, where the DAVID groups are sorted from oldest to youngest. Note that each diagonal term of the interset efficiency matrix is the set efficiency of that DAVID group. Similarly, Fig. 4.8 shows the interset efficiency between DAVID groups in HumanNet.

Purely locally, AML 2.3 and HumanNet look quite different from one another. AML 2.3 is composed of roughly 10,000 genes and 338,000 edges, and HumanNet is composed of 14,000

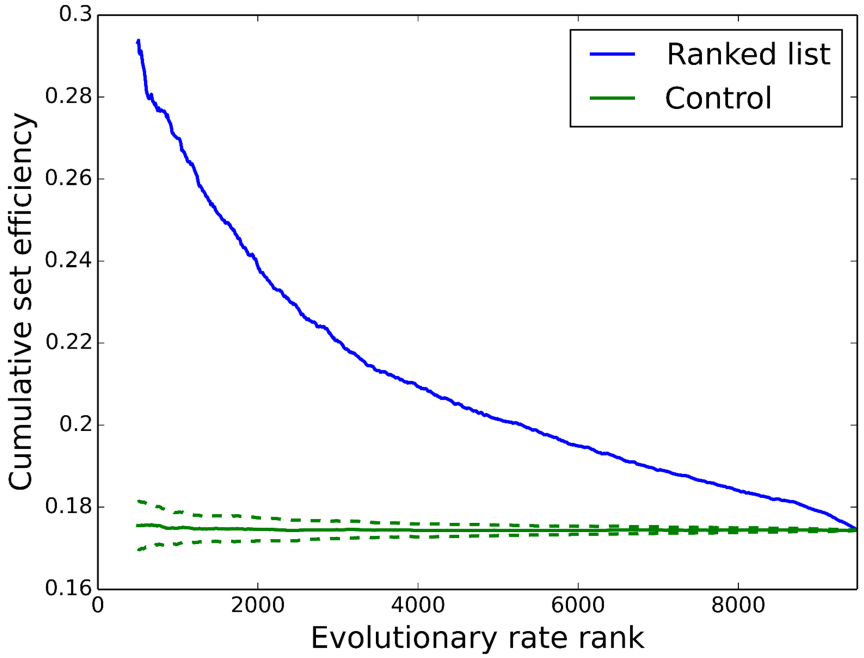


Figure 4.5 Set efficiency and evolutionary rate for AML 2.3. The cumulative set efficiency (SE) of all genes below a given evolutionary rate (ER) rank (lowest to highest ER, i.e. "coldest" to "hottest"). The SE of the 500 coldest genes is significantly higher than the control, and including hotter genes monotonically decreases the SE. This indicates that the coldest genes are located near each other, while the hottest genes are more dispersed.

genes and 876,000 edges. While they share roughly 9,000 genes, they share only 26,000 edges. However, modularity and interset efficiency, which are coarse-grained network measures that reveal the properties of sets of nodes rather than individual nodes or pairs of nodes, demonstrate that the same evolutionary signatures are present in both networks.

4.4 Conclusion

It was demonstrated that slowly evolving, old genes tend to interact with each other, and frequently evolving, young genes tend to interact with each other, whereas edges between those groups are less common. This naturally creates communities of genes with relatively homogeneous evolutionary attributes. Analyzing the networks in terms of communities and DAVID groups rather than single genes provided a new perspective which established clear relationships between network

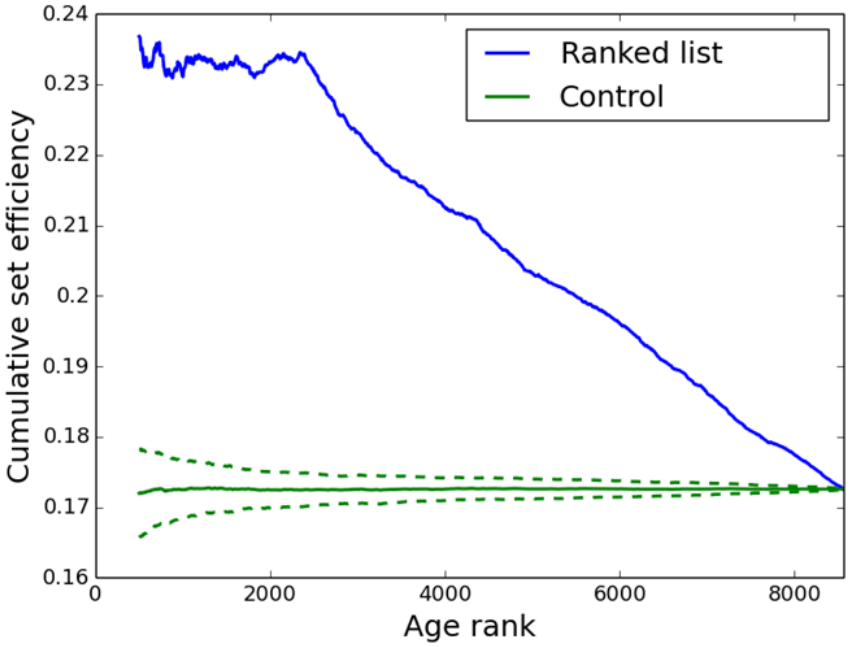


Figure 4.6 Set efficiency and age for AML 2.3. The cumulative set efficiency (SE) of all genes below a given age rank (oldest to youngest). The SE of the 500 oldest genes is significantly higher than the control, and including younger genes monotonically decreases the SE (after a transient period due to the discrete nature of the age data). This indicates that the oldest genes are located near each other, while the youngest genes are more dispersed.

topology and evolution. The abundance of connections between old DAVID groups and the relative scarcity between old-and-young and young-and-young DAVID groups suggests that during the course of human evolution, the primitive gene regulatory network present in early metazoans began as a core of fundamental genes and pathways. As genes duplicated and mutated, novel functions arose and eventually, through selective duplications, deletions, mutations, translocations, and rewirings, novel regulatory pathways emerged, growing outward from these ancient genes. This placed the oldest genes near the middle of the network and the youngest genes toward the periphery. These findings were mainly derived from an AML network and a general human network, and are consistent with previous reports [58].

No gene is an island. A real understanding of the evolution of a genome only comes from studying its constituent genes in the context of the underlying complex network of interactions

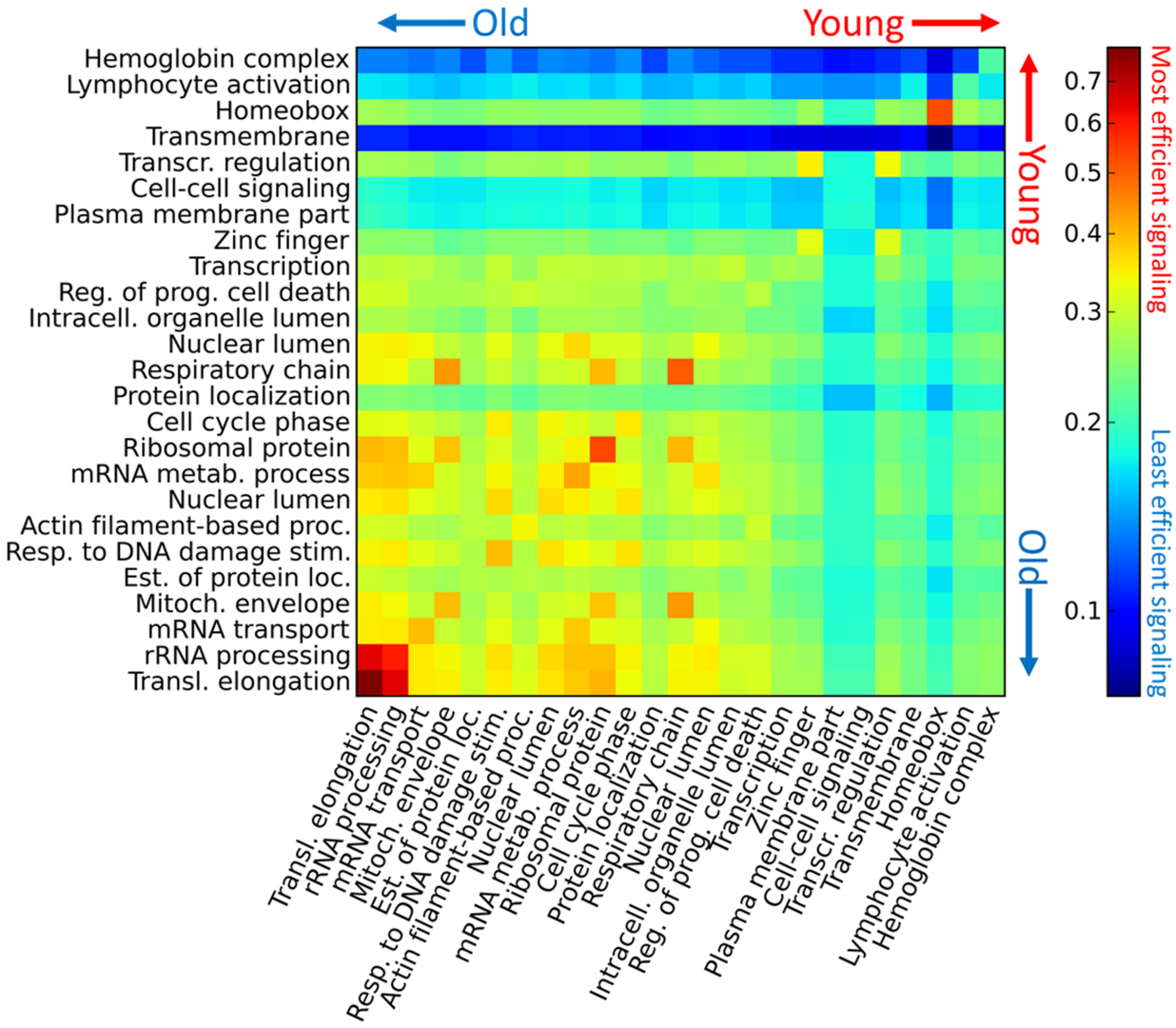


Figure 4.7 Interset efficiency and age for AML 2.3. Interset efficiency from DAVID group in column j to DAVID group in row i. The list of DAVID groups was sorted by average age from oldest (transcriptional elongation) to youngest (hemoglobin complex). Old DAVID groups exchange information efficiently, as indicated by the high interset efficiency values in the lower-left corner. Younger DAVID groups, particularly the blood cell-specific DAVID groups of lymphocyte activation and hemoglobin complex, are remote from most other DAVID groups. Note that the above matrix is asymmetric because the network is directed, and that the colors are log-scaled.

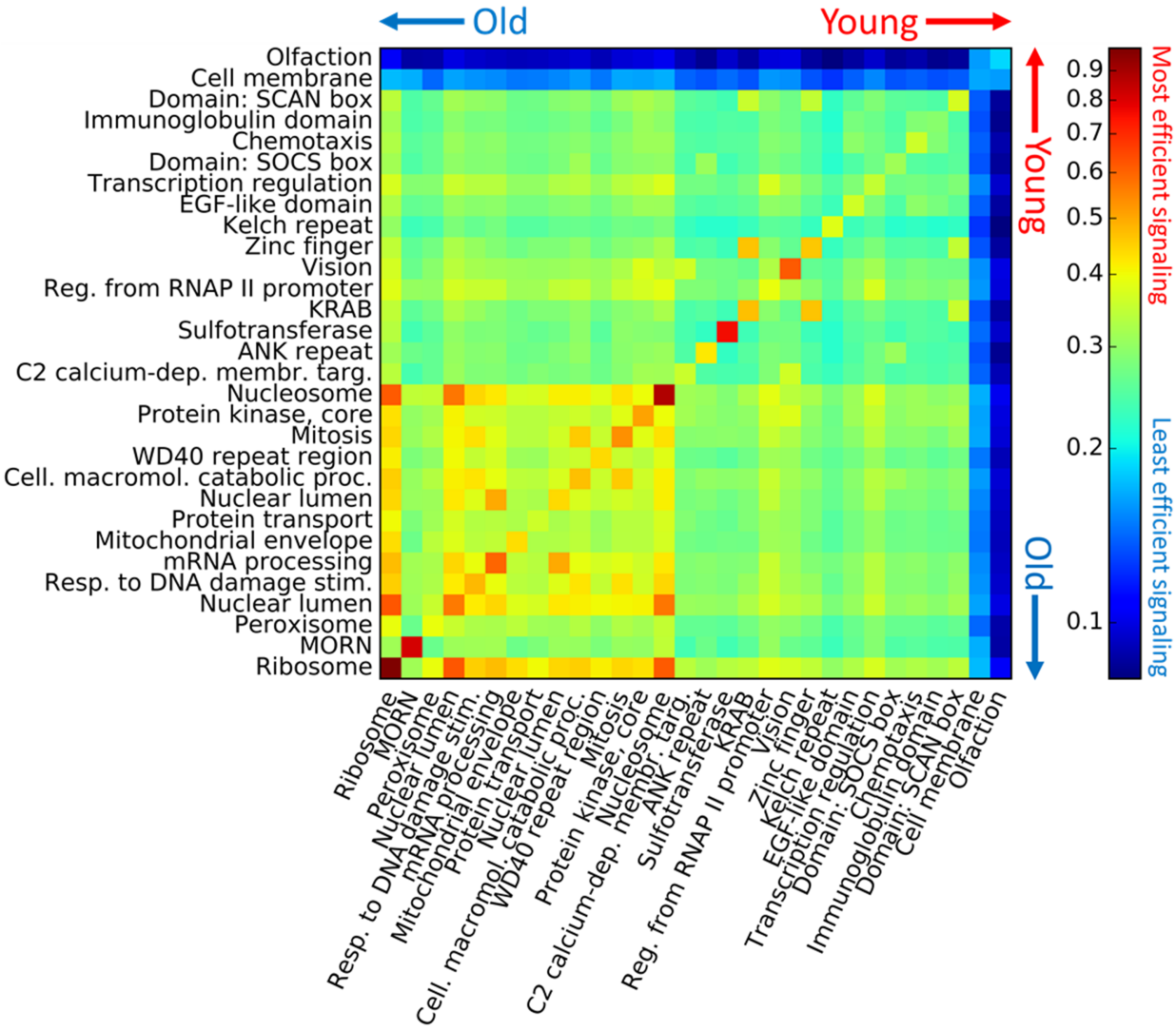


Figure 4.8 Interset efficiency and age for HumanNet. See Fig. 4.7 for explanation.

rather than as independent units. As network reconstruction methods continue to improve and more high quality networks become available, researchers will find more evidence of how evolution shaped, and continues to shape, the topology of gene regulatory networks.

Comm. index	*Num. genes*	*Comm. ER*	*Diff. in mean*	*p-value*	*Comm. age*	*Diff. in mean*	*p-value*	*DAVID group name*	*Group type*	*Num. genes*	*DAVID Benjamini*	*Group ER*	*Diff. in mean*	*p-value*	*Group age*	*Diff. in mean*	*p-value*
0	1760	hot	0.24	2.5E-81	young	1.01	6.7E-28	Zinc finger	P	275	1.2E-19	hot	0.32	6.0E-19	young	0.75	4.4E-05
								Transcription regulation	B	298	1.9E-16	hot	0.28	4.8E-14	young	1.17	9.3E-11
1	1579	average	-0.03	2.3E-02	old	-0.39	6.8E-07	Transcription	B	304	4.8E-24	average	-0.05	4.6E-02	young	0.69	5.6E-05
								Intracellular organelle lumen	L	239	9.7E-19	average	-0.07	7.1E-02	old	-1.18	1.7E-08
								Protein localization	B	145	2.9E-10	cold	-0.27	7.3E-05	old	-1.50	5.2E-08
2	1208	cold	-0.10	2.6E-10	old	-1.30	1.2E-58	Nuclear lumen	L	332	8.4E-107	cold	-0.21	2.6E-07	old	-1.94	1.6E-28
								Cell cycle phase	B	169	3.6E-98	average	0.06	1.0E-01	old	-1.67	5.4E-07
								Response to DNA damage stimulus	B	135	3.5E-58	average	0.03	3.1E-01	old	-1.97	3.2E-07
3	1055	average	0.01	2.4E-01	young	1.41	5.4E-53	Cell-cell signaling	B	146	3.6E-39	cold	-0.16	7.0E-04	young	1.12	1.3E-05
								Plasma membrane part	C	359	1.1E-67	cold	-0.11	2.5E-04	young	0.82	4.1E-06
4	867	cold	-0.27	3.7E-55	old	-0.94	1.7E-16	mRNA metabolic process	B	124	9.5E-66	cold	-0.46	2.7E-25	old	-1.84	2.7E-06
								Nuclear lumen	L	212	9.3E-60	cold	-0.34	5.3E-24	old	-1.45	1.2E-09
								mRNA transport	B	27	1.2E-11	average	-0.36	5.8E-03	average	-2.17	7.2E-03
5	780	cold	-0.11	1.1E-05	old	-0.91	5.1E-10	Establishment of protein localization	B	105	1.1E-19	cold	-0.33	5.2E-09	old	-2.03	1.6E-12
								Actin filament-based process	B	51	1.9E-16	cold	-0.41	9.0E-06	old	-1.96	9.0E-06
								Regulation of programmed cell death	B	79	2.6E-07	average	-0.03	4.2E-01	average	-0.63	3.0E-02
6	748	hot	0.18	3.5E-22	young	1.34	2.8E-29	Lymphocyte activation	B	40	1.2E-11	hot	0.68	1.2E-10	young	3.27	6.2E-11
								Hemoglobin complex	P	13	1.3E-12	average	0.54	5.1E-03	young	4.43	1.1E-04
								Transmembrane	P	252	1.8E-05	hot	0.51	2.0E-34	young	2.55	4.5E-28
7	417	average	-0.08	1.2E-02	old	-1.54	8.3E-22	Mitochondrial envelope	C	103	1.0E-68	average	-0.14	2.0E-02	old	-1.98	4.5E-10
								Respiratory chain	C	44	4.9E-47	average	0.03	3.4E-01	old	-1.57	9.3E-05
								Ribosomal protein	P	62	2.4E-53	average	0.18	2.3E-02	old	-1.73	3.2E-05
8	296	hot	0.16	1.5E-05	young	1.49	4.8E-14	Homeobox	P	27	2.2E-12	average	-0.22	7.3E-02	young	3.01	3.7E-07
9	270	cold	-0.28	5.7E-14	old	-1.96	2.9E-20	Translational elongation	B	77	1.7E-114	cold	-0.60	6.0E-18	old	-3.04	1.7E-13
								rRNA processing	B	27	1.1E-22	average	-0.24	3.2E-02	old	-2.93	1.7E-05

Table 4.2 Evolutionary properties of communities and DAVID groups in AML 2.3. Gene evolutionary rates (ERs) take real values from 0 (most conserved) to approximately 6.9 (most variable), and ages take integer values from 0 (oldest) to 12 (youngest). The table is organized as follows. "Comm. Index" is the index of the ten largest communities. "Num. genes" is the number of genes in the community. "Comm. ER" indicates whether the community is significantly hotter (i.e. has a higher ER) or colder (i.e. has a lower ER) than the mean of 300 equally-sized sets of genes randomly selected from the network, with a significance threshold of $p = 10^{-3}$. "Diff. in mean" is the difference between the mean ER of the community and the mean ER of the 300 randomly selected sets. "p-value" is the significance of the difference. "Comm. age", "Diff. in mean", and "p-value" are the same as previously stated, but for age rather than ER. "DAVID group name" is the name of the DAVID group that DAVID identified as enriched in each community. “Group type” states whether the DAVID group is a protein type (P), location of final gene product (L), biological process (B), or cellular component (C). "Num. genes" is the number of genes in the DAVID group. "DAVID Benjamini" is the significance of the enrichment of the DAVID group, as reported by DAVID. The remaining DAVID group columns are computed in the same manner as the community columns.

Comm. index	*Num. genes*	*Comm. ER*	*Diff. in mean*	*p-value*	*Comm. age*	*Diff. in mean*	*p-value*	*DAVID group name*	*Group type*	*Num. genes*	*DAVID Benjamini*	*Group ER*	*Diff. in mean*	*p-value*	*Group age*	*Diff. in mean*	*p-value*
0	2961	hot	0.38	1.0E-185	young	2.05	7.1E-198	Immunoglobulin domain	P	201	7.6E-66	hot	0.92	4.0E-82	young	4.16	2.2E-58
								EGF-like domain	P	122	6.8E-55	hot	0.25	1.5E-04	young	1.02	3.3E-04
								Chemotaxis	B	93	1.5E-40	hot	0.86	4.7E-31	young	4.09	1.9E-23
1	2849	cold	-0.14	8.7E-30	average	0.10	8.2E-03	Protein kinase core	P	344	3.2E-226	cold	-0.30	1.5E-15	old	-1.69	2.0E-18
								Pos. reg. of transcr. from RNAP II promoter	B	181	4.0E-57	cold	-0.30	1.3E-08	young	0.69	4.6E-03
								Transcription regulation	B	503	4.2E-59	cold	-0.22	2.0E-11	young	1.04	1.8E-12
2	2665	cold	-0.12	9.1E-21	old	-1.50	5.3E-97	Mitochondrial envelope	C	211	2.7E-81	cold	-0.16	5.0E-04	old	-2.24	1.0E-23
								Protein transport	B	243	1.2E-51	cold	-0.35	2.6E-15	old	-2.20	1.9E-21
								Peroxisome	C	54	2.6E-30	average	-0.11	1.2E-01	old	-2.39	2.0E-06
3	1402	cold	-0.07	1.7E-05	old	-1.05	2.7E-28	Response to DNA damage stimulus	B	145	1.9E-71	average	0.03	3.3E-01	old	-2.35	2.0E-16
								Cellular macromolecule catabolic process	B	195	1.6E-70	cold	-0.31	5.4E-10	old	-2.10	7.2E-16
								Mitosis	B	82	1.2E-58	average	0.01	3.8E-01	old	-1.83	1.9E-06
4	1115	cold	-0.24	7.1E-30	old	-2.20	8.3E-89	Ribosome	C	68	7.2E-89	cold	-0.64	9.6E-16	old	-3.50	1.8E-15
								Nuclear lumen	L	241	3.6E-59	cold	-0.29	5.6E-13	old	-2.38	5.6E-27
								Nucleosome	C	61	2.5E-29	cold	-0.50	9.6E-06	average	-0.09	4.0E-01
5	781	cold	-0.26	1.6E-24	old	-1.08	8.1E-16	mRNA processing	B	151	1.4E-144	cold	-0.51	9.8E-25	old	-2.32	3.5E-15
								Nuclear lumen	L	185	1.1E-75	cold	-0.39	1.9E-15	old	-2.15	1.1E-15
								WD40 repeat region	P	83	5.1E-60	cold	-0.30	5.3E-05	old	-1.95	9.3E-07
6	736	cold	-0.09	6.1E-04	young	0.75	1.3E-07	Sulfotransferase	P	26	1.6E-33	average	0.06	3.4E-01	average	0.23	3.6E-01
								Vision	B	32	1.2E-16	average	0.03	3.9E-01	average	0.84	1.1E-01
								C2 calcium-dependent membrane targeting	P	25	6.7E-11	average	-0.14	1.7E-01	average	0.14	4.5E-01
7	731	hot	0.20	9.7E-16	average	0.17	1.1E-01	KRAB	P	168	9.5E-208	hot	0.67	4.9E-36	average	0.66	1.1E-02
								Zinc finger	P	205	3.4E-245	hot	0.62	1.4E-32	average	0.84	1.3E-03
								Domain: SCAN box	P	22	1.4E-20	hot	0.54	3.2E-04	young	4.82	8.1E-10
8	320	average	0.05	1.2E-01	young	0.70	4.6E-04	ANK repeat	P	58	2.4E-72	average	0.07	2.4E-01	average	0.15	3.9E-01

Table 4.3 Evolutionary properties of communities and DAVID groups in HumanNet. See Table 4.2 for explanation of column headers.

CHAPTER 5

THE HOPFIELD MODEL

Children and lunatics cut the Gordian knot which the poet spends his life patiently trying to untie.

—Jean Cocteau

The following chapter introduces a dynamical mathematical system, the Hopfield model, and is adapted from my 2014 publication, *Control of asymmetric Hopfield networks and application to cancer attractors* [141]. It is organized as follows.

- Section 5.1, *Why Boolean?*: the motivation behind using Boolean models to understand gene regulatory networks
- Section 5.2, *The Hopfield model*: a description of the canonical Hopfield model and some of its basic properties
- Section 5.3, *Cancer signalling and the Hopfield model*: an application of the Hopfield model to signalling in cancerous gene regulatory networks
- Section 5.4, *Conclusion*: a summary of the chapter's findings

For further details concerning the Hopfield model and its applications, see *Introduction to the Theory of Neural Computation* by John Hertz, Anders Krogh, and Richard Palmer [67].

5.1 Why Boolean?

The boom in the volume of biological data over the past decade has made mathematical modeling in biology both more feasible and more important. Analyzing genome-wide expression data increasingly requires the use of advanced clustering techniques and network models (and even network-based clustering techinques [97, 148]). However, standard static networks provide only a sort of time-averaged representation of gene regulation. In reality, GRNs are dynamical systems

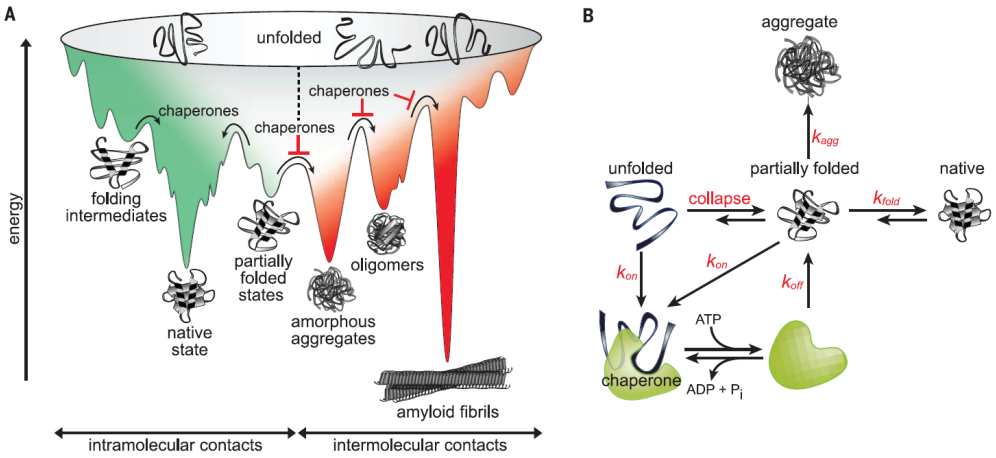


Figure 5.1 [A] Schematic of a protein energy landscape. The high dimensional configuration space has been collapsed to the horizontal axis, with isolated configurations drawn on the left and highly interacting configurations on the right. Local energetic minima are stable configurations, and transitions between configurations can occur due to thermal fluctuations, protein-protein interactions, and changes in the microenvironment. [B] Diagram of some of the possible transitions in the configuration of a protein. Image taken from [13].

that transmit information via molecular signals across networks. Predicting the effects of perturbations such as *gene knockouts* (disabling a gene or multiple genes in an organism to observe changes in expression and/or phenotype) or the application of *kinase inhibitors* (a class of drugs which blocks certain protein enzymes called *kinases*) requires a dynamical model.

It is perhaps natural for physical scientists, and particularly physicists, to try to design a first-principles mathematical model built from a handful of microscopic details to predict and explain macroscopic phenomena. The power of this approach is clear [158]: knowing the lattice structure of a crystal can lead to highly accurate predictions of X-ray diffraction patterns, and the critical temperature of a noninteracting Bose-Einstein condensate can be computed to great precision from just two parameters, the mass and the density of particles.

A similar *ab initio* approach is taken in the field of *protein folding*, the primary goal of which is predicting the folded shape of a protein from its sequence and local environment (including both small molecules and other macromolecules) alone [13]. Much effort has been devoted to understanding the *folding landscape*, shown schematically in Fig. 5.1A, where proteins transi-

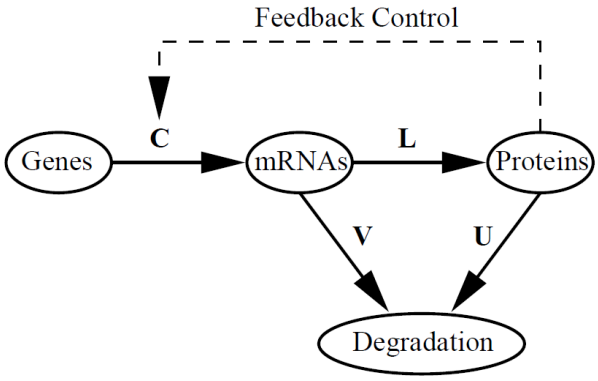


Figure 5.2 Schematic of Eq. 5.1. Genes produce mRNAs, mRNAs produce proteins, proteins regulate mRNA production rates, and mRNAs and proteins decay, all at varying rates. Image taken from [26].

tion between states on an energetic landscape and settle into local energetic minima, occasionally "hopping" to neighboring wells with the assistance of thermal fluctuations and interactions with *chaperone proteins* as shown in in Fig. 5.1B.

While the configuration of a protein is defined by many degrees of freedom, the vast majority of states are transient states. However, a great number of local minima pocket the landscape, some of which are energetically highly favorable but biologically damaging, such as the protein aggregates formed in mad cow disease. Furthermore, mutations which alter the amino acid sequence of encoded proteins can alter this landscape, making the normally functional configurations of the protein inaccessible, and creating new nonfunctional energetic minima. Given that a host of afflictions such as Huntington's disease and sickle cell disease are caused by the misfolding of a single protein (the proteins *huntingtin* and *HBB*, respectively), understanding the mechanics of protein folding is a crucial task for advancing medicine.

However, cells are highly complex, diverse, crowded, stochastic, non-equilibrium microenvironments whose processes and components span a broad spectrum of length and time scales, and the combinatorics of multi-gene diseases such as cancer make the problem far more difficult to analyze and understand from a molecular point of view. Even if all of the innumerable parameters that govern the dynamical processes of a particular cell were known, modelling the dynamics

of genome-wide GRNs atom-by-atom to predict the effects of mutations is computationally intractable. Accordingly, some approximations must be made.

An alternative to the molecular dynamics approach is modelling gene regulation with coupled differential equations [150]. One possible system of differential equations, shown schematically in Fig. 5.2, is [26]

$$\begin{aligned} \frac{d\vec{r}}{dt} &= f(\vec{p}) - V\vec{r} \\ \frac{d\vec{p}}{dt} &= L\vec{r} - U\vec{p} \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} N &= \text{the number of genes in the genome} \\ \vec{r} &= \text{mRNA concentrations, functions of } t \\ \vec{p} &= \text{protein concentrations, functions of } t \\ f(\vec{p}) &= \text{transcription functions} \\ L &= \text{translational constants, } N \times N \text{ diagonal matrix} \\ V &= \text{mRNA degradation rates, } N \times N \text{ diagonal matrix} \\ U &= \text{protein degradation rates, } N \times N \text{ diagonal matrix} \end{aligned} \tag{5.2}$$

While much simpler than large-scale molecular dynamics simulations, modelling the behavior of a signficant fraction of the genome using this approach requires knowledge of a great deal of parameters, and the model's nonlinear nature (buried in the $f(\vec{p})$ term) makes solving such systems computationally expensive. *Hill functions* are commonly used [75] to model protein-gene interactions. Activating Hill functions with a single regulatory protein, for example, take the form

$$f(p) = \frac{ap^n}{k^n + p^n} \tag{5.3}$$

for amplitude a, *Hill coefficient* n (a steepness parameter), and *apparent dissociation constant* k (the ratio of unbinding to binding rates for the protein/gene pair). A family of curves for $a = 1$ and $k = 5$ is shown in Fig. 5.3. For large n, the rate of transcription is sensitive to protein concentration only over a short range near k,[1] resulting in *switch-like* [47] gene regulation.

[1] For large n and given k, a Taylor expansion at $p = k$ shows that $f(p)$ has nearly zero slope everywhere except in the approximate range $k\left(1 - \frac{2}{n}\right) \lesssim p \lesssim k\left(1 + \frac{2}{n}\right)$.

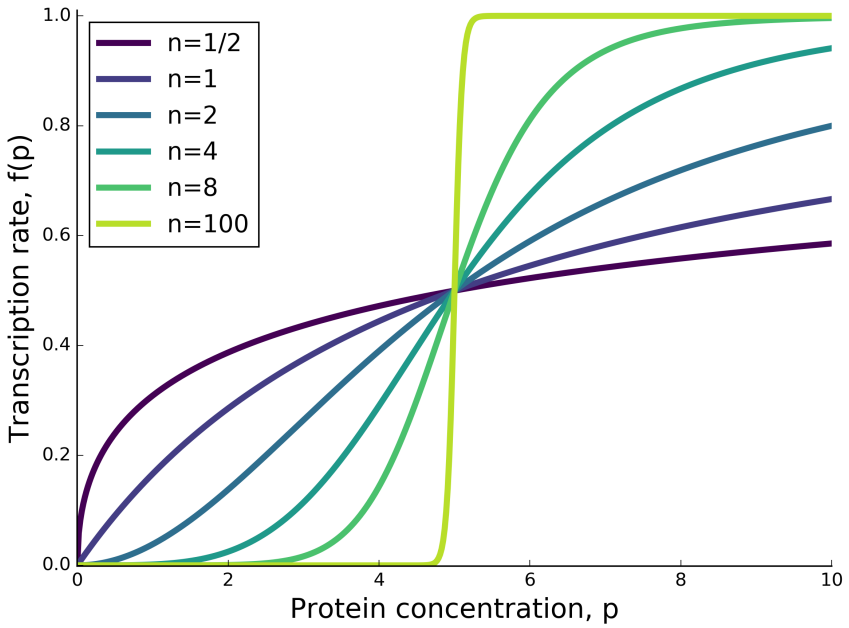


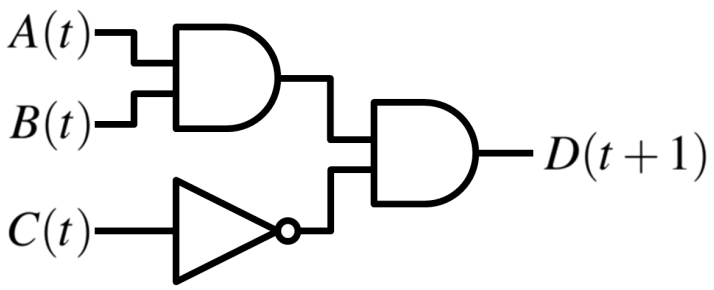
Figure 5.3 Family of Hill functions from Eq. 5.3 for $a = 1$ and $k = 5$.

As argued in Section 3.4, it may not be necessary to understand the precise details of every interaction in a GRN to understand some of its high level properties. Inspired by the approximate switch-like nature of some genes, *Boolean models* provide a simple way to interpret both the current state of a gene (either inactive or saturated) and the nature of interactions between multiple genes. Boolean models tend to be defined via discrete-time ($t \in \mathbb{Z}$) mappings of the form

$$\vec{\sigma}(t+1) = f(\vec{\sigma}(t)) \tag{5.4}$$

where N is the number of nodes, $\vec{\sigma}(t) \in \{0,1\}^N$, and the particular model is defined by the form of the mapping function $f: \{0,1\}^N \to \{0,1\}^N$ (mapping a Boolean vector to another Boolean vector).

One of the most famous Boolean models, the *Kauffman model* [74], also known as a *random Boolean network* (RBN), treats interactions between genes as random sets of digital logic gates embedded in a random network, and was created to demonstrate the complex dynamics that can emerge from a collection of many elements interacting via a simple set of rules. For example, a system in which the production of a protein D is activated by transcription factors A and B and is



$A(t)$	$B(t)$	$C(t)$	$D(t+1)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

Figure 5.4 Circuit and logic table for Eq. 5.5.

repressed by C could be modelled as

$$D(t+1) = [A(t) \text{ AND } B(t)] \text{ AND } [\text{NOT } C(t)] \tag{5.5}$$

as shown diagrammatically in Fig. 5.4.[2] Because the trajectory of an RBN is deterministic and confined to a finite number of configurations $\left(2^N\right)$, all states are guaranteed to eventually converge to *point attractors* (stable states) or *cyclic attractors* (periodic states).[3] Fig. 5.5 shows the dynamics of one such random Boolean NK network for $N = 100$, $K = 2$, and a random initial state. Although there are $2^{100} \approx 10^{30}$ unique configurations, the system settles into a cyclic attractor with period 4 in less than 20 time steps. The simple rules that govern RBNs generate rich, complex dynamical properties and have been studied extensively [41, 42, 135], including phase changes, stability (sensitivity to single bit flips), and capacity (average number of attractors per node) on various network topologies.

5.2 The Hopfield model

The Hopfield model is a kind of *associative memory model* that was introduced by the physicist JJ Hopfield in 1982 [68], and was designed to mimic the way a brain stores memories (attractors) and recalls them from particular stimuli, and has been applied to areas such as neural computation

[2] In Eq. 5.5, C is a *canalyzing variable*, since $C(t) = 1$ guarantees $D(t+1) = 0$ regardless of the state of the other inputs. Canalyzing variables have been shown to greatly increase the stability of the Kauffman model [73, 87, 121].

[3] A point attractor is a special case of a cyclic attractor with period 1.

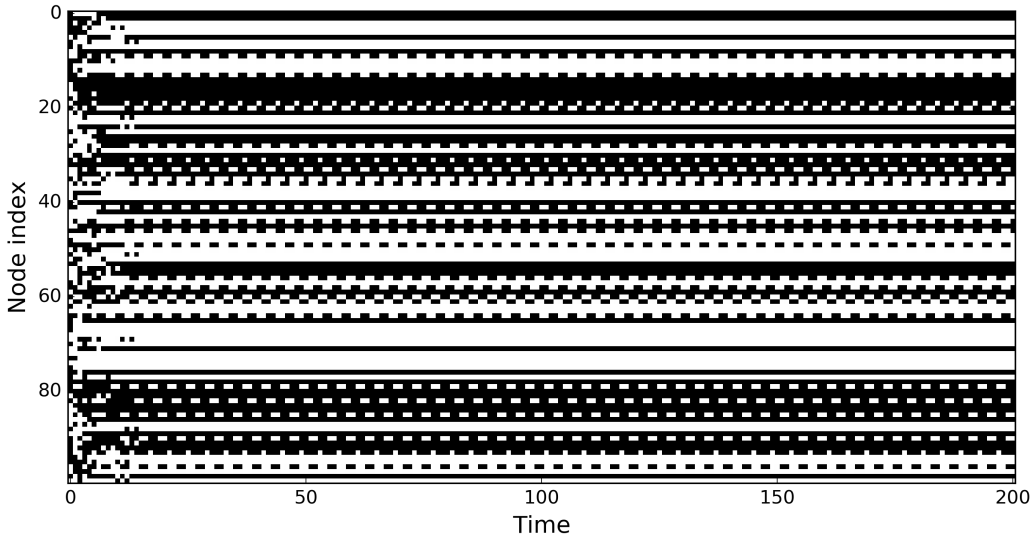


Figure 5.5 Trajectory of a random Boolean NK network with $N = 100$, $K = 2$, and a random initial state. Black and white pixels represent 0's and 1's. The system quickly settles into a cyclic attractor with period 4.

and image recognition. It belongs to the class of infinite-range *spin glass models* such as the the Sherrington-Kirkpatrick model [76]. Similarities and differences between the Kauffman and Hopfield models have been studied for many years [5, 83, 84, 126]. The most important distinction for the purposes of this dissertation is their differing abilities to encode data: RBNs define random interactions that give rise to emergent attractors, whereas *the Hopfield model begins with a set of desired attractor states and constructs a set of node-node couplings which guarantees them*. As with the protein folding energy landscape in Fig. 5.1A, the (symmetric) Hopfield model defines a high dimensional energy landscape with local energetic minima corresponding to programmed attractor configurations.[4]

Formally, the canonical Hopfield model is an Ising model whose configuration is defined by N spins $\sigma_i(t)$ at time $t \in \mathbb{Z}$. The state of each node takes one of two values, $\sigma_i(t) = \pm 1$ (on/off). The *coupling matrix*, which defines both the strength and the sign of the signal from node j to node i

[4]Only Hopfield systems with symmetric (i.e. undirected) connections can be visualized as energetic landscapes. Hopfield systems with directed edges still obey the same signaling rules, but may produce cyclic attractors.

(+ for ferromagnetic and − for antiferromagnetic), is constructed according to

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{5.6}$$

where p is the number of attractors, $\xi_i^\mu = \pm 1$ is the state of the i^{th} spin in the μ^{th} attractor, and the leading factor of $1/N$ is conventional. (Note that the superscript μ is an index, not an exponent.) A certain subset of interactions can be selected by introducing an adjacency matrix A_{ij} so that

$$J_{ij} = \frac{A_{ij}}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{5.7}$$

The total field at node i at time t is given by

$$h_i(t) = \sum_j J_{ij} \sigma_j(t) + h_i^{\text{ext}}(t) \tag{5.8}$$

where $\sum_j J_{ij}\sigma_j(t)$ is the internal field at node i due to its coupling with all nodes j, and $h_i^{\text{ext}}(t)$ is an optional external field applied to node i. The dynamical update rule is defined by

$$\sigma_i(t+1) = \begin{cases} +1 & \text{with probability } (1 + e^{-2h_i(t)/T})^{-1} \\ -1 & \text{otherwise} \end{cases} \tag{5.9}$$

(known as *Glauber dynamics*) where the factor of 2 in the exponent is conventional and T is the "temperature," i.e. the level of noise. Biologically, this noise represents the effects of all kinds of fluctuations present in cells. Note that for $h_i(t) \to \pm\infty$, $\sigma_i(t+1) = \pm 1$; and for $T \to \infty$, $\sigma_i(t+1) = \pm 1$ with equal probability. For $T \to 0$ (i.e. fully deterministic dynamics),

$$\sigma_i(t+1) = \text{sign}\,(h_i(t)) \tag{5.10}$$

A family of curves for Eq. 5.9 for various temperatures is shown in Fig. 5.6.

There are two main ways to apply the Hopfield update rule: the *synchronous* and *asynchronous* schemes. The synchronous scheme assumes that there is some sort of centralized "clock" that updates all nodes simultaneously at every time step, analogous to the clock signal used in synchronous digital circuits. In contrast, the asynchronous scheme selects one node i at random and

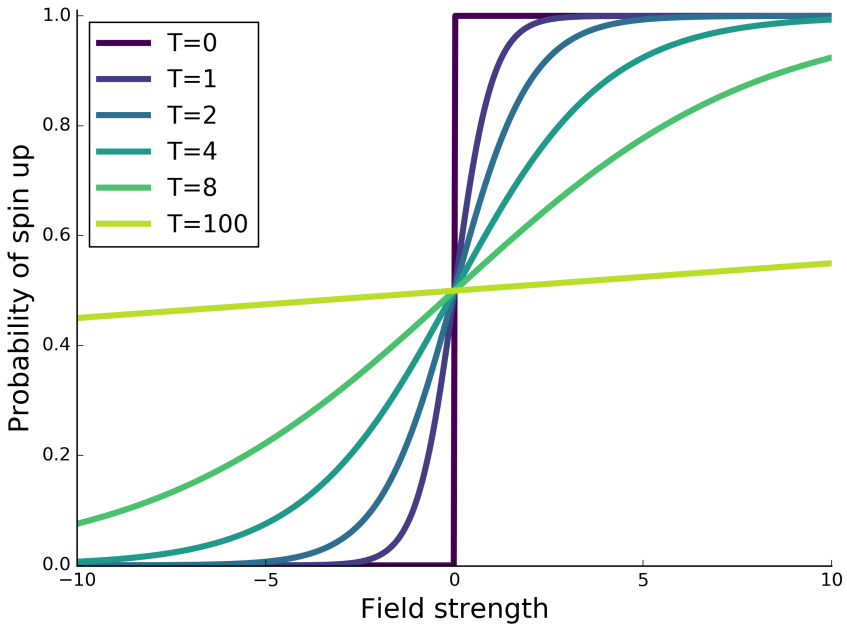


Figure 5.6 Family of functions from Eq. 5.9.

updates its state, then selects another (or possibly the same) node and updates its state, etc. Alternatively, a "quasisynchronous" scheme can be constructed by assigning a fixed probability q that each node is updated at each time step, so that the number of updates follows a binomial probability distribution and the expected number of updates at each time step is Nq. Note that for $q = 1$, the quasisynchronous scheme is the same as the synchronous scheme, whereas if $Nq \lesssim 1$, the asynchronous and quasisynchronous schemes give nearly identical results aside from statistical fluctuations around Nq. Because the Hopfield model was selected to simulate the dynamics of gene regulation, a stochastic system which lacks a central clock, the remainder of this chapter uses the asynchronous scheme.

For the sake of the following analysis, assume that all attractors are randomly and independently drawn from the space $\{-1,+1\}^N$ (so that $\xi_i^\mu = \pm 1$ with equal probability for any given i

and μ). Consider the special case of $p = 1$ and $T = 0$. If $\sigma_i(t) = \xi_i^1$, then

$$\begin{aligned}\sigma_i(t+1) &= \text{sign}\left(\sum_j J_{ij}\sigma_j(t)\right) \\ &= \text{sign}\left(\frac{1}{N}\xi_i^1 \sum_j \xi_j^1 \xi_j^1\right) \\ &= \text{sign}\left(\xi_i^1\right) \\ &= \xi_i^1\end{aligned} \tag{5.11}$$

which utilizes

$$\sum_{j=1}^{N}\left(\xi_j^1\right)^2 = \sum_{j=1}^{N} 1 = N \tag{5.12}$$

This means $\sigma_i = \xi_i^1$ is a *stationary point* in state space, but in order to be an attractor, it has to be a *stable point* as well (i.e. the system must relax to ξ_i^1 if perturbed). This is indeed the case: as long as more than half of the spins are in the state ξ_i^1, the field on all nodes will be in the direction of ξ_i^1, guaranteeing that $\sigma_i(t+1) = \sigma_i(t+2) = \ldots = \xi_i^1$. Note that if $+\xi_i^1$ is an attractor, $-\xi_i^1$ (its mirror state) is also an attractor. This is a basic symmetry of the system and is true for all attractors μ regardless of the number of attractors p. Only the application of a nonzero external field h_i^{ext} can break this symmetry.

More generally, the stability of an attractor ξ_i^μ in a system with an arbitrary number of attractors p can also be computed. It is helpful to break the coupling matrix into two terms, $J_{ij} = N^{-1}\left(\xi_i^\mu \xi_j^\mu + \sum_{\nu\neq\mu} \xi_i^\nu \xi_j^\nu\right)$. If $\sigma_i(t) = \xi_i^\mu$, then the internal field at node i is

$$\begin{aligned}h_i(t) &= \frac{1}{N}\xi_i^\mu \sum_j \xi_j^\mu \xi_j^\mu + \frac{1}{N}\sum_{\nu\neq\mu}\sum_j \xi_i^\nu \xi_j^\nu \xi_j^\mu \\ &= \xi_i^\mu + \frac{1}{N}\sum_{\nu\neq\mu}\sum_j \xi_i^\nu \xi_j^\nu \xi_j^\mu \\ &= \xi_i^\mu\left(1 + \frac{1}{N}\xi_i^\mu \sum_{\nu\neq\mu}\sum_j \xi_i^\nu \xi_j^\nu \xi_j^\mu\right) \\ &\equiv \xi_i^\mu\left(1 + C_i^\mu\right)\end{aligned} \tag{5.13}$$

This is the same as the $p = 1$ case, plus an additional *crosstalk* term $\xi_i^\mu C_i^\mu$ which depends on the patterns' particular values. As long as the term C_i^μ is greater than -1, it cannot change the

sign of the field at node i, making ξ_i^μ a stationary point. Furthermore, the pattern ξ_i^μ is stable to small perturbations as long as the self-term is not overwhelmed by the crosstalk term. Because the crosstalk term is a sum of many equally probable ± 1's for large N and p, the central limit theorem implies that its sum follows a normal distribution with mean zero and standard deviation $\sqrt{p/N}$. The probability that the crosstalk term flips a given bit i away from ξ_i^μ after one time step is thus the probability that $C_i^\mu < -1$,

$$P_{\text{single flip}} = \frac{1}{2}\left[1 - \text{erf}\left(\sqrt{\frac{N}{2p}}\right)\right] \tag{5.14}$$

This means that for $N \to \infty$, $p \to \infty$, and a single-node error tolerance of $P_{\text{single flip}} = 0.001$, the maximum number of attractors per node is $p/N = 0.105$. However, $NP_{\text{single flip}}$ only represents the expected number of flips after one time step. Each of these subsequent errors can also produce more errors in the next time step, and even more in the next time step, etc., leading to a cascade of spin flips that drives the system away from the desired attractor state ξ_i^μ. It can be shown via a laborious mean field solution [6, 67] that the *maximum capacity* (also known as the *load parameter*) is $\alpha_c \equiv p/N = 0.138$, beyond which errors compound and destabilize states intended to be attractors. The asymmetric case (i.e. using a directed network A_{ij}) can also be exacty solved in some limits [40]. See Appendix F for a derivation of the Hopfield mean field equation on directed Erdős-Rényi networks, as well as my unpublished derivation of the Hopfield mean field equation on modular directed Erdős-Rényi networks.

5.3 Cancer signalling and the Hopfield model

The Hopfield model has been used to model biological processes of current interest such as the reprogramming of pluripotent stem cells [85]. Moreover, it has been suggested that a biological system in a chronic or therapy-resistant disease state can be seen as a network that has become trapped in a pathological Hopfield attractor [7]. To a major extent, the final determinant of a cell's phenotype is its gene expression profile, which is relatively stable, reproducible, unique to cell types, and can differentiate cancer cells from normal cells, as well as differentiate between differ-

ent types of cancer [21, 46]. In fact, there is evidence that attractors exist in cells' gene expression, and as with protein folding, these attractors can be reached by different trajectories rather than only by a single transcriptional program [69]. While the dynamical attractors paradigm was originally proposed in the context of cellular development, the similarity between cellular *ontogenesis* (the development of different cell types) and oncogenesis (the process under which normal cells are transformed into cancer cells) has been recently emphasized [140]. The main hypothesis of this chapter is that cancer robustness is rooted in the dynamical robustness of signaling in an underlying cellular network. If the cancerous state of rapid, uncontrolled growth is an attractor state of the system [9], a goal of modelling therapeutic control could be to design complex therapeutic interventions based on drug combinations [51] that push the cell out of the cancer attractor basin [31].

Many authors have discussed the control of biological signaling networks using complex external perturbations. Calzolari and coworkers considered the effect of complex external signals on apoptosis signaling [23]. Agoston and coworkers [2] suggested that perturbing a complex biological network with partial inhibition of many targets could be more effective than the complete inhibition of a single target, and explicitly discussed the implications for multi-drug therapies [32]. In the traditional approach to control theory [137], the control of a dynamical system consists in finding the specific input temporal sequence required to drive the system to a desired output. This approach has been discussed in the context of RBNs [4] and their attractor states [27]. Several studies have focused on the intrinsic global properties of control and hierarchical organization in biological networks [52, 18]. A recent study has focused on the minimum number of nodes that needs to be addressed to achieve the complete control of a network [94]. This study used a linear control framework, a matching algorithm [124] to find the minimum number of controllers, and a replica method to provide an analytic formulation consistent with the numerical study. Finally, Cornelius et al. [30] discussed how nonlinearity in network signaling allows reprogramming a system to a desired attractor state even in the presence of constraints in the nodes that can be accessed by external control. This novel concept was explicitly applied to a T-cell survival signaling net-

work to identify potential drug targets in T-cell large granular lymphocyte (T-LGL) leukemia. The approach in the present paper is based on nonlinear signaling rules and takes advantage of some useful properties of the Hopfield formulation. In particular, by considering two attractor states it will be shown that the network separates into two types of domains which do not interact with each other. Moreover, the Hopfield framework allows for a direct mapping of a gene expression pattern into an attractor state of the signaling dynamics, facilitating the integration of genomic data in the modeling.

Originally in [141] and reproduced here, I consider an asymmetric Hopfield model using two GRNs, mapping gene expression data from normal and cancer cells to attractor states. Focus was placed on the question of controlling of a network's final state (after a transient period) using external local fields representing therapeutic interventions, e.g. gene knockouts or kinase inhibitors. General strategies aiming at selectively disrupting the signaling only in cells that are near a cancer attractor state are investigated. The strategies use the concept of *bottlenecks*, which identify single nodes or strongly connected clusters of nodes that have a large impact on the signaling. A theorem that places bounds on the minimum number of nodes that guarantee control of a bottleneck consisting of a strongly connected component is also provided. This theorem is useful for practical applications since it helps to establish whether an exhaustive search for such minimal set of nodes is practical. These strategies are then applied to lung and B cell cancers. Two different networks are used for this analysis. The first is an experimentally validated and non-specific network (that is, the observed interactions are compiled from many experiments conducted on heterogeneous cell types) obtained from a kinase interactome and phospho-protein database [161] combined with a database of interactions between transcription factors and their target genes [100]. The second network is cell-specific and was obtained using network reconstruction algorithms and transcriptional and post-translational data from mature human B cells [89]. The algorithmically reconstructed network is significantly more dense than the experimental one, and the same control strategies produce different results in the two cases.

5.3.1 Mathematical details

See Table 5.4 for a list and description of all mathematical symbols introduced in this chapter. In order to derive analytical results, the remainder of this chapter makes the simplifying assumption that the signaling in GRNs is fully deterministic, i.e. $T = 0$. Real biological systems are stochastic, however, and the importance of this stochasticity is covered in Chapter 6. It is convenient to introduce a new kind of notation for networks in addition to the adjacency matrix A_{ij}. Let $G = \{V(G), E(G)\}$ be a network composed of the set of vertices (nodes) $V(G)$ and the set of edges $E(G) = \{(j,i) : j \to i\}$. Note that because the networks of interest in this chapter are directed, care must be taken concerning *source nodes* (nodes with zero indegree). Source nodes are fixed to their initial states by a small external field so that $\sigma_q(t) = \sigma_q(0)$ for all $q \in Q$, where Q is the set of source nodes. However, the source nodes flip if directly targeted by an external field. Biologically, genes at the "top" of a network are assumed to be controlled by elements outside of the network.

In this application, two attractors are needed. Define these states as $\vec{\xi}^n$ and $\vec{\xi}^c$, the *normal state* and *cancer state*, respectively. The *magnetization* (also called the *alignment* or *overlap*) along attractor state a is given by

$$m^a(t) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i(t)\xi_i^a \ . \tag{5.15}$$

Note that if $m^a(t) = \pm 1$, $\vec{\sigma}(t) = \pm\vec{\xi}^a$. Also define the steady state magnetization along state a as

$$m_\infty^a = \lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=1}^{\tau} m^a(t) \ . \tag{5.16}$$

There are two ways to model normal and cancer cells. One way is to simply define a different coupling matrix for each attractor state a,

$$J_{ij}^a = \frac{A_{ij}}{N}\xi_i^a\xi_j^a \ . \tag{5.17}$$

Alternatively, both attractor states can be encoded in the same coupling matrix,

$$J_{ij} = \frac{A_{ij}}{N}\left(\xi_i^n\xi_j^n + \xi_i^c\xi_j^c\right) \ . \tag{5.18}$$

Systems using Eqs. 5.17 and 5.18 will be referred to as the one attractor state ($p = 1$) and two attractor state ($p = 2$) systems, respectively. An interesting property emerges when $p = 2$. Consider

a simple network composed of two nodes, with only one edge $1 \to 2$ with attractor states $\vec{\xi}^n$ and $\vec{\xi}^c$, and $T = 0$. The only nonzero entry of the matrix J_{ij} is

$$J_{21} = \frac{1}{N}\left(\xi_2^n \xi_1^n + \xi_2^c \xi_1^c\right) \ . \tag{5.19}$$

Note that if $\vec{\xi}^n = \pm\vec{\xi}^c$, $J_{21} = 2\xi_2^n \xi_1^n$. In either case, by Eq. 5.9,

$$\sigma_2(t+1) = \begin{cases} +\xi_2^n & \text{if } \sigma_1(t) = +\xi_1^n \\ -\xi_2^n & \text{if } \sigma_1(t) = -\xi_1^n \end{cases} \ , \tag{5.20}$$

that is, the spin of node 2 at a given time step will be driven to match the attractor state of node 1 at the previous time step. However, if $\xi_1^n = \pm\xi_1^c$ and $\xi_2^n = \mp\xi_2^c$, $J_{21} = 0$. This gives

$$\sigma_2(t) = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases} \tag{5.21}$$

In this case, node 2 receives no input from node 1. Nodes 1 and 2 have become effectively disconnected.

This motivates new designations for node types. Define *similarity nodes* as nodes with $\xi_i^n = \xi_i^c$, and *differential nodes* as nodes with $\xi_i^n = -\xi_i^c$. Additionally, define the set of similarity nodes $S = \left\{i : \xi_i^n = \xi_i^c\right\}$ and the set of differential nodes $D = \left\{i : \xi_i^n = -\xi_i^c\right\}$. Connections between two similarity nodes or two differential nodes remain in the network, whereas connections that link nodes of different types transmit no signals. The effective deletion of edges between nodes means that the original network fully separates into two subnetworks: one composed entirely of similarity nodes (the *similarity network*) and another composed entirely of differential nodes (the *differential network*), each of which can be composed of one or more separate weakly connected components (see Fig. 5.7). With this separation, new source nodes (*effective sources*) can be exposed in both the similarity and differential networks. For the remainder of this article, Q is the set of both source and effective source nodes in a given network.

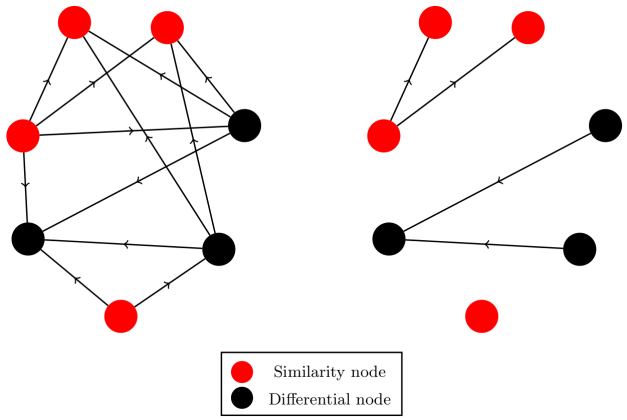


Figure 5.7 Network segregation for two attractor states ($p = 2$). Every edge that connects a similarity node to a differential node or a differential node to a similarity node transmits no signal. This means that the signaling in the right network shown above is identical to that of the left network. Because the goal is to leave normal cells unaltered while damaging cancer cells as much as possible, all similarity nodes can be safely ignored, and searches and simulations only need to be done on the differential subnetwork.

5.3.2 Control Strategies

The strategies presented below focus on selecting the best single nodes or small clusters of nodes to control, ranked by how much they individually change m_∞^a. In application, however, controlling many nodes is necessary to achieve a sufficiently changed m_∞^a. The effects of controlling a set of nodes can be more than the sum of the effects of controlling individual nodes, and predicting the truly optimal set of nodes to target is computationally difficult. Heuristic strategies are discussed for controlling large networks where the combinatorial approach is impractical.

For both $p = 1$ and $p = 2$, simulating a cancer cell means that $\vec{\sigma}(0) = +\vec{\xi}^c$ (i.e. starting in the cancer state), and likewise for normal cells. Although the normal and cancer states are mathematically interchangeable, biologically useful results are obtained by decreasing m_∞^c as much as possible while leaving $m_\infty^n \approx +1$. "Network control" thus means driving the system away from its initial state of $\vec{\sigma}(0) = \vec{\xi}^c$ with $\vec{h}^{\text{ext}}$. Controlling individual nodes is achieved by applying a strong field (stronger than the magnitude of the field due to the node's upstream neighbors) to a set

of targeted nodes T so that

$$h_\tau^{\text{ext}} = \begin{cases} \lim_{(u\to\infty)} -u\xi_\tau^c & \text{if } \tau \in T \\ 0 & \text{otherwise} \end{cases} . \tag{5.22}$$

This ensures that the drug field can always overcome the field from neighboring nodes.

In application, similarity nodes are never deliberately directly targeted, since changing their state would adversely affect both normal and cancer cells. Roughly 70% of the nodes in the networks surveyed are similarity nodes, so the search space is reduced. For $p = 2$, the effective edge deletion means that only the differential network in cancer cells needs to be simulated to determine the effectiveness of $\vec{h}^{\text{ext}}$. For $p = 1$, however, there may be some similarity nodes that receive signals from upstream differential nodes. In this case, the full effect of $\vec{h}^{\text{ext}}$ can be determined only by simulating all differential nodes as well as any similarity nodes downstream of differential nodes. All following discussion assumes that all nodes examined are differential, and therefore targetable, for both $p = 1$ and $p = 2$. The existence of similarity nodes for $p = 1$ only limits the set of targetable nodes.

5.3.2.1 Directed acyclic networks

Full control of a directed acyclic network is achieved by forcing $\sigma_q = -\xi_q^c$ for all $q \in Q$. This guarantees $m_\infty^c = -1$. Suppose that nodes $q \in Q$ in an acyclic network have always been fixed away from the cancer state, that is, $\sigma_q(t \to -\infty) = -\xi_q^c$. For any node i to have $\sigma_i(t) = \xi_i^n$, it is sufficient to have either $i \in Q$ or $\sigma_j(t-1) = \xi_j^n$ for all $j \to i$, $i \notin Q$. Because there are no cycles present, all upstream paths of sufficent length terminate at a source. Because the spin of all nodes $q \in Q$ point away from the cancer attractor state, all nodes downstream must also point away from the cancer attractor state. Thus, for acyclic networks, forcing $\sigma_q = -\xi_q^c$ guarantees $m_\infty^c = -1$. The complications that arise from cycles are discussed in the next section. However, controlling nodes in Q may not be the most efficient way to push the system away from the cancer basin of attraction and, depending on the control limitations, it may not be possible. If minimizing the number of controllers is required, searching for the most important bottlenecks is a better strategy.

Consider a directed network G and an initially identical copy, $G' = G$. If removing node i (and all connections to and from i) from G' decreases the indegree of at least one node $j \in V(G')$, $j \neq i$, to less than half of its indegree in network G, $\{i\}$ is a *size 1 bottleneck*. The *bottleneck control set* of bottleneck $\{i\}$, $L(i)$, is defined algorithmically as follows: (1) Begin a set $L(i)$ with the current bottleneck i so that $L = \{i\}$; (2) Remove bottleneck $\{i\}$ from network G'; (3) Append $L(i)$ with all nodes j with current indegree that is less than half of that from the original network G; (4) Remove all nodes j from the network G'. If additional nodes in G' have their indegree reduced to below half of their indegree in G, go to step 3. Otherwise, stop. The *impact of the bottleneck i*, $I(i)$, is defined as

$$I(i) = |L(i)| \,, \tag{5.23}$$

where $|X|$ is the cardinality of the set X. The impact of a bottleneck is the minimum number of nodes that are guaranteed to switch away from the cancer state when the bottleneck is forced away from the cancer state.

The impact is used to rank the size 1 bottlenecks by importance, with the most important as those with the largest impact. In application, when searching for nodes to control, any size 1 bottleneck $\{i\}$ that appears in the bottleneck control set of a different size 1 bottleneck $\{j\}$ can be ignored, since fixing j to the normal state fixes i to the normal state as well.

The network in Fig. 5.8, for example, has three sources (nodes 1, 2 and 3), but one important bottleneck (node 6). If maximum damage, i.e. $m_\infty^c = -1$, is required, then control of all source nodes is necessary. If minimizing the number of directly targeted nodes is important and $m_\infty^c > -1$ can be tolerated, then control of the bottleneck node 6 is a better choice.

5.3.2.2 Directed cycle-rich networks

Not all networks can be fully controlled at $T = 0$ by controlling the source nodes, however. If there is a cycle present, paths of infinite length exist and the final state of the system may depend on the initial state, causing parts of the network to be hysteretic. Controlling only sources in a general directed network thus does not guarantee $m_\infty^c = -1$ unless the system begins with $\sigma_i = -\xi_i^c$.

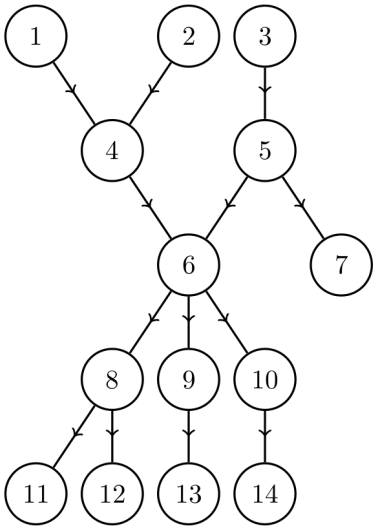


Figure 5.8 A directed acyclic network. Controlling all three source nodes (nodes 1, 2 and 3) guarantees full control of the network, but are ineffective when targeted individually. The best single node to control in this network is node 6 because it directly controls all downstream nodes.

Define a *cycle cluster*, C, as a strongly connected subnetwork of a network G. The network in Fig. 5.9, for example, has one cycle cluster with nodes $V(C) = \{4,5,6,7\}$. If the network begins with $\vec{\sigma}(0) = \vec{\xi}^c$, forcing both source nodes away from the cancer state does nothing to the nodes downsteam of node 3 (see Fig. 5.10). This is because the indegree $\deg^-(4) = 4$, and a majority of the nodes connecting to node 4 are in the cancer attractor state. At $T = 0$, cycle clusters with high connectivity tend to block incoming signals from outside of the cluster.

The most effective single node to control in this network is any one of nodes 4, 6 or 7. Forcing any of these away from the cancer attractor state will eventually cause the entire cycle cluster to flip away from the cancer state, and all nodes downstream will flip as well, as shown in Fig. 5.10. This cycle cluster act as a large, hysteretic bottleneck that motivates a generalization of size 1 bottlenecks.

Define a *size k bottleneck* in a network G to be a cycle cluster B with $|V(B)| = k$ which, when removed from G, reduces the indegree of at least one node $j \in V(G)$, $j \notin V(B)$ to less than half of its original indegree. Other than now using the set of nodes $V(B)$ rather than a single node set, the above algorithm for finding the bottleneck control set remains unchanged. In Fig. 5.9, for instance,

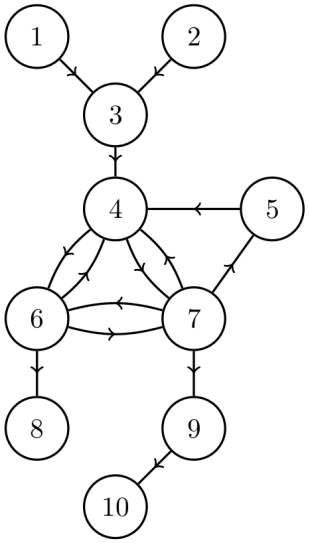


Figure 5.9 A network in which nodes 4, 5, 6 and 7 compose a single cycle cluster. The high connectivity of node 4 prevents any changes made to the spin of nodes 1-3 from propagating downstream. The only way to indirectly control nodes 8-10 is to target nodes inside of the cycle cluster. Targeting node 4, 6 or 7 will cause the entire cycle cluster to flip away from its initial state, guaranteeing control of nodes 4-10 (see Fig. 5.10).

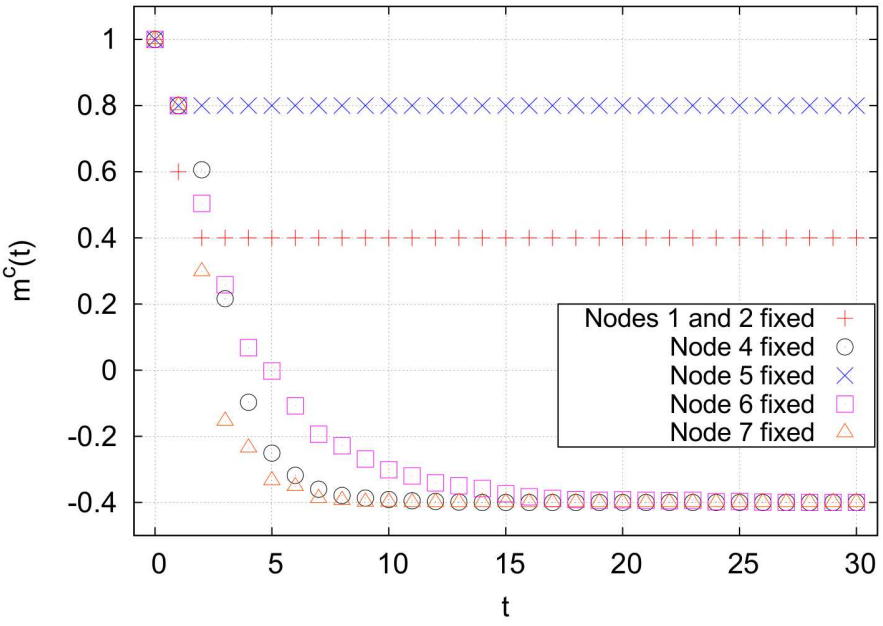


Figure 5.10 Cancer magnetization from targeting various nodes in the network shown in Fig. 5.9, averaged over 10,000 runs. The averaging removes fluctuations due to the random flipping of nodes with $h_i = 0$. Targeting node 7 results in the quickest stabilization, but targeting any one of nodes 4, 6 or 7 results in the same final magnetization.

$V(B) = \{4,5,6,7\}$, $k = 4$, $L(B) = \{4,5,6,7,8,9,10\}$, and $I(B) = 7$. Note that controlling any size k bottleneck B guarantees control of all size 1 bottlenecks B' in the control set of B for all $k \geq 1$.

For any bottleneck B of size $k \geq 1$ in a network G, define the *set of critical nodes*, $Z(B,G)$, as the set of nodes $Z(B,G) \subseteq V(B)$ of minimum cardinality that, when controlled, guarantees full control of all nodes $i \in V(B)$ after a transient period. Also define the *critical number of nodes* as $n_{\text{crit}}(B,G) = |Z(B,G)|$. Thus, for the network in Fig. 5.9, $Z(B,G) = \{4\}$, $\{6\}$, or $\{7\}$, and $n_{\text{crit}}(B,G) = 1$.

In general, however, more than one node in a cycle cluster may need to be targeted to control the entire cycle cluster. Fig. 5.11 shows a cycle cluster (composed of nodes 2-10) that cannot be controlled by targeting any single node. The precise value of n_{crit} for a given cycle cluster C depends on its topology as well as the edges connecting nodes from outside of C to the nodes inside of C, and finding $Z(C,G)$ can be difficult. The following theorem places bounds on n_{crit} to help determine whether a search for $Z(C,G)$ is practical.

Theorem:[5] Suppose a network G contains a cycle cluster C. Define the *set of externally influenced nodes*

$$R(C,G) = \{i \in V(C) : j \in V(G \setminus C), (j,i) \in E(G)\} , \tag{5.24}$$

the *set of intruder connections*

$$W(C,G) = \{(j,i) \in E(G) : i \in V(C), j \in V(G \setminus C)\} , \tag{5.25}$$

and the *reduced set of critical nodes*

$$Z_{\text{red}}(C,G) = Z(C, G \setminus W) . \tag{5.26}$$

If $N = |V(C)|$ and

$$\mu \equiv \min_{i \in V(C)} \deg^-(i) , \tag{5.27}$$

where $\deg^-(i)$ is computed ignoring intruder connections, then

$$\left\lceil \frac{\mu}{2} \right\rceil \leq n_{\text{crit}}(C,G) \leq \zeta , \tag{5.28}$$

[5]This is admittedly a rather technical theorem. It won't hurt my feelings if you don't want to parse it out.

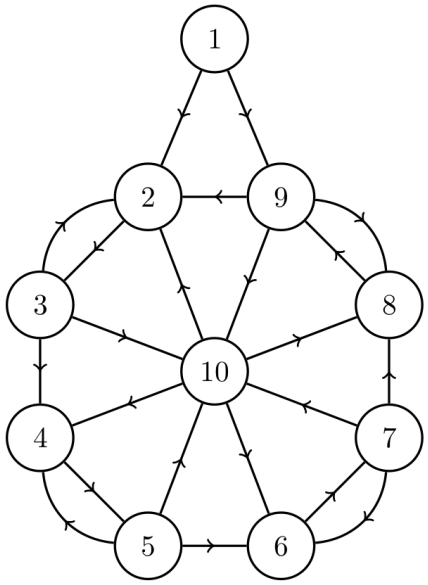


Figure 5.11 A network with a cycle cluster C, composed of nodes 2-10, that cannot be controlled at $T = 0$ by controlling any single node. Here, the set of externally influenced nodes is $R(C,G) = \{2,9\}$, the set of intruder connections is $W(C,G) = \{(1,2),(1,9)\}$, the reduced set of critical nodes is $Z_{\text{red}}(C,G) = \{9,10\}$, the minimum indegree is $\mu = 1$ and the number of nodes in the cycle cluster is $N = 9$. By Eq. 5.28, this gives the bounds of the critical number of nodes to be $1 \leq n_{\text{crit}} \leq 6$.

where

$$\zeta \equiv \min\left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\text{red}}(C,G)|, N\right) . \tag{5.29}$$

See Appendix E for the proof.

There can be more than one Z_{red} for a given cycle cluster. Note that the tightest constraints on n_{crit} in Eq. 5.28 come from using the Z_{red} with the largest overlap with R. If finding Z_{red} is too difficult, an overestimate for the upper limit of n_{crit} can be made by assuming that $R \cap Z_{\text{red}} = \{\varnothing\}$ so that

$$\left\lceil \frac{\mu}{2} \right\rceil \leq n_{\text{crit}}(C,G) \leq \min\left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G)|\ ,\ N\right) . \tag{5.30}$$

The cycle cluster in Fig. 5.11 has $N = 9$, $R = \{2,9\}$, $\mu = 1$, and one of the reduced sets of critical nodes is $Z_{\text{red}} = \{9,10\}$, so $1 \leq n_{\text{crit}} \leq 6$. It can be shown through an exhaustive search that for this network $n_{\text{crit}} = 2$, and the set of critical nodes is $Z = \{9,10\}$ (see Fig. 5.12). Here, $Z = Z_{\text{red}}$, although this is not always the case. Because the cycle cluster has 9 nodes and $1 \leq n_{\text{crit}} \leq 6$, at

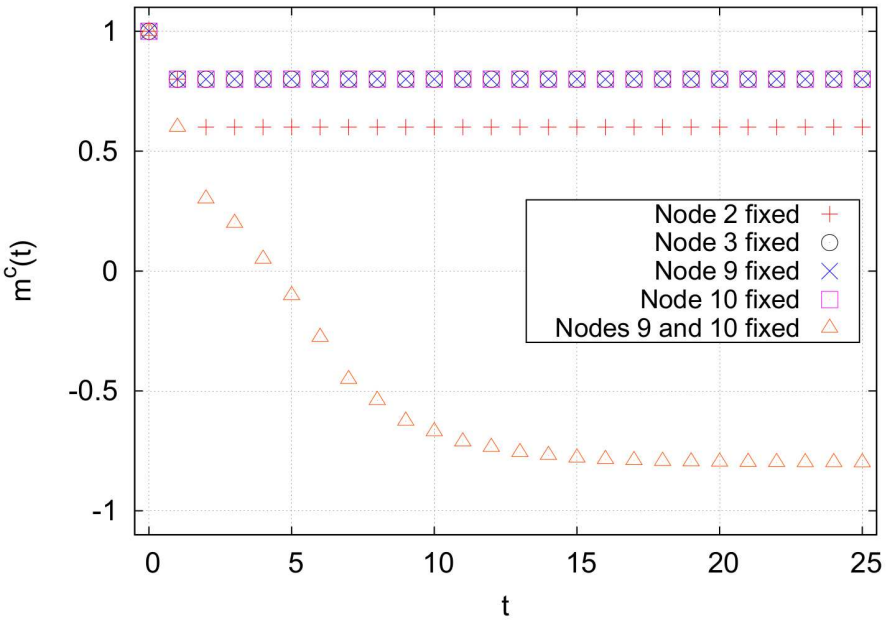


Figure 5.12 Magnetization for network from Fig. 5.11, averaged over 10,000 runs. There is no single node to target that will control the cycle cluster, but fixing nodes 9 and 10 results in full control of the cycle cluster, leaving only node 1 in the cancer state. This means $Z(C,G) = \{9,10\}$ and $n_{\text{crit}} = 2$.

most $\sum_{n=1}^{6} \binom{9}{n} = 465$ simulations are needed to find at least one solution for $Z(C,G)$. However, the maximum number of simulations required to find $Z(C,G)$ increases exponentially and for larger networks the problem quickly becomes intractable.

One heuristic strategy for controlling cycle clusters is to look for size $k' < |V(C)|$ bottlenecks inside of C. Bottlenecks of size $k \gg 1$ and average indegree $\langle \deg^-(B) \rangle \ll k$ can contain high impact size k' bottlenecks, where $k' < k$. Size $k \geq 1$ bottlenecks need to be compared to find the best set of nodes to target to reduce m_∞^C. Simply comparing the impact is insufficent because a cycle cluster with a large impact could also have a large n_{crit}, requiring much more effort than its impact merits. Define the *critical efficiency* of a bottleneck B as

$$e_{\text{crit}}(B) = \frac{I(B)}{n_{\text{crit}}(B,G)} \,. \tag{5.31}$$

If the critical efficiency of a cycle cluster is much smaller than the impacts of size 1 bottlenecks from outside of the cycle cluster, the the cycle cluster can be safely ignored.

For some cycle clusters, however, not all of the nodes need to be controlled in order for a large portion of the nodes in the cycle cluster's control set to flip. Define the *optimal efficiency* of a bottleneck B as

$$e_{\text{opt}}(B) = \max_{n=1,2,\ldots} \left(\frac{I\left(\bigcup_{i=1}^{n} B_i\right)}{n} \right) \tag{5.32}$$

where $B_i \subseteq V(B)$ are size 1 bottlenecks and $I(B_i) > I(B_{i+1})$ for all i. Note that for any size 1 bottleneck B, $e_{\text{opt}}(B) = e_{\text{crit}}(B) = I(B)$. This quantity thus allows bottlenecks with very different properties ($I(B)$, $n_{\text{crit}}(B,G)$, or $|V(B)|$) to be ranked against each other.

All strategies presented above are designed to select the best individual or small group of nodes to target. Significant changes in the biological networks' magnetization require targeting many nodes, however. Brute force searches on the effect of larger combinations of nodes are typically impossible because the required number of simulations scales exponentially with the number of nodes. A crude Monte Carlo search is also numerically expensive, since it is difficult to sample an appreciable portion of the available space. One alternative is to take advantage of the bottlenecks that can be easily found, and rank all size $k \geq 1$ bottlenecks B_i in an ordered list U such that

$$U = (B_1, B_2, B_3, \ldots) \tag{5.33}$$

where

$$e_{\text{opt}}(B_i) \geq e_{\text{opt}}(B_{i+1}) \text{ , } B_i \not\subset L(B_j) \tag{5.34}$$

for all $B_i, B_j \in U$ and fix the bottlenecks in the list in order. This is called the *efficiency-ranked* strategy. If all size $k > 1$ bottlenecks are ignored, it is called the *pure* efficiency-ranked strategy, and if size $k > 1$ bottlenecks are included it is called the *mixed* efficiency-ranked strategy.

An effective polynomial-time algorithm for finding the top z nodes to fix, which called the *best+1* strategy (a kind of greedy algorithm), works as follows: (1) Begin with a seed set of nodes to fix, F; (2) Test the effect of fixing $F \cup i$ for all allowed nodes $i \notin F$; (3) $F \leftarrow F \cup i_{\text{best}}$, where i_{best} is the best node from all i sampled; (4) If $|F| < z$, go to step (2). Otherwise, stop. The seed set of nodes could be the single highest impact size 1 bottleneck in the network, or it could be the best set of n nodes (where $n < z$) found from a brute force search.

5.3.3 Cancer signaling

In application to biological systems, the magnetization of cell type a is assumed to be related to the *viability* of cell type a, v^a; that is, the fraction of cells of type a that survives a drug treatment. It is reasonable to assume that v^a is a monotonically increasing function of m_∞^a. As few controllers as possible should be used to sufficiently reduce m_∞^c while leaving $m_\infty^n \approx +1$. In practical applications, however, the set of druggable targets is limited. All classes of drugs are constrained to act only on a specific set of biological components. For example, kinase inhibitors have two constraints: the only nodes that can be targeted are those that correspond to kinases, and they can only be inhibited, i.e. turned off. The example of kinase inhibitors will be used to show how control is affected by such constraints. In the real systems studied, many differential nodes have only similarity nodes upstream and downstream of them, while the remaining differential nodes form one large cluster. For $p = 2$, focus is placed on controlling only the largest weakly connected differential subnetwork, ignoring all *islets* (isolated nodes). All final magnetizations are normalized by the total number of nodes in the full network, even if the simulations are only conducted on small portion of the network.

5.3.3.1 Lung Cell Network

The network used to simulate lung cells was built by combining the kinase interactome from PhosphoPOINT [161] with the transcription factor interactome from TRANSFAC [100]. Both of these are general networks constructed from experimentally observed interactions. This bottom-up approach means that some edges may be missing, but those present are reliable. Because of this, the network is sparse ($\sim 0.057\%$ complete, see Table 5.1), resulting in the formation of many islets for $p = 2$. Note also that this network presents clear hierarchical structure characteristic of biological networks [56, 125], with many sink nodes [134] that are targets of transcription factors and a relatively large cycle cluster originating from the kinase interactome.

It is important to note that this is a non-specific network, whereas real gene regulatory networks can experience a sort of "rewiring" for a single cell type under various internal conditions [96]. In

Properties	Lung	B cell
Nodes	9073	4364
Edges	45635	55144
Sources	129	8
Sinks	8443	1418
Av. outdegree	5.03	12.64
Max outdegree	240	2372
Max indegree	68	196
Self-loops	238	0
Undirected edges	350	23386
Diameter	11	11
Max cycle cluster	401	2886
Av. clustering coeff. [50]	0.0544	0.2315

Table 5.1 General properties of the full networks. The network used for the analysis of lung cancer is a generic one obtained by combining the data sets in [100] and [161]. The B cell network is a curated version of the B cell interactome obtained from [89] using a network reconstruction method and gene expression data from B cells.

this analysis, the difference in topology between a normal and a cancer cell's regulatory network is assumed to be negligible. The methods described here can be applied to more specialized networks for specific cell types and cancer types as these networks become more widely available.[6]

The IMR-90 cell line [105, 108] was used for the normal attractor state, and the two cancer attractor states examined were from the A549 (adenocarcinoma) [70, 109, 129, 139, 154] and NCI-H358 (bronchioalveolar carcinoma) [154, 139] cell lines. Gene expression measurements from all referenced studies for a given cell line were averaged together to create a single attractor. The resulting magnetization curves for A549 and NCI-H358 are very similar, so the following analysis addresses only A549. The full network contains 9073 nodes, but only 1175 of them are differential nodes in the IMR-90/A549 model. In the unconstrained $p = 1$ case, all 1175 differential nodes are candidates for targeting. Exhaustively searching for the best pair of nodes to control requires investigating 689725 combinations simulated on the full network of 9073 nodes. However, 1094 of the 1175 nodes are sinks (i.e. nodes i with outdegree $\deg^+(i) = 0$, ignoring self loops) and therefore have $I(i) = e_{\text{opt}}(i) = 1$, which can be safely ignored. The search space is thus reduced to

[6]The original article containing these results was published before Ong's method was developed.

81 nodes, and finding even the best triplet of nodes exhaustively is possible. Including constraints, only 31 nodes are differential kinases with $\xi_i^c = +1$. This reduces the search space at the cost of increasing the minimum achievable m_∞^c.

There is one important cycle cluster in the full network, and it is composed of 401 nodes. This cycle cluster has an impact of 7948 for $p = 1$, giving a critical efficiency of at least ~ 19.8, and $1 \leq n_{\text{crit}} \leq 401$ by Eq. 5.30. The optimal efficiency for this cycle cluster is $e_{\text{opt}} = 29$, but this is achieved for fixing the first bottleneck in the cluster. Additionally, this node is the highest impact size 1 bottleneck in the full network, and so the mixed efficiency-ranked results are identical to the pure efficiency-ranked results for the unconstrained $p = 1$ lung network. The mixed efficiency-ranked strategy was thus ignored in this case.

Fig. 5.13 shows the results for the unconstrained $p = 1$ model of the IMR-90/A549 lung cell network. The unconstrained $p = 1$ system has the largest search space, so the Monte Carlo strategy performs poorly. The best+1 strategy is the most effective strategy for controlling this network. The seed set of nodes used here was simply the size 1 bottleneck with the largest impact. The best+1 method works better than effeciency-ranked method because best+1 includes the synergistic effects of fixing multiple nodes, while efficiency-ranked assumes that there is no overlap between the set of nodes downstream from multiple bottlenecks. Importantly, however, the efficiency-ranked method works nearly as well as best+1 and much better than Monte Carlo, both of which are more computationally expensive than the efficiency-ranked strategy.

Fig. 5.14 shows the results for the unconstrained $p = 2$ model of the IMR-90/A549 lung cell network. The search space for $p = 2$ is much smaller than that for $p = 1$. The largest weakly connected differential subnetwork contains only 506 nodes (see Table 5.2) , and the remaining differential nodes are islets or are in subnetworks composed of two nodes and are therefore unnecessary to consider. Of these 506 nodes, 450 are sinks. Fig. 5.15 shows the largest weakly connected component of the differential subnetwork, and the top five bottlenecks in the unconstrained case are shown in red. If limiting the search to differential kinases with $\xi_i^c = +1$ and ignoring all sinks, $p = 2$ has 19 possible targets. There is only one cycle cluster in the largest differential subnetwork,

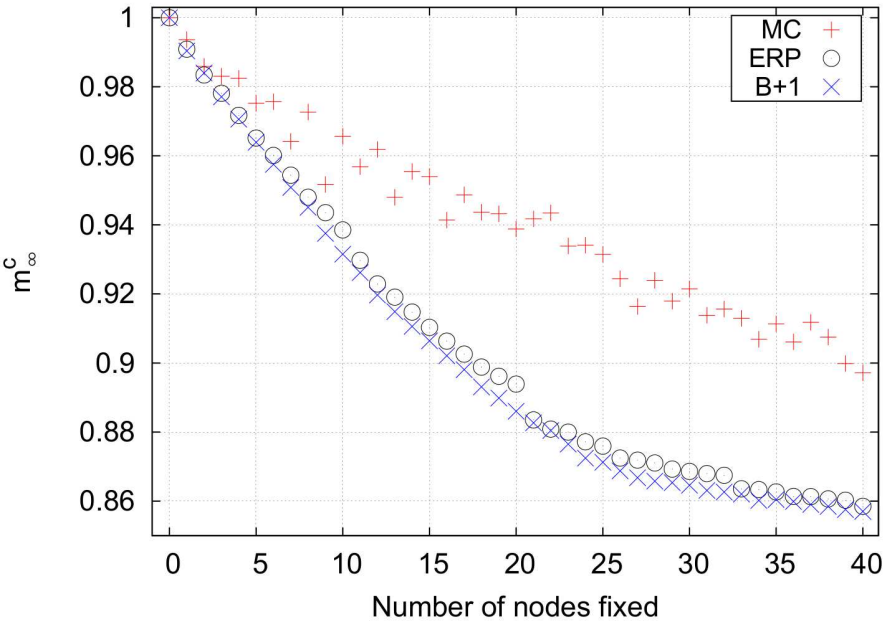


Figure 5.13 Final cancer magnetizations for an unconstrained search on the lung cell network using $p = 1$. The efficiency-ranked strategy outperforms the relatively expensive Monte Carlo strategy. The best+1 strategy works best, although it requires the largest computational time. Note that the mixed efficiency-ranked curve is not shown because it is identical to the pure efficiency-ranked curve. Key for magnetization curves: MC = Monte Carlo, B+1 = best+1, ERP = pure efficiency-ranked.

containing 6 nodes. Like the $p = 1$ case, the optimal efficiency occurs when targeting the first node, which is the highest impact size 1 bottleneck. Because the mixed efficiency-ranked strategy gives the same results as the pure efficiency-ranked strategy, only the pure strategy was examined. The Monte Carlo strategy fares better in the unconstrained $p = 2$ case because the search space is smaller. Additionally, the efficiency-ranked strategy does worse against the best+1 strategy for $p = 2$ than it did for $p = 1$. This is because the effective edge deletion decreases the average indegree of the network and makes nodes easier to control indirectly. When many upstream bottlenecks are controlled, some of the downstream bottlenecks in the efficiency-ranked list can be indirectly controlled. Thus, controlling these nodes directly results in no change in the magnetization. This gives the plateaus shown for fixing nodes 9-10 and 12-15, for example.

The only case in which an exhaustive search is possible is for $p = 2$ with constraints, which is

	Lung		B					
Properties	I/A	I/H	N/D	N/F	N/L	M/D	M/F	M/L
Nodes	506	667	684	511	841	621	457	742
Edges	846	1227	2855	1717	3962	2525	1501	3401
Sources and eff. sources	30	34	12	11	9	9	9	12
Sinks and eff. sinks	450	598	286	198	369	275	204	333
Av. outdegree	1.67	1.84	4.17	3.36	4.71	4.07	3.28	4.58
Max outdegree	52	51	155	143	336	138	132	292
Max indegree	8	10	40	29	49	35	27	44
Self-loops	27	31	0	0	0	0	0	0
Undirected edges	0	4	1238	738	1468	1000	596	1214
Diameter	9	9	12	15	12	13	14	12
Max cycle cluster size	6	3	351	280	397	305	199	337
Av. clustering coeff	0.035	0.042	0.188	0.197	0.245	0.175	0.194	0.239

Table 5.2 Properties of the largest weakly connected differential subnetworks for all cell types. I = IMR-90 (normal), A = A549 (cancer), H = NCI-H358 (cancer), N = Naïve (normal), M = Memory (normal), D = DLBCL (cancer), F = Follicular lymphoma (cancer), L = EBV-immortalized lymphoblastoma (cancer).

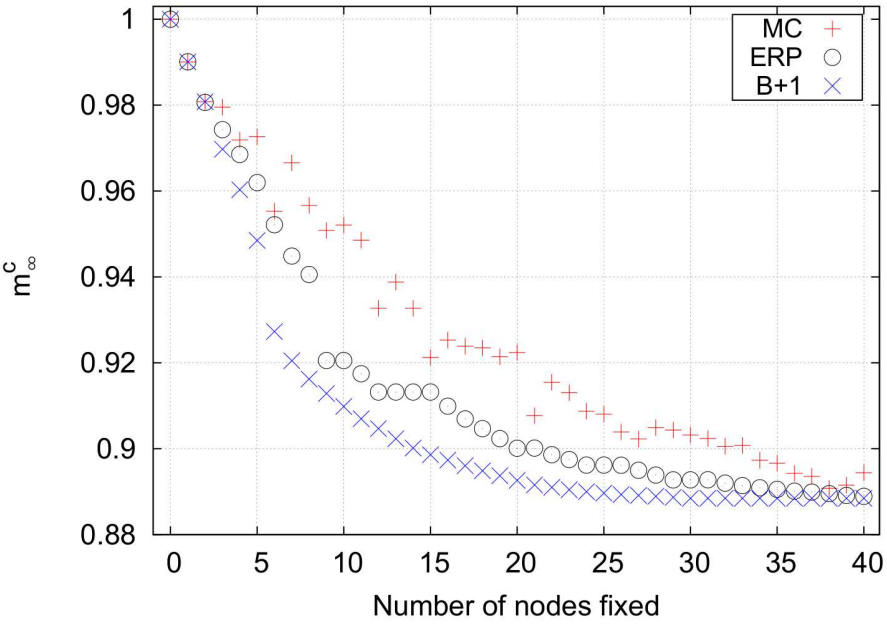


Figure 5.14 Final cancer magnetizations for an unconstrained search on the lung cell network using $p = 2$. As in the $p = 1$ case, the efficiency-ranked strategy outperforms the expensive Monte Carlo search. The plateaus in the efficiency-ranked strategy when fixing 9-10, 12-15, 20-21, etc. nodes are a result of targeting bottlenecks that are already indirectly controlled.

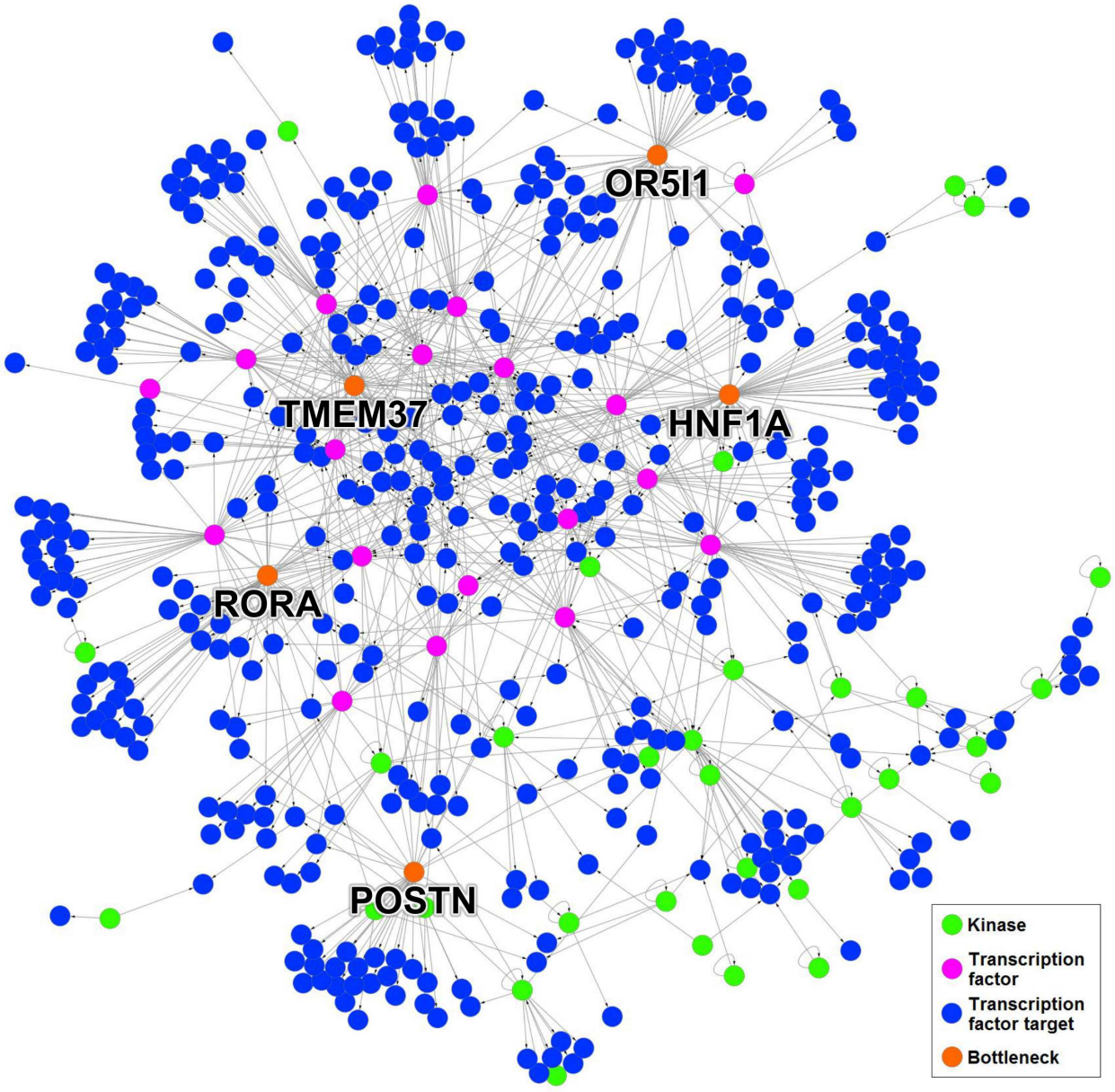


Figure 5.15 Largest weakly connected differential subnetwork for IMR-90/A549 and $p = 2$. Out of the 506 pictured nodes, 450 are sinks and therefore have an impact equal to one. The top five bottlenecks are labeled with their gene names and colored orange.

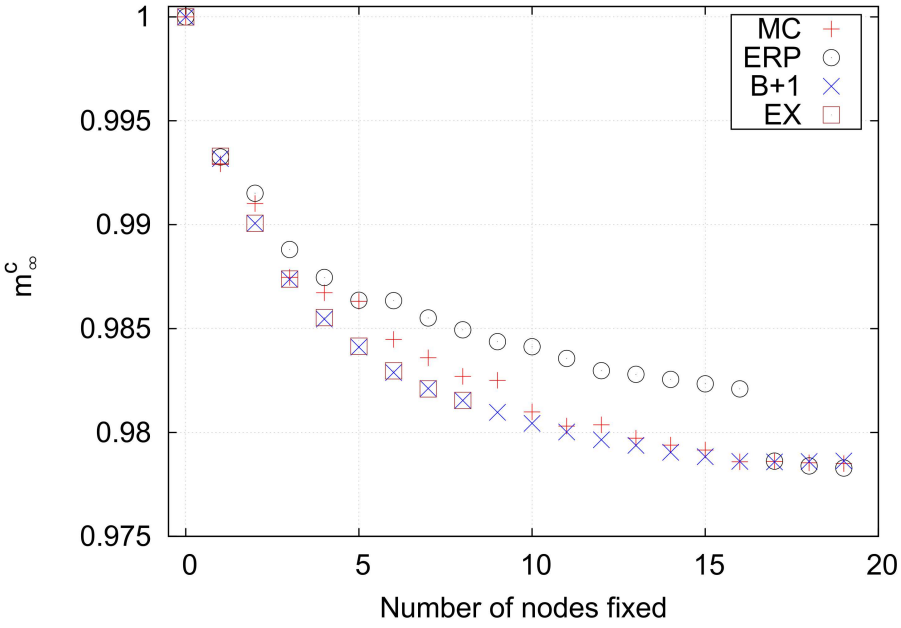


Figure 5.16 Final cancer magnetizations for a constrained search on the lung cell network using $p = 2$. This is the only case in which a limited exhaustive search is possible. Interestingly, the exhaustive search locates the same nodes as the best+1 strategy for fixing up to eight nodes. The efficiency-ranked strategy performs poorly compared to the Monte Carlo strategy because the search space is small and a large portion of the available space is sampled by the Monte Carlo search.

shown in Fig. 5.16. Note that the polynomial-time best+1 strategy identifies the same set of nodes as the exponential-time exhaustive search. This is not surprising, however, since the constraints limit the available search space. This means that the Monte Carlo also does well. The efficiency-ranked method performs worst. The efficiency-ranked strategy is designed to be a heuristic strategy that scales gently, however, and is not expected to work well in such a small space when compared with more computationally expensive methods.

5.3.3.2 B Cell Network

The B cell network was derived from the B cell interactome of Ref. [89]. The reconstruction method used in Ref. [89] removes edges from an initially complete network depending on pair-wise gene expression correlation. Additionally, the original B cell network contains many protein-

protein interactions (PPIs) as well as transcription factor-gene interactions (TFGIs). TFGIs have definite directionality: a transcription factor encoded by one gene affects the expression level of its target gene(s). PPIs, however, do not have obvious directionality. These PPIs were filtered by checking if the genes encoding these proteins interacted according to the PhosphoPOINT/TRANSFAC network of the previous section, and if so, kept the edge as directed.

Because of the network construction algorithm and the inclusion of many undirected edges, the B cell network is more dense ($\sim$0.290% complete, see Table 5.1) than the lung cell network. This higher density leads to many more cycles than the lung cell network, and many of these cycles overlap to form one very large cycle cluster containing about two thirds of all nodes in the full network. All gene expression data used for B cell attractors was taken from Ref. [29]. Two types of normal B cells (naïve and memory) and three types of B cell cancers (diffuse large B-cell lymphoma (DLBCL), follicular lymphoma, and EBV-immortalized lymphoblastoma) were analyzed, giving six combinations in total. Results are presented for only the naïve/DLBCL combination below, but Tables 5.2 and 5.3 list the properties of all normal/cancer combinations. Again, all gene expression measurements for a given cell type were averaged together to produce a single attractor. The full B cell network is composed of 4364 nodes. For $p = 1$, there is one cycle cluster C composed of 2886 nodes. This cycle cluster has $1 \leq n_{\text{crit}}(C) \leq 1460$, $I(C) = 4353$, and $3.0 \leq e_{\text{crit}}(C) \leq 4353$. Finding $Z(C)$ was deemed too difficult.

Fig. 5.17 shows the results for the unconstrained $p = 1$ case. Again, the pure efficiency-ranked strategy gave the same results as the mixed efficiency-ranked strategy, so only the pure strategy was analyzed. As shown in Fig. 5.17, the Monte Carlo strategy is out-performed by both the efficiency-ranked and best+1 strategies. The synergistic effects of fixing multiple bottlenecks slowly becomes apparent as the best+1 and efficiency-ranked curves separate.

Fig. 5.18 shows the results for the unconstrained $p = 2$ case. The largest weakly connected subnetwork contains one cycle cluster with 351 nodes, with $1 \leq n_{\text{crit}} \leq 208$. Although finding a set of critical nodes is difficult, the optimal efficiency for this cycle cluster is 62.2 for fixing 10 bottlenecks in the cycle cluster. This makes targeting the cycle cluster worthwhile. The efficiency

	I/A				I/H			
	$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I
UNC	HNF1A	29	OR5I1	35	HNF1A	29	HMX1	41
	TMEM37	22	TMEM37	25	MAP3K3	18	PBX1	38
	OR5I1	20	HNF1A	23	TP53	18	MYB	25
	MAP3K14	19	POSTN	21	RUNX1	17	ITGB2	20
	MAP3K3	18	RORA	18	RORA	16	TNFRSF10A	18
CON	MAP3K14	19	SRC	15	TTN	16	BMPR1B	18
	SRC	14	BMPR1B	7	RIPK3	6	LCK	8

	N/D				N/F				N/L			
	$p=1$		$p=2$		$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I
UNC	BCL6	12	NFIC	22	BCL6	12	NCOA1	20	RBL2	11	RBL2	22
	MEF2A	5	TGIF1	19	MEF2A	5	NFATC3	15	FOXM1	8	ATF2	12
	NCOA1	5	BCL6	14	NCOA1	5	BCL6	11	ATF2	7	NFATC3	11
	TGIF1	4	FOXJ2	12	TGIF1	4	CEBPD	8	RXRA	5	RXRA	9
	NFATC3	4	NFATC3	12	NFATC3	4	RELA	8	NFATC3	4	PATZ1	8
CON	BUB1B	2	CSNK2A2	2	BUB1B	2	WEE1	2	BUB1B	2	PRKCD	2
	AAK1	1	AKT1	2	AAK1	1	CSNK2A2	2	AAK1	1	AURKB	2

	M/D				M/F				M/L			
	$p=1$		$p=2$		$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I
UNC	BCL6	12	FOXJ2	12	BCL6	12	NCOA1	18	RBL2	11	RBL2	16
	MEF2A	5	NFIC	12	MEF2A	5	BCL6	13	FOXM1	8	ATF2	10
	NCOA1	5	BCL6	11	NCOA1	5	E2F3	9	ATF2	7	ZNF91	8
	NFATC3	4	NCOA1	9	NFATC3	4	RUNX1	9	RXRA	5	STAT6	8
	SMAD4	4	MEF2A	8	RELA	4	TFE3	7	TGIF1	4	FOXM1	8
CON	AAK1	1	RIPK2	1	AAK1	1	ROCK2	2	AAK1	1	AURKB	2
	RIPK2	1	MAST2	1	RIPK2	1	RIPK2	1	SCYL3	1	RIPK2	1

Table 5.3 Best single genes and their impacts for the p=1 and p=2 models. The unconstrained (UNC) and constrained (CON) case are shown. The constrained case refer to target that are kinases and are expressed in the cancer case. I = IMR-90 (normal), A = A549 (cancer), H = NCI-H358 (cancer), N = Naïve (normal), M = Memory (normal), D = DLBCL (cancer), F = Follicular lymphoma (cancer), L = EBV-immortalized lymphoblastoma (cancer).

of this set of 10 nodes is larger than the efficiencies of the first 10 nodes from the pure efficiency-ranked strategy, so the m_∞^c from the mixed strategy drops earlier than the pure strategy. Both strategies quickly identify a small set of nodes capable of controlling a significant portion of the differential network, however, and the same result is obtained for fixing more than 10 nodes. The

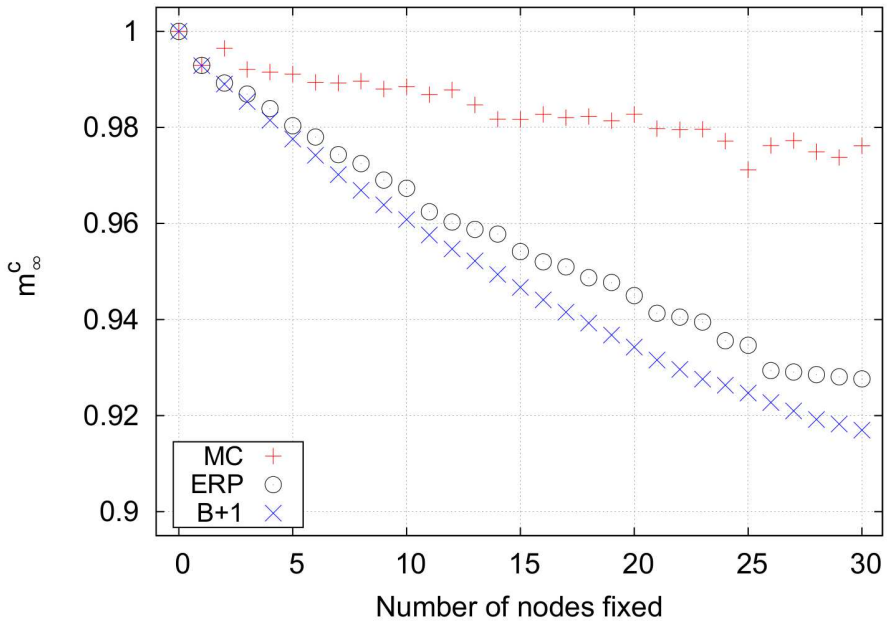


Figure 5.17 Final cancer magnetizations for an unconstrained search on the B cell network using $p = 1$. The Monte Carlo strategy is ineffective for fixing any number of nodes. The efficiency-ranked and best+1 curves slowly separate because synergistic effects accumulate faster for best+1.

best+1 strategy finds a smaller set of nodes that controls a similar fraction of the cycle cluster, and fixing more than 7 nodes results in only incremental decreases in m^c_∞. The Monte Carlo strategy performs poorly, never finding a set of nodes adequate to control a significant fraction of the nodes in the cycle cluster.

5.4 Conclusion

Signaling models for large and complex biological networks are becoming important tools for designing new therapeutic methods for complex diseases such as cancer. Even if knowledge of biological networks is incomplete, rapid progress is currently being made using reconstruction methods that use large amounts of publicly available omic data [36, 64]. The Hopfield model allows mapping of gene expression patterns of normal and cancer cells into stored attractor states of the signaling dynamics in directed networks. The role of each node in disrupting the network

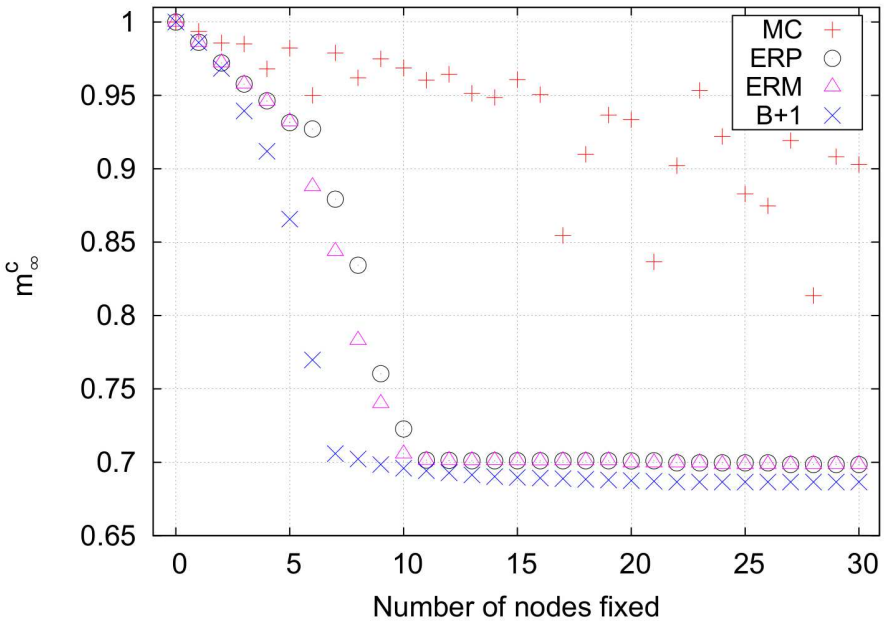


Figure 5.18 Final cancer magnetizations for an unconstrained search on the B cell network using $p = 2$. The rather sudden drop in the magnetization between controlling 5 and 10 nodes in the efficiency-ranked strategies comes from flipping a significant portion of a cycle cluster. This is the only network examined in which the mixed efficiency-ranked strategy produces results different from the pure efficiency-ranked strategy.

signaling can therefore be explicitly analyzed to identify isolated genes or sets of strongly connected genes that are selective in their action. The concept of *size k bottlnecks* to identify such genes led to the formulation of two heuristic strategies, *efficiency-ranked* and *best+1*, to find nodes that efficiently reduce the cancer state's magnetization. Using this approach, small sets of nodes in lung and B cancer cells were located which, when forced away from their initial states with local magnetic fields representing targeted drugs, disrupt the signaling of the cancer cells while leaving normal cells in their original state. For networks with few targetable nodes, exhaustive searches or Monte Carlo searches can locate effective sets of nodes. For larger networks, however, these strategies become too cumbersome and the discussed heuristic strategies represent a feasible alternative. For tree-like networks, the pure efficiency-ranked strategy works well, whereas the mixed efficiency-ranked strategy could be a better choice for networks with high-impact cycle clusters.

Two important assumptions are made in applying this analysis to real biological systems. First,

it is assumed that genes are either fully off or fully on, with no intermediate state. Modelling the state of a neuron as "all-or-none" has long been accepted as a reasonable assumption [102], which inspired the spin glass framework for the Hopfield model. While similar switch-like behavior in gene regulatory networks has been proposed as an explanation of, for example, segmentation in *Drosophila* embryos [81], assigning a Boolean value to gene expression may be overly simplistic in many cases. A model which uses spins with more than two projections could prove to be more realistic and predictive. Second, all nodes are assumed to update their status with a single timescale and with a single interaction strength. If the signaling timescale τ_{ij} (i.e. how long it takes for a gene j to produce a protein that regulates gene i) for each edge in the biological network were known, simulations could be conducted in which a signal traveling along an edge (j,i) reaches its target after τ_{ij} time steps. This would amount to a "queue" of signals moving between nodes.

Despite these issues, the model shows promise. Some of the genes identified in Table 5.3 are consistent with current clinical and cancer biology knowledge. For instance, one of the effective targets in the lung cancer list is the well known tumor suppressor gene TP53 [12] that is frequently mutated in many cancer types including lung cancer [143]. Mutations in PBX1 have recently been detected in non-small-cell lung cancer and this gene is now being considered as a target for therapy and prognosis [103]. MAP3K3 and MAP3K14 are in the MAPK/ERK pathway which is a target of many novel therapeutic agents [104], and SRC is a well known oncogene and a candidate target in lung cancer [128]. BCL6 (B-cell lymphoma 6) is the most common oncogene in DLBCL, and it is known that its expression can predict prognosis and response to drug therapy [62, 127, 159]. BCL6 is also frequently mutated in follicular lymphoma [3, 43]. This analysis identified BCL6 as an important drug target for both DLBCL and follicular lymphoma using either naive or memory B-cells as a control for both $p = 1$ and $p = 2$. RBL2 disregulation has been recently associated with many types of lymphoma [35, 123, 156]. FOXM1 is a potential therapeutic target in mature B cell tumors [147] and ATF2 has been recently found to be highly disregulated in lymphoma [151, 155]. Besides BCL6 discussed above, the N/D list for DLBCL contains genes (MEF2A [11], NCOA1 [49, 163], TGIF1 [17, 60, 92], and NFATC3 [57]) that are all known to have a functional

role in cancer, even if they have not been associated to the specific B-cell cancer types considered here. All of these predictions were made from data from cell lines commonly used in *in vitro* testing in many laboratories. RNAi and targeted drugs could be used in these cell lines against the top scoring genes in Table 5.3 to test the disruption of survival or proliferative capacity. If experimentally validated, this analysis based on attractor states and bottlenecks could be applied to patient-derived cancer cells by integrating in the model patient gene expression data to identify patient-specific targets.

The above unconstrained searches assume that there exists some set of "miracle drugs" which can turn any gene on and off at will. This limitation can be partially taken into account by using constrained searches that restrict the set of nodes that can be addressed. However, even the constrained search results are unrealistic, since most drugs directly target more than one gene. Inhibitors, for example, could target differential nodes with $\xi_i^c = -1$ and $\xi_i^n = +1$, which would damage only normal cells. Additionally, drugs would not be restricted to target only differential nodes, and certain combinations could be toxic to both normal and cancer cells. Few cancer treatments involve the use of a single drug, and the synergistic effects of combining multiple drugs adds yet another level of complication to finding an effective treatment [52]. On the other hand, the intrinsic nonlinearity of a cellular signaling network, with its inherent structure of attractor states, enhances control [30] so that a properly selected set of druggable targets might be sufficient for robust control.

Symbol	Explanation
G	Set of nodes and directed edges (network)
N	Number of nodes
A_{ij}	Adjacency matrix
$V(G)$	Set of nodes in G
$E(G)$	Set of edges in G
$\deg^{+/-}(i)$	Outdegree/indegree of node i
σ_i	Spin of node i, $= \pm 1$
ξ^a	a^{th} attractor
$\xi^{n/c}$	Normal/cancer attractor
J_{ij}	Coupling matrix
h_i	Total field at node i
h_i^{ext}	External field applied to node i
T	Temperature
Q	Set of source and effective source nodes
$m^a(t)$	Magnetization along attractor a at time t
m_∞^a	Steady-state magnetization along attractor a
p	Number of attractors in coupling matrix
S	Set of similarity nodes
D	Set of differential nodes
$L(B)$	Control set of bottleneck B
$I(B)$	Impact of bottleneck B
C	Cycle cluster
B	Size k bottleneck, where $k = \lvert B \rvert$
$Z(B,G)$	Set of critical nodes for bottleneck B in network G
$n_{\text{crit}}(B,G)$	Critical number of nodes in bottleneck B in network G
$R(C,G)$	Set of externally influenced nodes
$W(C,G)$	Set of intruder connections
$Z_{\text{red}}(C,G)$	Reduced set of critical nodes
μ	Minimum indegree of all nodes in a cycle cluster
$e_{\text{crit}}(B)$	Critical efficiency of bottleneck B
$e_{\text{opt}}(B)$	Optimal efficiency of bottleneck B

Table 5.4 Reference table of all symbols used in this chapter.

CHAPTER 6

GENERALIZED HOPFIELD MODELS

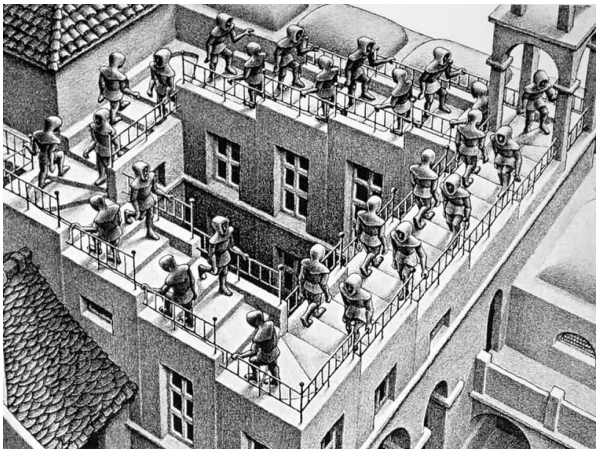


Figure 6.1 *Ascending and Descending* by MC Escher.

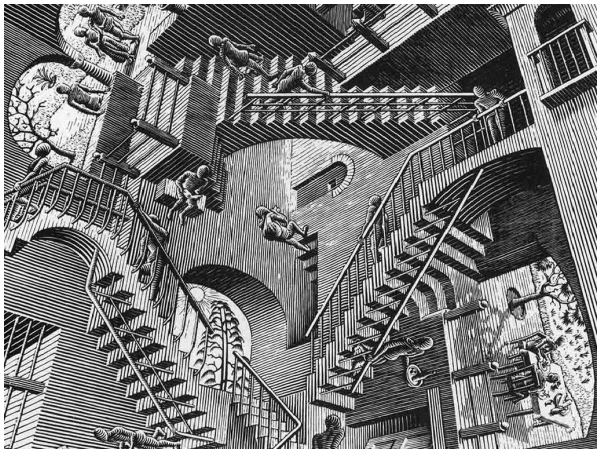


Figure 6.2 *Relativity* by MC Escher.

The following chapter contains unpublished but exciting new directions for the Hopfield model in general and for its application to gene regulation in particular, and is organized as follows.

- Section 6.1, *Stability in Hopfield GRNs*: a discussion of the starkly different behavior of the magnetization of modular Erdős-Rényi networks and real gene regulatory networks as a function of the level of noise in the Hopfield model
- Section 6.2, *Programmable nonequilibrium Hopfield systems*: a discussion of one known extension of the Hopfield model which encodes cyclic attractors, and one novel extension which enables more complex shapes to be programmed into state space
- Section 6.3, *Cell cycle and the Hopfield model*: an application of the cyclic Hopfield model to cell cycle

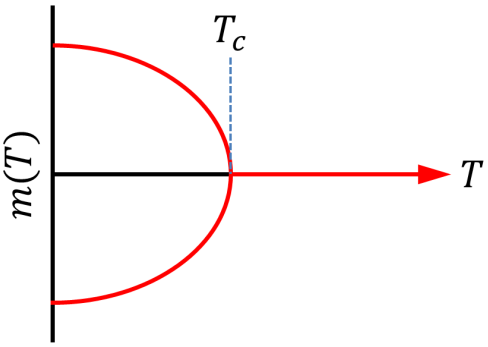


Figure 6.3 Schematic of the magnetization phase transition for Hopfield systems on Erdős-Rényi networks. Note that the symmetric (mirror) solution is also shown below the T axis.

6.1 Stability in Hopfield GRNs

The Hopfield model is designed to balance stability and flexibility, which is an indispensable attribute of all biological systems: a cell temporarily driven away from its native gene expression state by temporary or weak external perturbations or noise (e.g. fluctuations in the environmental temperature) is able to self-correct and return to its native state, but is also capable of switching its state in response to particular sets of signals and stresses (e.g. an increased concentration of growth hormones or other chemical signals in the cell's environment may lead to upregulation of proliferation). Furthermore, the modular structure of gene regulatory networks means that individual communities of genes may respond differently to the same stresses. This means that perturbations to genes in a particular community may transmit only weakly or not at all to neighboring communities due to the smaller number of connections between them (by the definition of maximum modularity). Understanding how community structures affect signaling is thus important for designing control techniques for GRNs.

The mean field solution for Hopfield systems on asymmetric Erdős-Rényi networks [40] is rederived in Appendix F. Because in the limit $N \to \infty$ the vast majority of nodes in Erdős-Rényi networks are average nodes (they all have nearly identical degree and are wired together in a uniformly random way), there is a sharp phase transition in the magnetization as a function of temperature, $m(T)$, at a critical temperature T_c, as shown schematically in Fig. 6.3.

In the same appendix, I generalize this result to a class of model networks called *stochastic block model networks* (SBMNs). Each of the Ω communities in a SBMN is its own Erdős-Rényi network, and these communities are also randomly wired together. A SBMN is defined by a *connectivity matrix*, $C^{\Omega\times\Omega}$, where C_{IJ} is the average number of connections per node from nodes in community J to nodes in community I (and so C_{II} is the average number of connections per node within community I). This allows for different levels of stability within communities and asymmetric couplings between communities, meaning that the state of J may influence the state of I more strongly than the reverse case.

Fig. 6.4 shows the mean field solution (dashed lines) and a numerical solution (circles with error bars) for a SBMN constructed using the connectivity matrix from AML 2.3 (using level-1 communities only) and random attractor states, with $T_c \approx 20$. There is good agreement between the analytical and numerical solutions, and this agreement improves when larger networks are used. However, simulating the dynamics on the real AML 2.3 network across the same range of temperatures reveals very different behavior, as shown in Fig. 6.5 (which includes the dashed mean field solution for reference). While the zero temperature system has lower magnetization than the SBMN case, it maintains finite magnetization even for $T \gg T_c$, showing no sign of a phase transition. This is perhaps because AML 2.3 has a power law rather than a binomial degree distribution, as verified by simulations using *modular power law networks*, which show qualitatively similar $m(T)$ curves. The nested modular structure of AML 2.3 may also play a part in the long tailed behavior of $m(T)$, as the lack of a phase transition could be caused by a level-by-level breakdown of order as the temperature is increased. This is an open question that merits further investigation.

6.2 Programmable nonequilibrium Hopfield systems

Note: in the previous chapter, the patterns $\left\{\xi_i^\mu\right\}$ were indexed according to $\mu = 1,2,\ldots,p$. For mathematical convenience, this chapter simply relabels these indices to $\mu = 0,1,\ldots,p-1$, and the limits on the summations have been disregarded for simplicity. Additionally, the previous chapter used an asynchronous update scheme, in which a node was randomly selected to be updated at

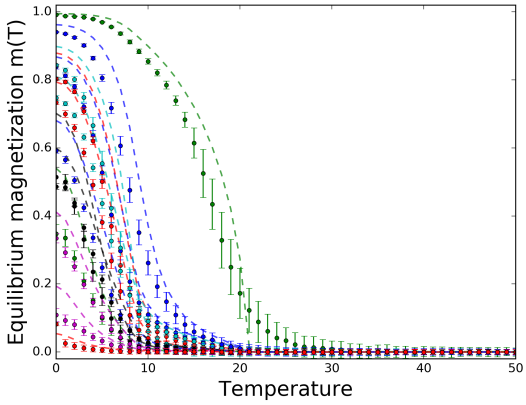


Figure 6.4 Mean field (dashed) and explicit simulations (circles with error bars) of the magnetization of communities in SBMNs over a range of temperatures. There is a phase change in the mean field solution at $T_c \approx 20$, and the explicit simulation is in good agreement.

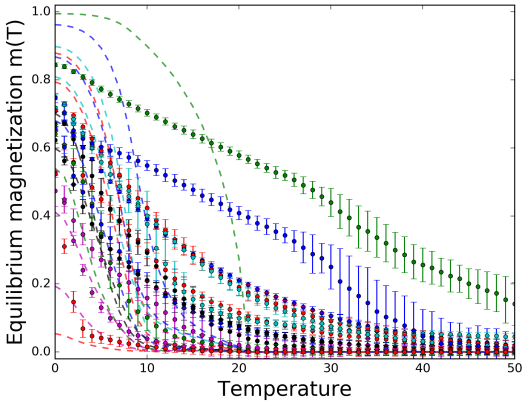


Figure 6.5 Mean field SBMN (dashed) and explicit AML 2.3 simulations (circles with error bars) of the magnetization of communities over a range of temperatures. Unlike the explicit simulations from Fig. 6.4, the real network shows no apparent phase change, instead maintaining finite magnetization over all temperatures examined.

each time step. The following models require using the synchronous scheme, in which the state of all nodes is simultaneously updated at each time step. The background and original publications for this section can be found in [67].

The original definition of the coupling matrix,

$$J_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^{\mu}\xi_j^{\mu} \tag{6.1}$$

encodes each of the p patterns as point attractors for $N, p \to \infty$, $p/N \equiv \alpha < \alpha_c \approx 0.138$, and $\xi_i^{\mu} = \pm 1$ with equal probability. As explained in Appendix G, a more stable form of the coupling matrix which accomodates correlated patterns may be constructed according to

$$J_{ij} = \frac{1}{N}\sum_{\mu\nu} \xi_i^{\mu}\left(Q^{-1}\right)_{\mu\nu}\xi_j^{\nu} \tag{6.2}$$

where

$$Q_{\mu\nu} = \frac{1}{N}\sum_{i} \xi_i^{\mu}\xi_i^{\nu} \tag{6.3}$$

A slight alteration to Eq. 6.2 enables cyclic attractors to be encoded as well. A simple cyclic attractor of period p can be created by

$$\widetilde{J}_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^{\mathrm{mod}p(\mu+1)} \xi_j^{\mu} \tag{6.4}$$

or, as futher explained in Appendix G, the matrix $Q_{\mu\nu}$ can be included to orthogonalize the patterns in the cyclic case by using the form

$$\widetilde{J}_{ij} = \frac{1}{N}\sum_{\mu\nu} \xi_i^{\mathrm{mod}p(\mu+1)} \left(Q^{-1}\right)_{\mu\nu} \xi_j^{\nu} \tag{6.5}$$

At zero temperature, this coupling matrix cyclically maps through the p patterns (or the mirror cycle, depending on the initial configuration); in other words, if $\sigma_i(t) = \pm\xi_i^{\mu}$, then $\sigma_i(t+\Delta t) = \pm\xi_i^{\mathrm{mod}p(\mu+\Delta t)}$ for all $\Delta t = 0, 1, 2, \ldots$. Cyclic attractors may be envisioned as a ball falling down a flight of Penrose stairs, such as MC Escher's *Ascending and Descending* shown in Fig. 6.1.

A delay may be introduced by combining the point attractor coupling matrix and the cyclic attractor coupling matrix into a single coupling matrix,

$$J'_{ij} = J_{ij} + \lambda\widetilde{J}_{ij} \tag{6.6}$$

for an adjustable *transition strength* parameter $\lambda \geq 0$. If $\sigma(t) = \xi_i^{\mu}$, $\lambda \ll 1$, and $T = 0$, the point attractor term dominates and $\sigma(t) = \sigma(t+1) = \sigma(t+2) = \ldots$. If $T > 0$, however, stochastic fluctuations will eventually begin to push the configuration out of the basin of attraction of the μ^{th} attractor and initiate the cascade of spin flips that causes the configuration to transition to $\xi_i^{\mathrm{mod}p(\mu+1)}$. The system will remain there until another stochastic fluctuation pushes it into the next basin of attraction, and so on. One of the possible applications of this delayed cycle model will be covered in Section 6.3.

The point and cyclic attractor coupling matrices may be rewritten as

$$J_{ij} = \sum_{\mu} \xi_i^{\mu} \left[\frac{1}{N}\sum_{\nu} \left(Q^{-1}\right)_{\mu\nu} \xi_j^{\nu}\right] \tag{6.7}$$

and

$$\widetilde{J}_{ij} = \sum_{\mu} \lambda\, \xi_i^{\mathrm{mod}p(\mu+1)} \left[\frac{1}{N}\sum_{\nu} \left(Q^{-1}\right)_{\mu\nu} \xi_j^{\nu}\right] \tag{6.8}$$

I define the common bracketed term

$$S_j^\mu = \frac{1}{N}\sum_\nu \left(Q^{-1}\right)_{\mu\nu} \xi_j^\nu \tag{6.9}$$

to be the *source term*, and the ξ_i^μ and $\lambda\xi_i^{\mathrm{mod}_p(\mu+1)}$ terms to be particular examples of *target terms*. A more general way to write J'_{ij} is

$$J'_{ij} = \sum_{\mu\omega} \xi_i^\omega M_{\omega\mu} S_j^\mu \tag{6.10}$$

for some $p \times p$ *mapping matrix* M. The point attractor coupling matrix is constructed by setting $M_{\mu\omega} = \delta_{\mu\omega}$ (the identity matrix). For $p = 3$, the cycle $\xi_i^0 \to \xi_i^1 \to \xi_i^2 \to \xi_i^0$ is constructed by

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{6.11}$$

and for a delayed cyclic attractor, M takes the form

$$M = \begin{pmatrix} 1 & 0 & \lambda \\ \lambda & 1 & 0 \\ 0 & \lambda & 1 \end{pmatrix} \tag{6.12}$$

M may be used to store transition maps which are more complex than simple point and cyclic attractors. For example, a system with $p = 3$ can encode the transitions $\xi_i^0 \to \xi_i^2$, $\xi_i^1 \to \xi_i^2$, and $\xi_i^2 \to \xi_i^2$ using

$$M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \tag{6.13}$$

This ensures that all states map to ξ_i^2. M can also be designed to store multiple point and cyclic attractors in the same system. In fact, any target term may be used to generate complex transition maps in state space, as long as each state transitions to exactly one other state (in addition to an optional self-loop that delays transitions). This enables the encoding of maps such as *deterministic random maps* [39], an example of which is shown in Fig. 6.6. Furthermore, different transition

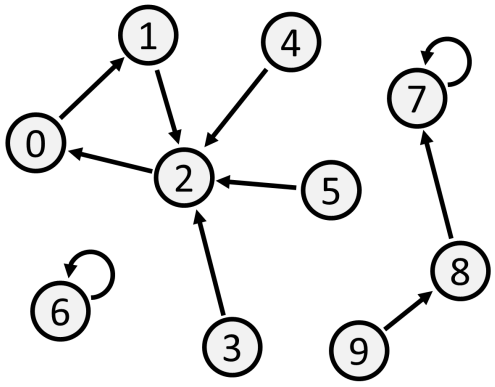


Figure 6.6 An example deterministic random map that can be stored in a Hopfield coupling matrix using an appropriate mapping matrix M. Each node $\mu = 0, 1, \ldots, 9$ represents a stored pattern $\vec{\xi}^{\mu}$ and each edge represents a transition.

strengths can be set to values other than 0 and 1, meaning that transition rates can be tuned as well. Similar to cyclic attractors, this complex landscape of attractors may be envisioned as a ball falling down a labyrinth of stairs such as in MC Escher's *Relativity* shown in Fig. 6.2, where doorways lead to closets (point attractors) or flights of Penrose stairs (cyclic attractors). The concept of mapping matrices could be used in the future to model gene expression changes in complex biological processes such as cell differentiation.

6.3 Cell cycle and the Hopfield model

One of the most important misregulated processes in cancer is *cell cycle* (CC), in which a cell replicates its DNA and divides into two daughter cells. Understanding CC in both normal and cancer cells is undoubtedly important for medical research, and the ability to slow CC in cancer cells without affecting CC in normal cells could be therapeutically beneficial. Below, I present the preliminary results of encoding *time series* gene expression data (gene expression measured at regular time intervals) for the process of CC in a delayed cyclic coupling matrix J'_{ij}, with the eventual goal of predicting effective perturbations which control CC.

Dominguez et al. [44] recently released time series bulk (i.e. many cells) RNA-seq data of HeLa cells (human cervical cancer). The CC processes in the many cells in the sample were

synchronized using a *double thymidine block*, freezing all cells in the G1 phase and then releasing the block. RNA-seq measurements were conducted on subsamples at regular intervals for the duration of two full cycles, producing a raw (continuous) expression matrix $D_{i\mu} \geq 0$.

I used the continuous matrix $D_{i\mu}$ to construct Boolean patterns as follows. For convenience, each row i of the raw matrix was linearly rescaled so that its minimum value was set to -1 and its maximum was set to $+1$. Because CC is a periodic process, only genes whose expressions were found to be sufficiently periodic were kept, and all non-periodic genes were discarded. The resulting data set had $N = 592$ genes and $p = 8$ gene expression patterns per CC period. The raw expression was then converted to Boolean form by assigning $\xi_i^\mu = +1$ (overexpressed) for samples in which $D_{i\mu} > 0$ and $\xi_i^\mu = -1$ (underexpressed) otherwise. These patterns were encoded cyclically in J'_{ij} according to Eq. 6.5.

In order to conduct simulations, the free parameters λ and T must be defined. $\lambda = 0.23$ and $T = 0.07$ were chosen here because it placed the system in a regime in which the fluctuations were great enough to cause transitions from one attractor to the next in a delayed, somewhat periodic manner. The system was initialized with $\sigma_i(t = 0) = \xi_i^0$ (representing a single isolated cell in the G1 phase) and was allowed to evolve. The resulting trajectory spanning nearly three periods is shown in Fig. 6.7. The bottom plot shows the overlap (magnetization) between the system's configuration and each of the $p = 8$ patterns as a function of time, where an overlap of $+1$ with attractor μ signifies perfect alignment. The top plot shows which pattern has maximum overlap at any point in time, and represents the cell's phenotype.

Unlike a purely cyclic system, the delay term in the coupling matrix causes the duration of each pattern to fluctuate from cycle to cycle. Real bulk gene expression measurements of CC show a decoherence of initially phase-matched cells because the length of each stage of CC for each cell fluctuates somewhat. This decoherence is shown for the randomly selected gene RNFT1 in Fig. 6.8. This effect can be observed using the Hopfield model by running parallel, independent simulations representing many cells and averaging the expression results. For example, Fig. 6.9 shows the resulting mean expression of RNFT1 across 50 isolated cells, each of which is governed

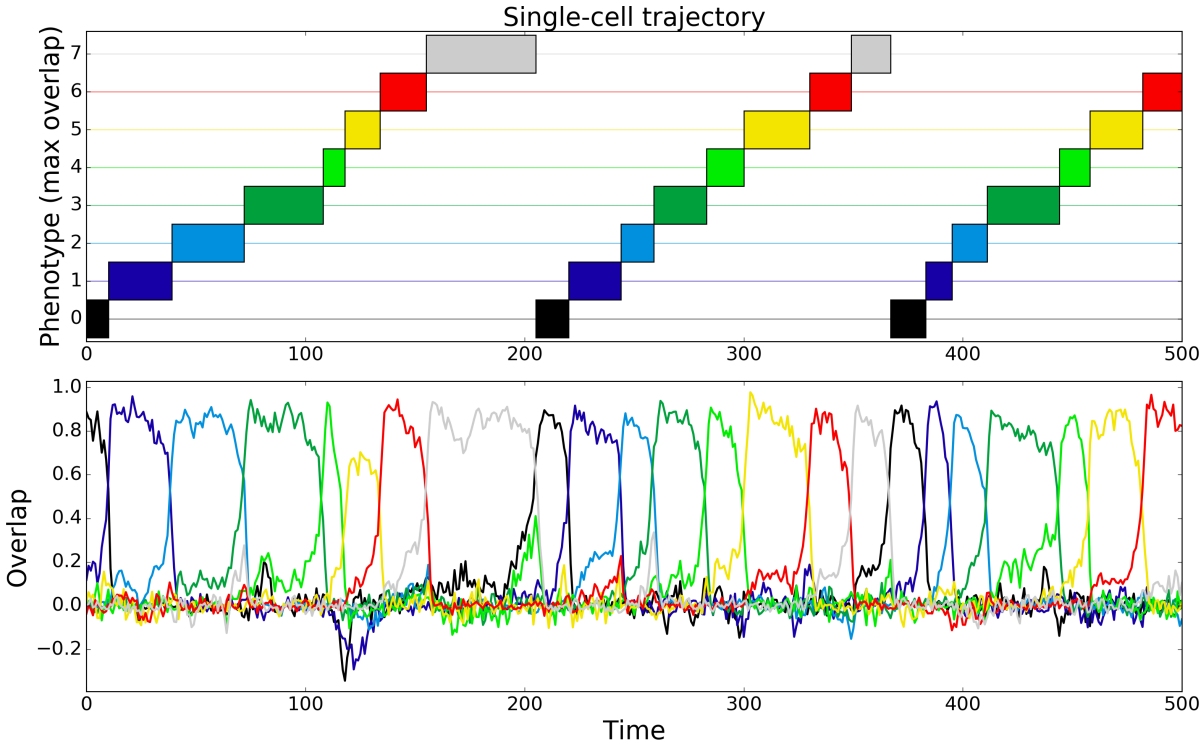


Figure 6.7 Simulated trajectory of a single HeLa cell. The bottom plot shows the overlap between the system's configuration and each of the $p = 8$ patterns as a function of time. The top plot shows which pattern has maximum overlap at any point in time, and represents the cell's phenotype.

by the same coupling matrix and parameters but subject to its own unique stochastic fluctuations. Although the expression of any given gene from any cell is either ± 1, the mean expression of a given gene across many cells reproduces the expected decaying sinusoid.

A heat map of J'_{ij} is shown in Fig. 6.10, separated into symmetric (point attractor) and anti-symmetric (transition) parts, and sorted into nested communities according to the algorithm from Section 2.8 applied to the symmetric part. Efficient control of the entire CC system (for example, arresting the system in the G1 phase) could begin by searching for control sets across different levels of the hierarchy of nested communities. If this method proves to be effective, it could significantly reduce the combinatoral complexity of searching for optimal or nearly optimal sets of genes to control.

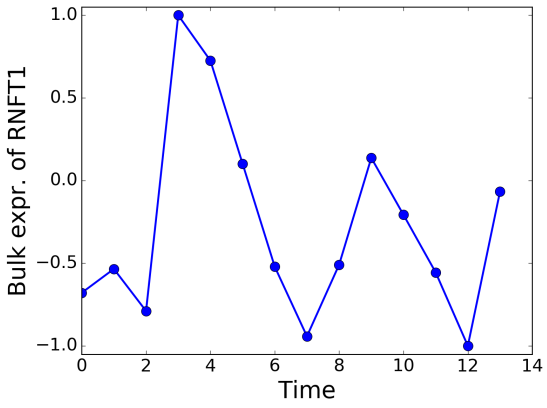


Figure 6.8 Experimental bulk mean expression for the gene RNFT1 as a function of time, taken from [44]. Although all cells were initially synchronized to the G1 phase, progression through cell cycle takes slightly different amounts of time for each cell and their gene expression profiles becomes less and less correlated, leading to a decaying sinusoid.

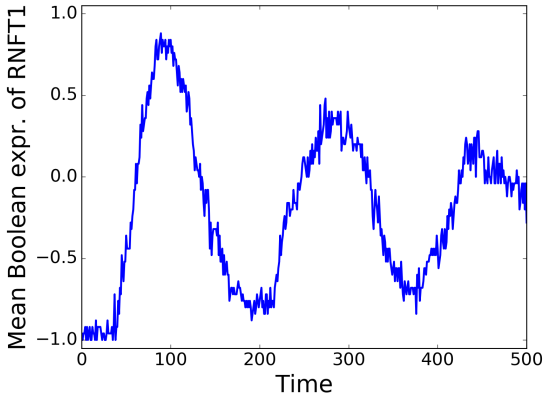


Figure 6.9 Mean of simulated Boolean expression for the gene RNFT1 across 50 isolated cells as a function of time. All cells began in identical states, but because transitions occur stochastically, the 50 cells decohere over time, leading a decaying sinusoid similar to Fig. 6.8.

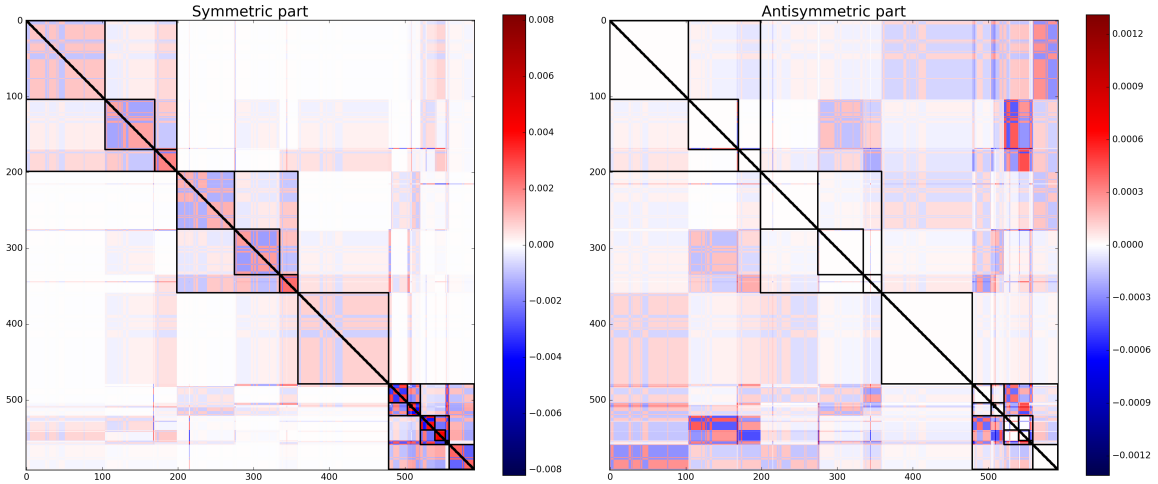


Figure 6.10 Cell cycle coupling matrix, separated into symmetric (point attractor) and antisymmetric (transition) parts on the left and right, respectively, and sorted into nested communities applied to the symmetric part. Examining the nested community structres may aid in the search for sets of nodes which, when targeted, could control cell cycle.

CHAPTER 7

CONCLUSIONS

This dissertation covered a wide variety of topics, but all were unified by the theme of investigating complex biological systems with relatively simple mathematical models. While my experiences have taught me a great deal, much work remains to be done, and my research has led me to ask more questions than I have answered.

The relationship between nested communities and Hopfield dynamics is intriguing. It has been shown for coupled oscillator models that nested community structures create distinct timescales wherein the lowest levels communities with the densest sets of connections synchronize first, followed by the next lowest, etc. up the hierarchy until all of the nodes in the entire network finally synchronize [136]. This could also happen in Hopfield systems; perhaps starting in a random initial state at low temperature causes a similar hierarchical synchronization. This deserves futher consideration.

As stated in Section 6.3, the ultimate goal of the cell cycle model is predicting the effects of perturbations such as drugs and searching for sets of genes which, when controlled, drive the system toward a desired state. Our collaborators in San Diego can test these predictions experimentally, and we can use the experimental results to further refine our mathematical models.

Another cell cycle data set using *S. cerevisiae* (yeast) cells [48] will also be examined in the future, and its results will be compared with the human data set. Because cell cycle is a fundamental process in many forms of life, I expect to find similarities between the models for humans and yeast, but it is possible that the evolutionary distance between humans and yeast resulted in very different cell cycle mechanisms. This will also provide two separate systems to test experimentally.

I recently unearthed a publication which demonstrates that cyclic attractors can be stored using an asynchronous update scheme [114], which is biologically more realistic than the synchronized case discussed in Section 6.3. This will certainly be integrated into the cell cycle model if possible.

Finally, all of the material presented in this document assumes that the simulated cells are isolated, but it is well known that cells communicate with one another using chemical signals. It may be possible to model cell-cell communication effects (such as the release of growth factors) in addition to the intracellular process of gene regulation. Furthermore, cells may compete over the limited space and resources in the environment. I have done some work with modelling population dynamics using competitive Lotka-Volterra equations with additional terms representing the different effects a set of drugs has on, for example, normal and cancer cells. Understanding not only the genetic heterogeneity of cells within populations, but how these different cells communicate and compete is critical to designing therapies to fight cancer. I plan to continue studying these and similar complex biological systems in my postdoctoral research.

APPENDICES

APPENDIX A

EULER'S SOLUTION TO THE KÖNIGSBERG BRIDGE PROBLEM

Assume an Eulerian path exists. Aside from the starting and ending nodes, every node that was entered via one bridge must be exited via another bridge, guaranteeing that all of the intermediate nodes have even degree. If the starting node was distinct from the ending node, then each of them must have odd degree, and if the starting node was the same as the ending node, then that node must have even degree. Because all four nodes in Fig. 2.2 have odd degree, no such path exists.

APPENDIX B

INTERSET EFFICIENCY NORMALIZATION

The interset efficiency,

$$E_{IJ} = \frac{1}{|I||J| - |I \cap J|} \sum_{\substack{i \in I \\ j \in J \\ i \neq j}} \frac{1}{d_{ij}},$$

measures how efficiently nodes in set J signal to nodes in set I. Below is the derivation of the number of terms in the summation, $(|I||J| - |I \cap J|)$.

Fig. B.1 shows an Euler diagram for the relationship between sets I and J, allowing for a non-empty intersection between them. For convenience, define

$$A = J \setminus I$$

$$B = J \cap I$$

$$C = I \setminus J$$

All that is left is to count the number of distinct pairs (i, j) such that

1. $j \in A$ and $i \in B$: $|A| \times |B| = |J \setminus I| \times |I \cap J|$
2. $j \in B$ and $i \in C$: $|B| \times |C| = |I \cap J| \times |I \setminus J|$
3. $j \in A$ and $i \in C$: $|A| \times |C| = |J \setminus I| \times |I \setminus J|$
4. $j \in B$, $i \in B$, and $i \neq j$: $|B| \times (|B| - 1) = |I \cap J| \times (|I \cap J| - 1)$

The total number of terms is the sum of these pieces,

$$\begin{aligned} \text{Normalization} = \; & |J \setminus I| \times |I \cap J| + && (A \text{ to } B) \\ & |I \cap J| \times |I \setminus J| + && (B \text{ to } C) \\ & |J \setminus I| \times |I \setminus J| + && (A \text{ to } C) \\ & |I \cap J| \times (|I \cap J| - 1) && (B \text{ to } B) \end{aligned}$$

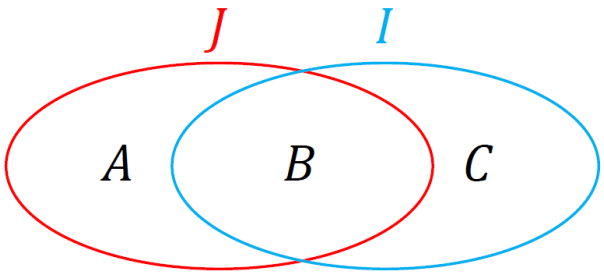


Figure B.1 Two overlapping sets, I and J.

Performing the substitutions

$$|I \setminus J| = |I| - |I \cap J|$$

$$|J \setminus I| = |J| - |I \cap J|$$

and canceling terms results in

$$\text{Normalization} = |I||J| - |I \cap J|. \blacksquare$$

APPENDIX C

NESTED COMMUNITY PLOTS

This appendix contains the spy plots referenced in Section 2.8.

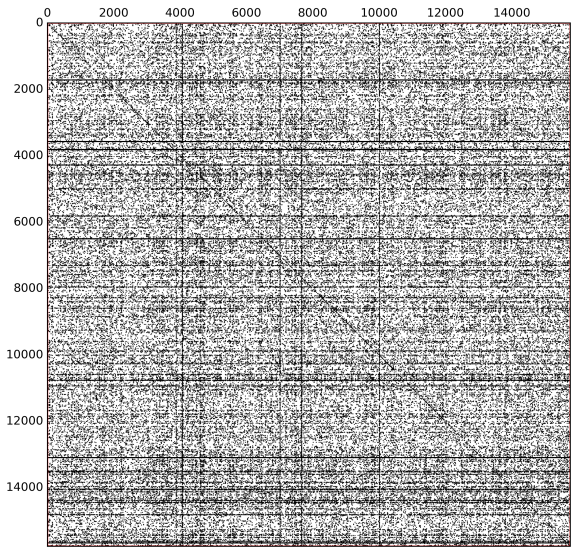


Figure C.1 Spy plot of Google's internal web page network [119]. Each node (row and column indicies) is a web page owned by Google, and each edge (a black dot located at (i, j)) is a hyperlink from page j to page i. Nodes were assigned a random index.

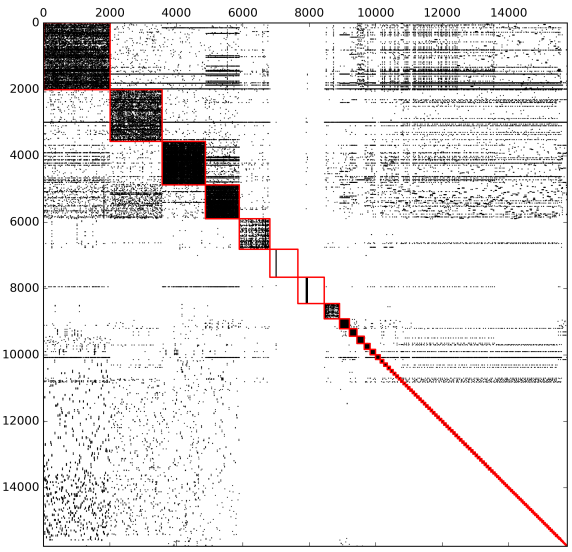


Figure C.2 Same network as Fig. C.1, but the nodes indices have been sorted into level-1 communities (boxed in red). This is the standard result from normal community detection.

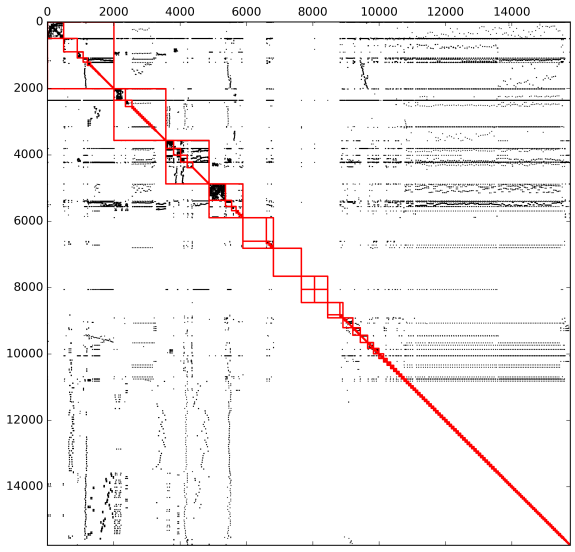


Figure C.3 Same network as Fig. C.1, but the nodes indices have been sorted into level-2 communities. Note that the new level-2 communities are inset in the level-1 communities from Fig. C.2 (boxes within boxes).

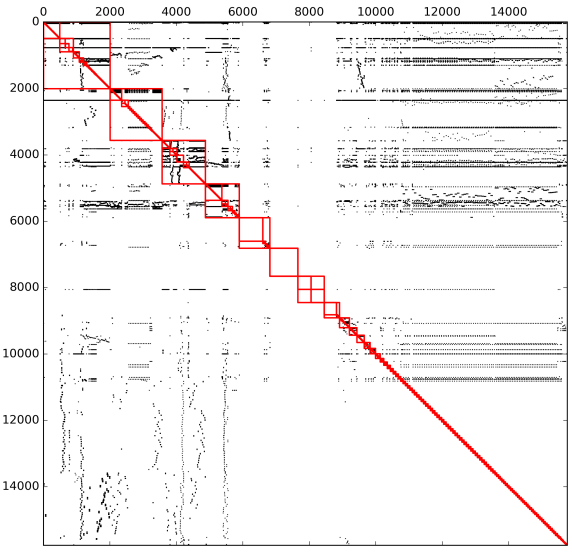


Figure C.4 Same network as Fig. C.1, but the nodes indices have been sorted into level-3 communities.

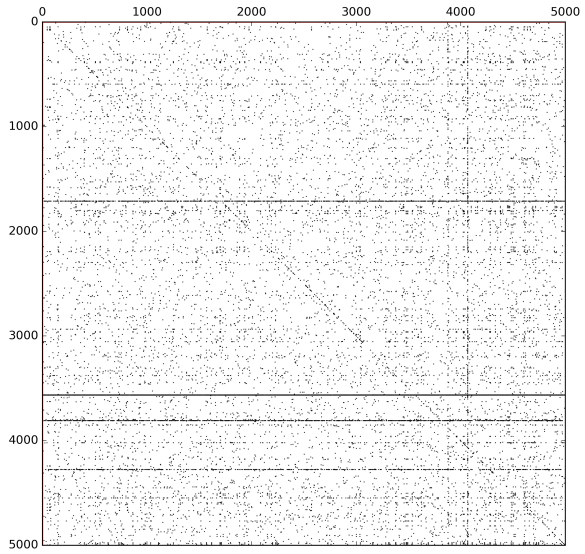


Figure C.5 Same network as shown in Fig. C.1, zoomed in to focus on the first 5,000 nodes.

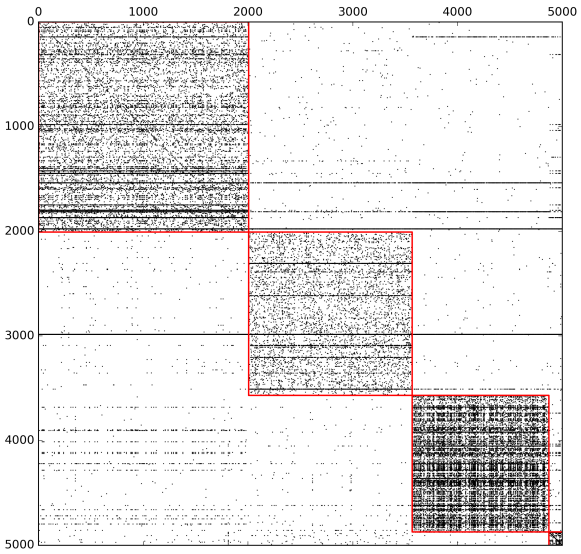


Figure C.6 Same network as shown in Fig. C.2, zoomed in to focus on the first 5,000 nodes.

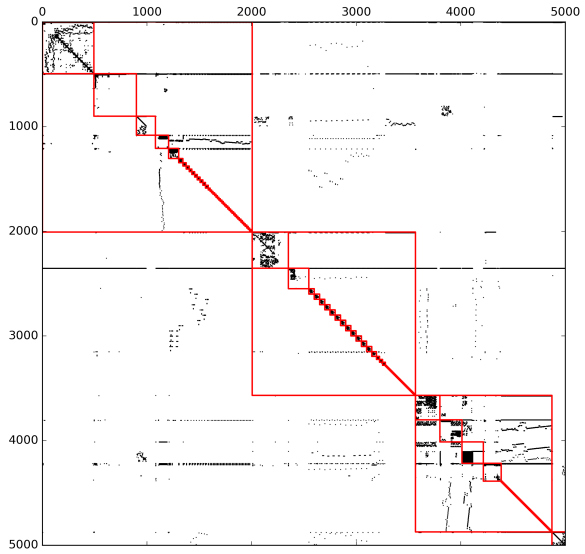


Figure C.7 Same network as shown in Fig. C.3, zoomed in to focus on the first 5,000 nodes.

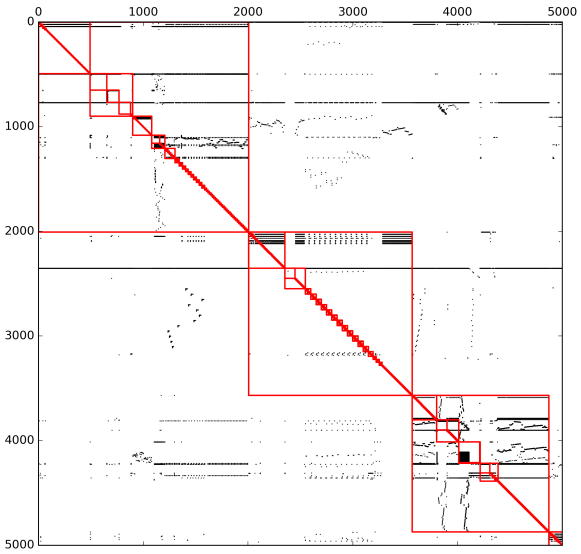


Figure C.8 Same network as shown in Fig. C.4, zoomed in to focus on the first 5,000 nodes.

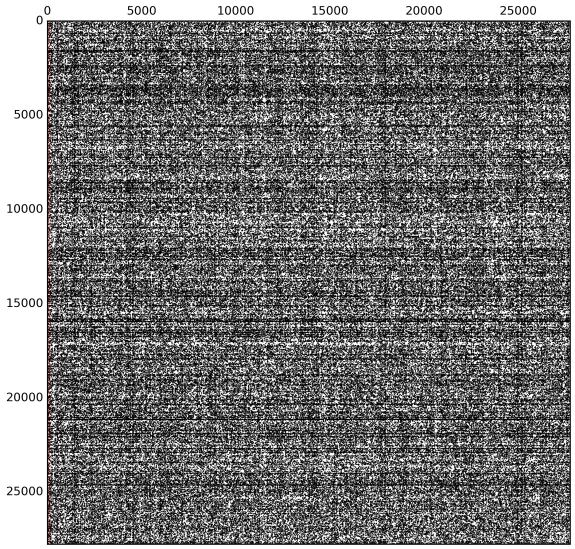


Figure C.9 Spy plot of an arXiv theoretical high energy physics citation network [90]. Each node (row and column indicies) is an author of at least one high energy physics publication, and an edge (a black dot located at (i, j)) from author j to author i means that one of j's articles on arXiv cited at least one of i's articles. Nodes were assigned a random index.

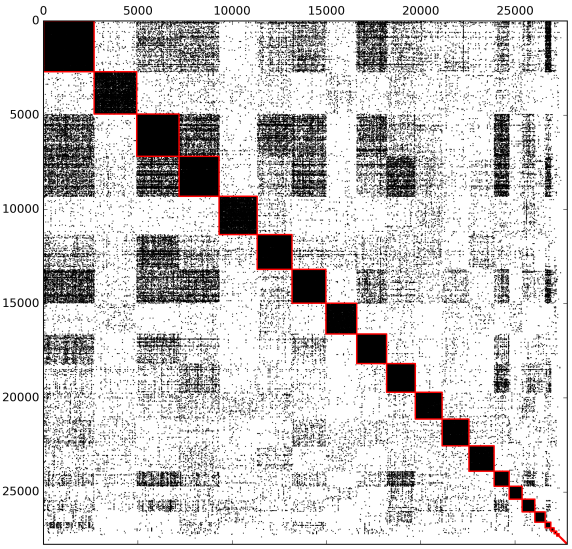


Figure C.10 Same network as Fig. C.9, but the nodes indices have been sorted into level-1 communities. This is the standard result from normal community detection.

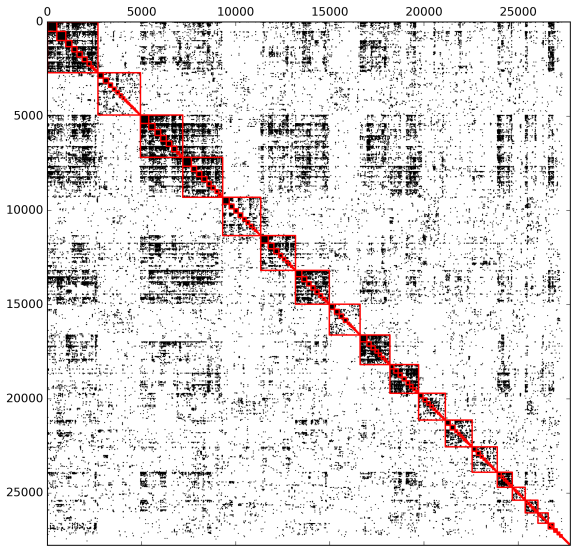


Figure C.11 Same network as Fig. C.9, but the nodes indices have been sorted into level-2 communities.

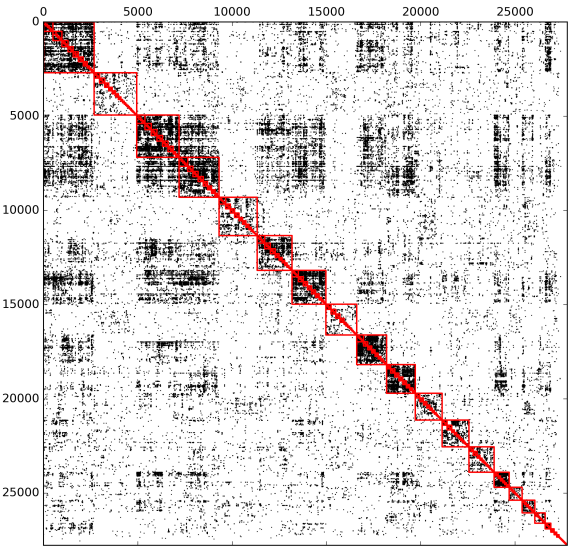


Figure C.12 Same network as Fig. C.9, but the nodes indices have been sorted into level-3 communities.

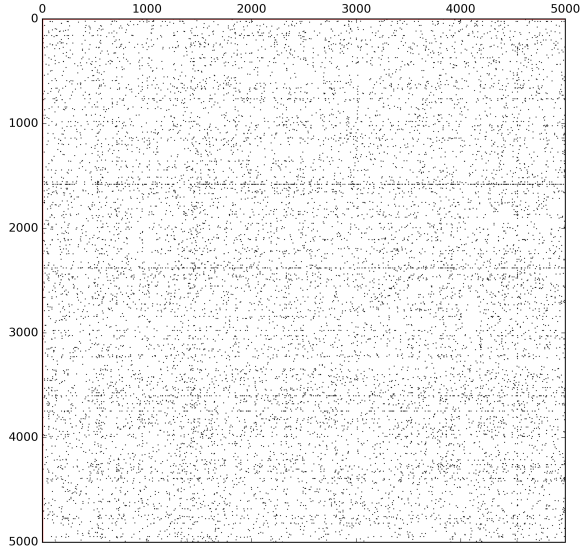


Figure C.13 Same network as shown in Fig. C.9, zoomed in to focus on the first 5,000 nodes.

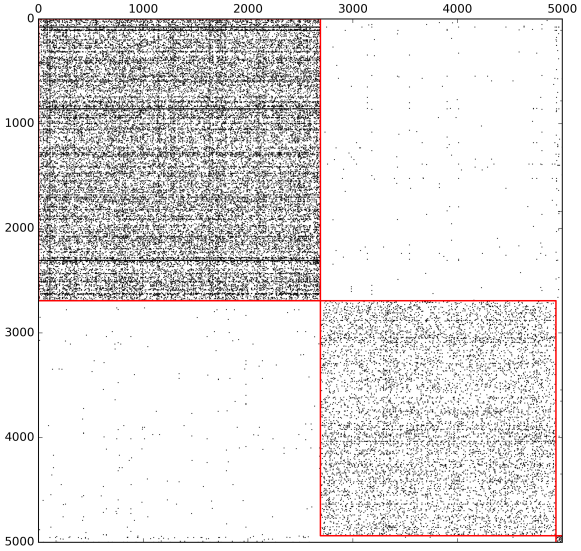


Figure C.14 Same network as shown in Fig. C.10, zoomed in to focus on the first 5,000 nodes.

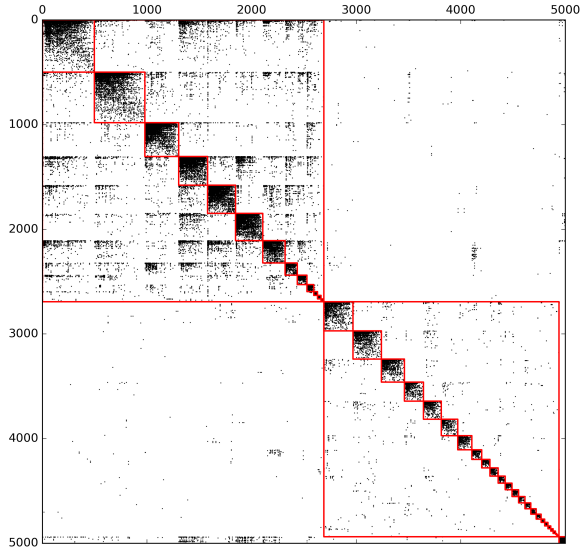


Figure C.15 Same network as shown in Fig. C.11, zoomed in to focus on the first 5,000 nodes.

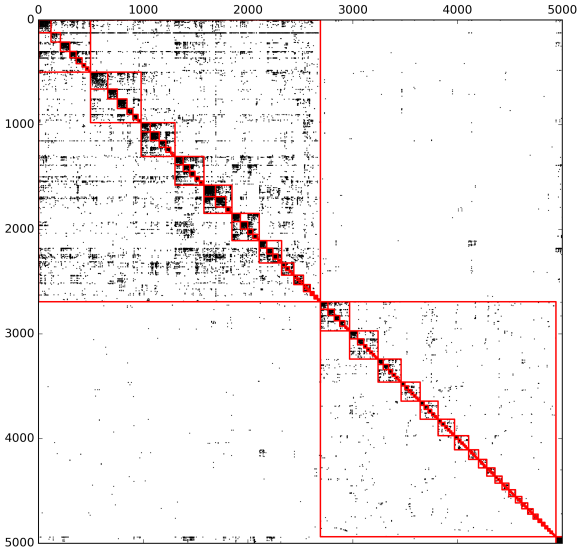


Figure C.16 Same network as shown in Fig. C.12, zoomed in to focus on the first 5,000 nodes.

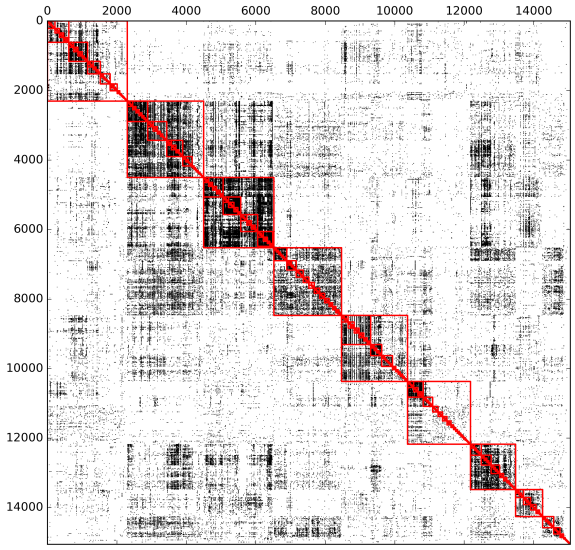


Figure C.17 Spy plot of AML 2.3, a gene regulatory network, sorted into nested communities. Each node (row and column indicies) is gene, and an edge (a black dot located at (i, j)) from gene j to gene i means that j regulates the expression of i. The maximum depth of this network is 7.

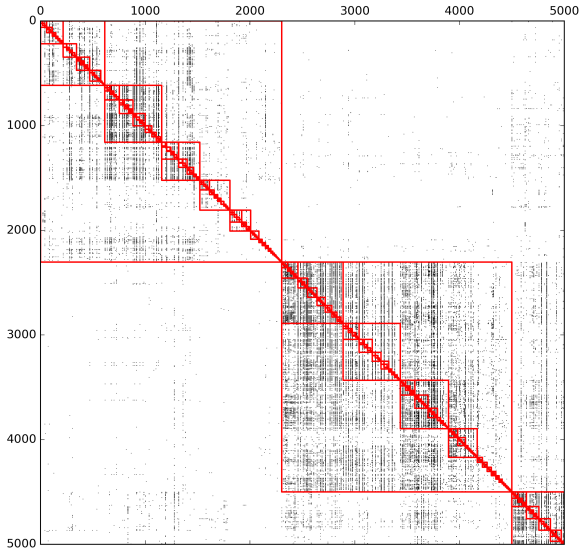


Figure C.18 Same network as shown in Fig. C.17, zoomed in to focus on the first 5,000 nodes.

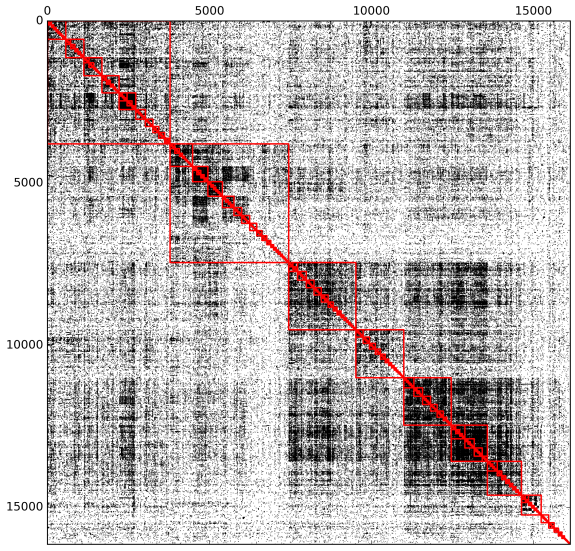


Figure C.19 Spy plot of HumanNet, a gene regulatory network, sorted into nested communities. Each node (row and column indicies) is gene, and an edge (a black dot located at (i, j)) from gene j to gene i means that j regulates the expression of i. The maximum depth of this network is 8.

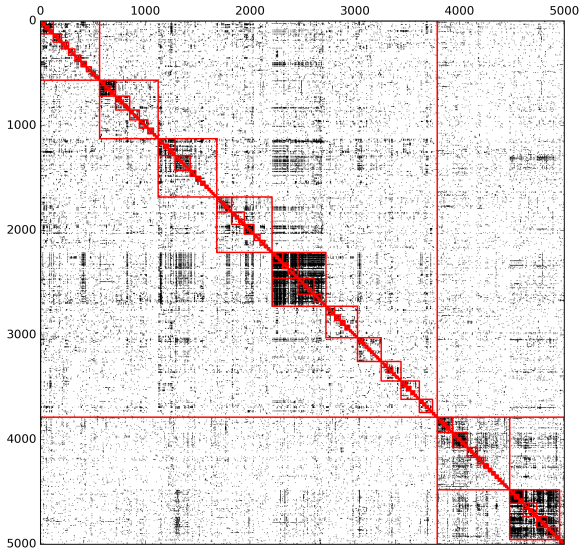


Figure C.20 Same network as shown in Fig. C.19, zoomed in to focus on the first 5,000 nodes.

APPENDIX D

MEASURING EXPRESSION

This appendix is intended as a basic introduction to two of the multitude of methods available for measuring bulk (i.e. many cells pooled together) genome-wide expression profiles. The recent breakneck speed of technological innovation in biological science will likely render this appendix hopelessly obsolete before the author finishes writing it. Any reader interested in detailed coverage of modern gene expression measurement techniques should consult the most recently published review articles available.

Two of the most popular methods today for measuring gene expression profiles are *microarrays* and *mRNA sequencing* (RNA-seq). First introduced around 1983 [144] and firmly established in 1995 [132], typical microarrays use a substrate, called a *chip*, which is subdivided into a lattice of *spots*. Each spot hosts many single-stranded DNA probes corresponding to a given gene from a given organism. Cells of interest are lysed and their mRNA is extracted from the lysate. *Complementary DNA* (cDNA) is synthesized from the mRNA sample by an enzyme derived from retroviruses called *reverse transcriptase* in a process called *reverse transcription* (making cDNA from mRNA, the reverse of making mRNA from DNA). Fluorescent tags are also attached to the resulting cDNA, and the cDNA count is amplified via polymerase chain reaction. The resulting single-stranded cDNA is then passed over the probes, and the cDNA binds to probes with complementary sequences. Finally, a light source is shone over the microarray and light intensities are recorded for each spot, with a bright spot meaning higher expression for that gene. Because the technology relies on bulk measurements of fluorescence rather than counting individual segments of cDNA, microarrays can only quantify relative differences in gene expression. Thus, all measurements (e.g. the gene expression of a sample of acute myeloid leukemia cells) must be paired with controls (e.g. healthy myeloblast cells) to identify which genes are over- or underexpressed. Additionally, different cDNA segments have different binding affinities with their matching probes,

and some can *cross hybridize*, or bind to mismatched probes. Even within a single experiment, the relative expression of a pair of genes cannot be reliably compared: a bright spot can be explained as either a high concentration of mRNA from the given gene, or a low concentration with high binding affinity. This experimental design makes combining microarray measurements from different experiments quite difficult, although more modern microarrays use tens of probes per gene to circumvent this issue. Furthermore, microarrays must be manufactured with a specific set of probes ahead of time, making them unsuitable for searching for novel transcripts in a known organism and unsuitable for less well studied organisms. Although there is a large amount of publicly available microarray data across many organisms and experimental conditions, microarrays are fading in popularity due to their limited quality and the rise of RNA-seq methods.

RNA-seq was introduced around 2008 [93], and was designed to take advantage of the cheap and highly accurate DNA sequencing methods and tools developed for the Human Genome Project in the 1990s. As with microarray technology, RNA-seq requires that cDNA is derived from the mRNA sample via reverse transcription. After the first round of cDNA synthesis and duplication via PCR, however, a second round of PCR is performed in four batches. Each of these four batches contains both normal free nucleotides as well as modified versions of one of the four DNA bases, A, C, G, and T. If a modified base is added to a given cDNA template, sythesis of that strand is halted. For example, a template strand of DNA in the T batch with the sequence ACTGCTTGA would produce the strands ACT, ACTGCT, ACTGCTT, and ACTGCTTGA in the final round of PCR, since there is a fixed probability that the DNA polymerase will pull a modified T from its environment. The same sequence in the C batch would produce the strands AC, ACTGC, and ACTGCTTGA. These reads are then sequenced by a machine via electrophoresis through narrow capillaries, with shorter segments of DNA moving through the gel more quickly than longer segments. To sequence with high fidelity, either the mRNA or the cDNA needs to be fragmented into small pieces (hundreds or thousands of base pairs each, depending on the platform) before being read. It then falls to computers to reconstruct the sequence of strands of cDNA, and thus the original mRNA, from the millions of resulting fragmented reads. RNA-seq is much more accurate

than microarrays because there are no artifacts from varying binding affinity or cross hybridization (but still may suffer from amplification and cDNA conversion artifacts), but the experimental and computational costs are much higher than microarrays. RNA-seq has the added advantage of being able to detect previously unknown transcripts. This has enabled researchers to learn about the regulatory and other roles of non-coding RNAs such as long non-coding RNAs (lncRNA), short interfering RNAs (siRNA), and micro RNAs (miRNA).

APPENDIX E

PROOF OF THEOREM FROM SECTION 5.3.2.2

Theorem: Suppose a network G contains a cycle cluster C. Define the *set of externally influenced nodes*

$$R(C,G) = \{i \in V(C) : j \in V(G \setminus C), (j,i) \in E(G)\} , \tag{E.1}$$

the *set of intruder connections*

$$W(C,G) = \{(j,i) \in E(G) : i \in V(C), j \in V(G \setminus C)\} , \tag{E.2}$$

and the *reduced set of critical nodes*

$$Z_{\text{red}}(C,G) = Z(C, G \setminus W) . \tag{E.3}$$

If $N = |V(C)|$ and

$$\mu \equiv \min_{i \in V(C)} \deg^{-}(i) , \tag{E.4}$$

where $\deg^{-}(i)$ is computed ignoring intruder connections, then

$$\left\lceil \frac{\mu}{2} \right\rceil \leq n_{\text{crit}}(C,G) \leq \zeta , \tag{E.5}$$

where

$$\zeta \equiv \min \left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\text{red}}(C,G)|, N \right) . \tag{E.6}$$

Proof: First, prove the lower limit of Eq. E.5. Let C be a cycle cluster in a network G with $R(C,G) = \{\varnothing\}$. (A cycle cluster in a network with $|R(C,G)| > 0$ will have the same or higher activation barrier for any node in the cluster than the same cycle cluster in a network with $R = \{\varnothing\}$. Since the lower limit of Eq. E.5 is being examined, the case with the lowest activation barrier is considered. Any externally influenced nodes cause n_{crit} to either increase or remain the same.) For any node i to be able to flip away from the cancer state (although not necessarily remain there), it must be that $h_i = -a\xi_i^c$ for $a \geq 0$, meaning that at least half of the nodes upstream of i must point

away from the cancer state. The node i requiring the smallest number of upstream nodes to be in the normal state is the node that satisfies $\deg^-(i) = \mu$. Controlling less than $\mu/2$ nodes will leave all uncontrolled nodes with a field in the cancer direction, and no more flips will occur. Thus,

$$n_{\text{crit}} \geq \left\lceil \frac{\mu}{2} \right\rceil . \tag{E.7}$$

For the upper limit of Eq. E.5, consider a complete *clique* on N nodes, $C = K_N$ (that is, $A_{ij} = 1$ for all $i, j \in V(K_N)$, including self loops) in a network G. First, let there be no connections to any nodes in C from outside of C so that $R(C,G) = \{\varnothing\}$. For odd N, forcing $(N+1)/2$ nodes away from the cancer state will result in the field

$$\sum_j J_{ij}\sigma_j = \left(\frac{N-1}{2} - \frac{N+1}{2}\right)\xi_i^c = -\xi_i^c \tag{E.8}$$

for all nodes i. After one time step, all nodes will flip away from the cancer state. For even N, forcing $N/2$ nodes away from the cancer state will result in the field

$$\sum_j J_{ij}\sigma_j = \left(\frac{N}{2} - \frac{N}{2}\right)\xi_i^c = 0 \tag{E.9}$$

for all nodes i. At the next time step, the unfixed nodes will pick randomly between the normal and cancer state. If at least one of these nodes makes the transition away from the cancer state, the field at all other nodes will point away from the cancer direction. The system will then require one more time step to completely settle to $\sigma_i = -\xi_i^c$. Thus, for $C = K_N$ in a network G with $R(C,G) = \{\varnothing\}$,

$$n_{\text{crit}}(K_N, G) = \left\lceil \frac{N}{2} \right\rceil . \tag{E.10}$$

K_N with $\sigma_i(0) = \xi_i^c$ gives the largest activation barrier for any cycle cluster on N nodes with $R(C,G) = \{\varnothing\}$ to switch away from the cancer attractor state. A general cycle cluster C with any topology on N nodes with $R(C,G) = \{\varnothing\}$ in a network G will have $\deg^-(i) \leq N$ for all nodes i, and so the upper bound is

$$n_{\text{crit}}(C, G) \leq \left\lceil \frac{N}{2} \right\rceil , \tag{E.11}$$

thus proving Eq. E.5 for the special case of $R(C,G) = \{\varnothing\}$.

Now consider a cycle cluster C on N nodes in a network G with $|R(C,G)| \geq 0$. Suppose all nodes in $Z_{\text{red}}(C,G)$ are fixed away from the cancer state. By Eq. E.11, $|Z_{\text{red}}(C,G)| \leq \lceil N/2 \rceil$. For any node $i \in (R(C,G) \cap Z_{\text{red}}(C,G))$, $\sigma_i(t \to \infty) = -\xi_i^c$ is guaranteed because it has already been directly controlled. Any node $i \in (R(C,G) \setminus Z_{\text{red}}(C,G))$ has some incoming connections from nodes $j \notin V(C)$, and these connections could increase the activation barrier enough such that fixing $Z_{\text{red}}(C,G)$ is not enough to guarantee $\sigma_i(t \to \infty) = -\xi_i^c$. To ensure that any node $l \in V(C)$ points away from the cancer state, it is sufficient to fix all nodes $i \in (R(C,G) \setminus Z_{\text{red}}(C,G))$ as well as $Z_{\text{red}}(C,G)$ away from the cancer state. This increases n_{crit} by at most $|R(C,G) \setminus Z_{\text{red}}(C,G)|$, leaving

$$n_{\text{crit}}(C,G) \leq \left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\text{red}}(C,G)| \,. \tag{E.12}$$

n_{crit} can never exceed N, however, because directly controlling every node results in controlling C. It can thus be said that

$$n_{\text{crit}}(C,G) \leq \min\left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\text{red}}(C,G)|, N \right) \,. \tag{E.13}$$

Finally, combining the upper limit in Eq. E.13 with the lower limit from Eq. E.7 gives Eq. E.5. ■

APPENDIX F

HOPFIELD MEAN FIELD SOLUTION

F.1 Mean field for Erdős-Rényi networks

I now rederive Derrida's mean field result [40]. Without loss of generality, define m to be the overlap between attractor 1 and the state vector at time t,

$$m = \frac{1}{N}\sum_{i=1}^{N} \xi_i^1 \sigma_i(t). \tag{F.1}$$

where N is the number of nodes in the network, and $N \to \infty$. Assume that $\{\sigma_i(t)\}$ has macroscopic overlap with attractor 1, and microscopic overlap with all other attractors $\mu \neq 1$. Define the indegree of node i to be k, and label its upstream neighbors $\{j_1, j_2, \ldots, j_k\}$. At a given time t, let n be the number of upstream nodes which have $\sigma_{j_r} = -\xi_i^1$ (and so $k-n$ have $\sigma_{j_r} = +\xi_i^1$). Similarly, define s to be the number of attractors across all upstream nodes which have $\xi_{j_r}^1 = -\xi_{j_r}^{\mu\neq 1}$ (and so $(kp-k-s)$ have $\xi_{j_r}^1 = +\xi_{j_r}^{\mu\neq 1}$). The total field at node i is given by

$$h_i = kp - 2n - 2s. \tag{F.2}$$

The probability for a randomly chosen node to have a given triplet (k,n,s) is given by

$$\begin{aligned} P(k,n,s) &= P(k)P(n,s|k) \\ &= P(k)P(n|k)P(s|k). \end{aligned} \tag{F.3}$$

Directed Erdős-Rényi random networks have a Poisson indegree (and identical outdegree) distribution,

$$P(k) = \frac{c^k e^{-c}}{k!}, \tag{F.4}$$

where c is the mean indegree (which equals the mean outdegree). Since $(m+1)/2$ is the expected fraction of nodes in state $\xi_{j_r}^1$, $P(n|k)$ follows a binomial distribution,

$$P(n|k) = \binom{k}{n}\left(\frac{1+m}{2}\right)^{k-n}\left(\frac{1-m}{2}\right)^{n}. \tag{F.5}$$

Each ξ_{jr}^{μ} is an independent random variable, so for the remaining $p-1$ attractors, $P(s|k)$ also follows a binomial distribution,

$$P(s|k) = \binom{k(p-1)}{s}\left(\frac{1}{2}\right)^{k(p-1)}. \tag{F.6}$$

This means that the expected spin of node i can be computed from

$$\begin{aligned}\langle\sigma_i\rangle = m &= \sum_{k=0}^{\infty}\sum_{n=0}^{k}\sum_{s=0}^{k(p-1)} P(k,n,s)\sum_{\sigma_i=\pm 1}\frac{\sigma_i}{1+e^{-(kp-2n-2s)\sigma_i/T}} \\ &= \sum_{k=0}^{\infty}\sum_{n=0}^{k}\sum_{s=0}^{k(p-1)} P(k,n,s)\tanh\left(\frac{kp-2n-2s}{T}\right)\end{aligned} \tag{F.7}$$

This gives Derrida's mean field equation for a Hopfield system on a dilute directed Erdős-Rényi network,

$$m = \sum_{k=0}^{\infty}\sum_{n=0}^{k}\sum_{s=0}^{k(p-1)}\frac{c^k e^{-c}}{k!}\binom{k}{n}\frac{(1+m)^{k-n}(1-m)^n}{2^{kp}}\binom{k(p-1)}{s}\tanh\left(\frac{kp-2n-2s}{T}\right), \tag{F.8}$$

where c is the mean indegree, p is the number of attractors, and T is the temperature.

In principle, m can be computed from Eq. F.8 numerically. However, the nested summations produce many non-negligible terms for even modest values of c and p (e.g. for $c=5$ and $p=3$, Eq. F.8 is composed of 4940 separate terms when keeping all terms that satisfy $P(k) > 10^{-3}$). Some approximations are now made to solve for m. Recall that

$$\lim_{a\to 0}\frac{1}{\sqrt{\pi a}}e^{-x^2/a} = \delta(x), \tag{F.9}$$

where $\delta(x)$ is the Dirac delta function. This means that for large c and p (but keeping the ratio $(p-1)/c$ fixed),

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{c^k e^{-c}}{k!} &\approx \int_{-\infty}^{+\infty}\frac{dk}{\sqrt{2\pi c}}\exp\left(-\frac{(k-c)^2}{2c}\right) \\ &= \int_{-\infty}^{+\infty}\frac{(dk/c)}{\sqrt{2\pi/c}}\exp\left(-\frac{(\frac{k}{c}-1)^2}{2/c}\right) \\ &\approx \int_{-\infty}^{+\infty} d\left(\frac{k}{c}\right)\delta\left(\frac{k}{c}-1\right).\end{aligned} \tag{F.10}$$

Similarly for large k,

$$\begin{aligned}\sum_{n=0}^{k}\binom{k}{n}(1-q)^{k-n}q^{n} &\approx \int_{-\infty}^{+\infty}\frac{dk}{\sqrt{2\pi kq(1-q)}}\exp\left(-\frac{(n-kq)^2}{2kq(1-q)}\right)\\ &\approx \int_{-\infty}^{+\infty} d\left(\frac{n}{cq}\right)\delta\left(\frac{n}{cq}-1\right).\end{aligned} \tag{F.11}$$

The summation over s can also be approximated as

$$\begin{aligned}\sum_{s=0}^{k(p-1)}\binom{k(p-1)}{s}\left(\frac{1}{2}\right)^{k(p-1)} &\approx \int_{-\infty}^{+\infty}\frac{ds}{\sqrt{\pi\frac{c(p-1)}{2}}}\exp\left(-\frac{(s-\frac{c(p-1)}{2})^2}{\frac{c(p-1)}{2}}\right)\\ &= \int_{-\infty}^{+\infty}\frac{d\left(\frac{s}{c}\right)}{\sqrt{\pi\frac{(p-1)}{2c}}}\exp\left(-\frac{(\frac{s}{c}-\frac{(p-1)}{2})^2}{\frac{(p-1)}{2c}}\right)\\ &= \int_{-\infty}^{+\infty}\frac{dy}{\sqrt{\pi}}e^{-y^2},\end{aligned} \tag{F.12}$$

where $\alpha=(p-1)/c$ and a variable substitution was performed by setting

$$y=\frac{\frac{s}{c}-\frac{p-1}{2}}{\sqrt{\alpha/2}}. \tag{F.13}$$

Combining these approximations results in Derrida's mean field integral equation,

$$m=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}dy\,e^{-y^2}\tanh\left(\frac{cm-cy\sqrt{2\alpha}}{T}\right). \tag{F.14}$$

F.2 Mean field for modular networks

The topology of a stochastic block model network (SBMN) with Ω communities is defined by the nonnegative connectivity matrix $C^{\Omega\times\Omega}$, where C_{IJ} is the mean number of edges from nodes in community J to nodes in community I. Note that while AML 2.3's C_{IJ} is diagonally dominant (i.e. there are more edges internal to communities than between communities), it is not necessary for the following results to hold.

Eq. F.8 will now be generalized for modular networks. Consider a SBMN with connectivity matrix C_{IJ}. Without loss of generality, define the overlap between all spins $i\in I$ and attractor 1 to be

$$m_I=\frac{1}{N_I}\sum_{i\in I}\xi_i^1\sigma_i(t), \tag{F.15}$$

where N_I is the number of nodes in community I, and $N_I \to \infty$. The variables k and n now gain indices to signify the source and target communities for each edge. $P(k_{IJ})$ is the probability that a randomly chosen node $i \in I$ has k_{IJ} upstream neighbors from community J, and is given by a Poisson distribution,

$$P(k_{IJ}) = \frac{C_{IJ}^{k_{IJ}} e^{-C_{IJ}}}{k_{IJ}!}. \tag{F.16}$$

n_{IJ} is the number of these upstream neighbors with a current spin antialigned with attractor 1 (and therefore $k_{IJ} - n_{IJ}$ is the number of upstream neighbors with current spin aligned with attractor 1). n_{IJ} follows a binomial distribution which depends on the overlap of community J with attractor 1,

$$P(n_{IJ}|k_{IJ}) = \binom{k_{IJ}}{n_{IJ}} \left(\frac{m_J+1}{2}\right)^{k_{IJ}-n_{IJ}} \left(\frac{m_J-1}{2}\right)^{n_{IJ}}. \tag{F.17}$$

Since all attractors are composed of independent random variables, the number of upstream attractor diagreements follows a binomial distribution and depends only on the total indegree across all communities,

$$P(s_I|k_I^{\text{tot}}) = \binom{k_I^{\text{tot}}(p-1)}{s_I} \left(\frac{1}{2}\right)^{k_I^{\text{tot}}(p-1)}, \tag{F.18}$$

where $k_I^{\text{tot}} = \sum_{J=1}^{\Omega} k_{IJ}$. The joint probability distribution can now be written as

$$\begin{aligned} P(k_{I1},\ldots,k_{I\Omega};n_{I1},\ldots,n_{I\Omega};s_I) &= \left[\prod_{J=1}^{\Omega} P(k_{IJ},n_{IJ})\right] P(s_I|k_I^{\text{tot}}) \\ &= \left[\prod_{J=1}^{\Omega} P(k_{IJ})P(n_{IJ}|k_{IJ})\right] P(s_I|k_I^{\text{tot}}). \end{aligned} \tag{F.19}$$

Substituting Eqs. F.16, F.17, and F.18 into Eq. F.19 gives the probability density

$$P(\{k_{IJ}\},\{n_{IJ}\},s_I) = \left[\prod_{J=1}^{\Omega} \frac{C_{IJ}^{k_{IJ}} e^{-C_{IJ}}}{k_{IJ}!} \binom{k_{IJ}}{n_{IJ}} (m_J+1)^{k_{IJ}-n_{IJ}} (m_J-1)^{n_{IJ}}\right] \binom{k_I^{\text{tot}}(p-1)}{s_I} \left(\frac{1}{2}\right)^{k_I^{\text{tot}}p}. \tag{F.20}$$

This means that the modular form of the mean field equation is given by

$$m_I = \left[\sum_{k_{I1}=0}^{\infty}\sum_{k_{I2}=0}^{\infty}\cdots\sum_{k_{I\Omega}=0}^{\infty}\right]\left[\sum_{n_{I1}=0}^{k_{I1}}\sum_{n_{I2}=0}^{k_{I2}}\cdots\sum_{n_{I\Omega}=0}^{k_{I\Omega}}\right]\left[\sum_{s_I=0}^{(p-1)\sum_{J=1}^{\Omega}k_{IJ}}\right]$$
$$\times\left[\prod_{J=1}^{\Omega}\frac{C_{IJ}^{k_{IJ}}e^{-C_{IJ}}}{k_{IJ}!}\binom{k_{IJ}}{n_{IJ}}(1+m_J)^{k_{IJ}-n_{IJ}}(1-m_J)^{n_{IJ}}\right]$$
$$\times\binom{\sum_{J=1}^{\Omega}k_{IJ}(p-1)}{s_I}\left(\frac{1}{2}\right)^{\sum_{J=1}^{\Omega}k_{IJ}p}\tanh\left(\frac{\sum_{J=1}^{\Omega}(k_{IJ}p-2n_{IJ})-2s_I}{T}\right). \quad \text{(F.21)}$$

Note that for $\Omega = 1$, Eq. F.21 reduces to Eq. F.8. Again, there are too many terms in Eq. F.21 for computation of m_I to be practical. For large values of p and C_{IJ} for all I and J, the approximations in Eqs. F.10, F.11, and F.12 hold. This means that Eq. F.21 becomes

$$m_I = \left[\prod_{J=1}^{\Omega}\int_{-\infty}^{+\infty}\frac{dk_{IJ}}{\sqrt{2\pi C_{IJ}}}\exp\left(-\frac{(k_{IJ}-C_{IJ})^2}{2C_{IJ}}\right)\int_{-\infty}^{+\infty}\frac{dn_{IJ}}{\sqrt{2\pi k_{IJ}q_J(1-q_J)}}\exp\left(-\frac{(n_{IJ}-k_{IJ}q_J)^2}{2k_{IJ}q_J(1-q_J)}\right)\right]$$
$$\times\int_{-\infty}^{+\infty}\frac{ds_I}{\sqrt{\pi\frac{p-1}{2}\sum_J k_{IJ}}}\exp\left(-\frac{(s_I-\frac{p-1}{2}\sum_J k_{IJ})^2}{\frac{p-1}{2}\sum_J k_{IJ}}\right)\tanh\left(\frac{\sum_J(k_{IJ}p-2n_{IJ})-2s_I}{T}\right). \quad \text{(F.22)}$$

where $q_J \equiv (1-m_J)/2$. Letting $C_I \equiv \sum_J C_{IJ}$, $\alpha_I \equiv (p-1)/C_I$, and taking $p \to \infty$, $C_{IJ} \to \infty$ such that α_I remains constant and the ratios $C_{IJ}/C_{I'J'}$ remain constant for all I, I', J, and J', the Gaussians in k_{IJ} and n_{IJ} turn into delta functions. Only the s_I integral remains, so

$$m_I = \int_{-\infty}^{+\infty} d\left(\frac{s_I}{C_I}\right)\sqrt{\frac{2}{\pi\alpha_I}}\exp\left(-\frac{(2\frac{s_I}{C_I}-p+1)^2}{2\alpha_I}\right)\tanh\left(\frac{\sum_J(k_{IJ}p-2n_{IJ})-2s_I}{T}\right) \quad \text{(F.23)}$$

Performing the substitution

$$y_I = \frac{2\frac{s_I}{C_I}-p+1}{\sqrt{2\alpha_I}} \quad \text{(F.24)}$$

gives the final form for m_I,

$$m_I = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} dy_I\, e^{-y_I^2}\tanh\left(\frac{\sum_{J=1}^{\Omega}C_{IJ}m_J - C_I y_I\sqrt{2\alpha_I}}{T}\right). \quad \text{(F.25)}$$

This system of Ω coupled self-consistent integral equations can be solved numerically. The variance of the magnetization is given by

$$\begin{aligned}(\Delta m_I)^2 &= \left\langle\sigma_I^2\right\rangle - \langle\sigma_I\rangle^2 \\ &= 1 - m_I^2.\end{aligned} \quad \text{(F.26)}$$

Introducing a control term Q_i, our system's magnetization vector becomes

$$m_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \tanh\left(\frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x - y C_i \sqrt{2\alpha_i}}{T} \right). \tag{F.27}$$

Note that

$$\lim_{T \to 0} m_i(T) = \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{w_i} dy\, e^{-y^2} - \int_{w_i}^{\infty} dy\, e^{-y^2} \right) \tag{F.28}$$

where

$$w_i \equiv \frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x}{C_i \sqrt{2\alpha_i}}. \tag{F.29}$$

Define χ_{ij} as the total derivative

$$\chi_{ij} \equiv \left. \frac{dm_i}{dQ_j} \right|_{m_x = m_x^*,\, Q_x = 0\ \forall x}, \tag{F.30}$$

where m_x^* is the numerical solution to (F.27). χ_{ij} is thus the expected change in the magnetization of community i when community j is targeted by an external field. I define this using the total derivative rather than the partial derivative because I'm making no assumptions about m_i remaining constant when m_j changes for $i \neq j$. This is required to see off-target effects. I drop all of the "evaluated at" bars hereafter, but keep in mind that all derivatives are evaluated at $m_x = m_x^*$, $Q_x = 0$ $\forall x$. In general, if $f = f(x_1(t), x_2(t), \ldots, t)$, the total derivative is given by

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}, \tag{F.31}$$

where I am being very careful to correctly distinguish between partial and total derivatives. In our case, since $m_i = m_i(m_1(Q_i), m_2(Q_i), \ldots, m_\Omega(Q_i), Q_i)$,

$$\begin{aligned} \frac{dm_i}{dQ_j} &= \frac{\partial m_i}{\partial Q_j} + \sum_{k=1}^{\Omega} \frac{\partial m_i}{\partial m_k} \frac{dm_k}{dQ_j} \\ &= \frac{\partial m_i}{\partial Q_j} + \sum_{k=1}^{\Omega} \frac{\partial m_i}{\partial m_k} \chi_{kj}, \end{aligned} \tag{F.32}$$

Using the definition from (F.27),

$$\begin{aligned} \frac{\partial m_i}{\partial Q_j} &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \frac{\partial}{\partial Q_j} \tanh\left(\frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x - y C_i \sqrt{2\alpha_i}}{T} \right) \\ &= \delta_{ij} \left(\frac{1}{T\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \operatorname{sech}^2\left(\frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x - y C_i \sqrt{2\alpha_i}}{T} \right) \right) \\ &= \delta_{ij} a_i, \end{aligned} \tag{F.33}$$

where, recalling that $m_x = m_x^*$, $Q_x = 0\ \forall x$,

$$a_i \equiv \frac{1}{T\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \operatorname{sech}^2 \left(\frac{\sum_{x=1}^{\Omega} C_{ix} m_x^* - y C_i \sqrt{2\alpha_i}}{T} \right). \tag{F.34}$$

Note that $a_i = a_i(T)$, so all of the temperature dependence is in a single term. Also note that in general,

$$\lim_{T\to 0} \frac{1}{T} \operatorname{sech}^2 \left(\frac{y_0 - y\gamma}{T} \right) = 2\delta(y_0 - y\gamma) = \frac{2}{|\gamma|} \delta \left(\frac{y_0}{\gamma} - y \right), \tag{F.35}$$

so

$$\lim_{T\to 0} a_i(T) = \sqrt{\frac{2}{C_i^2 \alpha_i \pi}} \exp \left[- \left(\frac{\sum_{x=1}^{\Omega} C_{ix} m_x^*}{C_i \sqrt{2\alpha_i}} \right)^2 \right]. \tag{F.36}$$

Also,

$$\begin{aligned} \frac{\partial m_i}{\partial m_k} &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \frac{\partial}{\partial m_k} \tanh \left(\frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x - y C_i \sqrt{2\alpha_i}}{T} \right) \\ &= C_{ik} \left(\frac{1}{T\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\, e^{-y^2} \operatorname{sech}^2 \left(\frac{Q_i + \sum_{x=1}^{\Omega} C_{ix} m_x - y C_i \sqrt{2\alpha_i}}{T} \right) \right) \\ &= C_{ik} a_i. \end{aligned} \tag{F.37}$$

Putting everything together,

$$\chi_{ij} = a_i \delta_{ij} + \sum_{k=1}^{\Omega} a_i C_{ik} \chi_{kj}. \tag{F.38}$$

If $A_{ij} \equiv a_i \delta_{ij}$ and $B_{ij} \equiv a_i C_{ij}$ (matrices which only depend on T, which is fixed) are defined, then

$$\chi_{ij} = A_{ij} + \sum_{k=1}^{\Omega} B_{ik} \chi_{kj}, \tag{F.39}$$

or in simpler notation,

$$\boldsymbol{\chi} = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{\chi}. \tag{F.40}$$

This can be solved by

$$\boldsymbol{\chi} = (\mathbf{1} - \boldsymbol{B})^{-1} \boldsymbol{A}. \tag{F.41}$$

Since I defined χ_{ij} as the derivative evaluated at equilibrium, χ_{ij} is the solution to the linear equation (F.41). However, χ_{ij} is a nonlinear function of the temperature T and the topology C_{ij} because it depends on a_i.

APPENDIX G

CORRELATED ATTRACTORS

Note: for mathematical convenience, this appendix labels the attractors $\left\{\xi_i^\mu\right\}$ using the scheme $\mu = 0, 1, \ldots, p-1$.

The basic Hopfield model was designed under the assumption that $\xi_i^\mu = \pm 1$ with equal probability for any given (i, μ) pair, and if $N \to \infty$ and $\alpha \ll \alpha_c \approx 0.138$, then $\sum_i \xi_i^\mu \xi_i^\nu / N \approx 0$ for all $\mu \neq \nu$, and the crosstalk term from Eq. 5.13 can be ignored. However, the capacity for correlated attractors is less than 0.138 and depends on the degree of the correlations. A simple trick from linear algebra called a *Moore-Penrose pseudoinverse* can be used to stabilize correlated patterns. If the coupling matrix is constructed as

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \left(Q^{-1}\right)_{\mu\nu} \xi_j^\nu \tag{G.1}$$

for the matix

$$Q_{\mu\nu} = \frac{1}{N} \sum_i \xi_i^\mu \xi_i^\nu \tag{G.2}$$

then for $\sigma_i(t) = \xi_i^\lambda$,

$$\begin{aligned} h_i(t) &= \frac{1}{N} \sum_{j\mu\nu} \xi_i^\mu \left(Q^{-1}\right)_{\mu\nu} \xi_j^\nu \xi_j^\lambda \\ &= \sum_{\mu\nu} \xi_i^\mu \left(Q^{-1}\right)_{\mu\nu} Q_{\nu\lambda} \\ &= \sum_\mu \xi_i^\mu \delta_{\mu\lambda} \\ &= \xi_i^\lambda \end{aligned} \tag{G.3}$$

Note that this only works if the matrix is invertible. If the patterns are uncorrelated, then $Q_{\mu\nu} = \delta_{\mu\nu}$, reducing J_{ij} to the basic form from Eq. 5.6. The matrix $Q_{\mu\nu}$ may also be integrated into the cyclic Hopfield matrix by

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^{\mathrm{mod}p(\mu+1)} \left(Q^{-1}\right)_{\mu\nu} \xi_j^\nu \tag{G.4}$$

Again, if $\sigma_i(t) = \xi_i^{\lambda}$,

$$\begin{aligned} h_i(t) &= \frac{1}{N} \sum_{j\mu\nu} \xi_i^{\mathrm{mod}_p(\mu+1)} \left(Q^{-1}\right)_{\mu\nu} \xi_j^{\nu} \xi_j^{\lambda} \\ &= \sum_{\mu\nu} \xi_i^{\mathrm{mod}_p(\mu+1)} \left(Q^{-1}\right)_{\mu\nu} Q_{\nu\lambda} \\ &= \sum_{\mu} \xi_i^{\mathrm{mod}_p(\mu+1)} \delta_{\mu\lambda} \\ &= \xi_i^{\mathrm{mod}_p(\lambda+1)} \end{aligned} \tag{G.5}$$

In general, using this formulation for the coupling matrix ensures higher fidelity point and cyclic attractors than the original form which omits $Q_{\mu\nu}$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] West Wing, 28 Feb 2001. Season 2, Episode 16, 23:00 mark. CJ Cregg is shocked to learn that the surface area of Africa is 14 times larger than that of Greenland.

[2] V. Ágoston, P. Csermely, and S. Pongor. Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example. *Phys. Rev. E*, 71(5):051909, 2005.

[3] T. Akasaka, I.S. Lossos, and R. Levy. Bcl6 gene translocation in follicular lymphoma: a harbinger of eventual transformation to diffuse aggressive lymphoma. *Blood*, 102(4):1443–1448, 2003.

[4] T. Akutsu, M. Hayashida, W.K. Ching, and M.K. Ng. Control of boolean networks: hardness results and algorithms for tree structured networks. *J. Theor. Biol.*, 244(4):670–679, 2007.

[5] S. Amari, H. Ando, T. Toyoizumi, and N. Masuda. State concentration exponent as a measure of quickness in kauffman-type networks. *Phys. Rev. E*, 87:022814, Feb 2013.

[6] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Phys. Rev. A*, 32(2):1007, 1985.

[7] Ron C. Anafi and Jason H. T. Bates. Balancing robustness against the dangers of multiple attractors in a hopfield-type model of biological attractors. *PLoS ONE*, 5(12):e14413, 12 2010.

[8] Vivek Anantharaman, Eugene V. Koonin, and L. Aravind. Comparative genomics and evolution of proteins involved in rna metabolism. *Nucleic Acids Research*, 30(7):1427–1464, 2002.

[9] P. Ao, D. Galas, L. Hood, and X. Zhu. Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. *Med. Hypotheses*, 70(3):678–684, 2008.

[10] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):1, 2003.

[11] X. Bai, L. Wu, T. Liang, Z. Liu, J. Li, D. Li, H. Xie, S. Yin, J. Yu, Q. Lin, et al. Overexpression of myocyte enhancer factor 2 and histone hyperacetylation in hepatocellular carcinoma. *J. Canc. Res. Clinic. Oncol.*, 134(1):83–91, 2008.

[12] S.J. Baker, E.R. Fearon, J. M. Nigro, A.C. Preisinger, J.M. Jessup, D.H. Ledbetter, D.F. Barker, Y. Nakamura, R. White, B. Vogelstein, et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244(4901):217–221, 1989.

[13] David Balchin, Manajit Hayer-Hartl, and F Ulrich Hartl. In vivo aspects of protein folding and quality control. *Science*, 353(6294):aac4354, 2016.

[14] Paola Ballario, Paola Vittorioso, Armando Magrelli, Claudio Talora, Andrea Cabibbo, and Giuseppe Macino. White collar-1, a central regulator of blue light responses in neurospora, is a zinc finger protein. *The EMBO Journal*, 15(7):1650, 1996.

[15] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[16] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

[17] M.T. Bengoechea-Alonso and J. Ericsson. Tumor suppressor fbxw7 regulates tgfβ signaling by targeting tgif1 for degradation. *Oncogene*, 29(38):5322–5328, 2010.

[18] N. Bhardwaj, M.B. Carson, A. Abyzov, K.-K. Yan, H. Lu, and M.B. Gerstein. Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. *PLoS Comp. Biol.*, 6(5):e1000755, 2010.

[19] Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960, 2012.

[20] Dennis Bray. Protein molecules as computational elements in living cells. *Nature*, 376(6538):307–312, 1995.

[21] L. Bullinger, K. Döhner, E. Bair, S. Fröhling, R.F. Schlenk, R. Tibshirani, H. Döhner, and J.R. Pollack. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New Engl. J. Med.*, 350(16):1605–1616, 2004.

[22] Carlos Caldas. Cancer sequencing unravels clonal evolution. *Nature biotechnology*, 30(5):408–410, 2012.

[23] D. Calzolari, G. Paternostro, P.L. Harrington Jr., C. Piermarocchi, and P.M. Duxbury. Selective control of the apoptosis signaling network in heterogeneous cell populations. *PLoS ONE*, 2(6):e547, 2007.

[24] Neil A. Campbell and Jane B. Reece. *Biology, 6th Edition*. Benjamin Cummings, 2002.

[25] Han Chen, Fangqin Lin, Ke Xing, and Xionglei He. The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nature Communications*, 6, 2015.

[26] Ting Chen, Hongyu L He, George M Church, et al. Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, volume 4, page 4, 1999.

[27] A. Choudhary, A. Datta, M.L. Bittner, and E.R. Dougherty. Intervention in a family of boolean networks. *Bioinformatics*, 22(2):226–232, 2006.

[28] S Ciftci-Yilmaz and R Mittler. The zinc finger network of plants. *Cellular and Molecular Life Sciences*, 65(7-8):1150–1160, 2008.

[29] Mara Compagno, Wei Keat Lim, Adina Grunn, Subhadra V Nandula, Manisha Brahmachary, Qiong Shen, Francesco Bertoni, Maurilio Ponzoni, Marta Scandurra, Andrea Califano, et al. Mutations of multiple genes cause deregulation of nf-κb in diffuse large b-cell lymphoma. *Nature*, 459(7247):717–721, 2009.

[30] S.P. Cornelius, W.L. Kath, and A.E Motter. Realistic control of network dynamics. *Nature Commun.*, 4:1–9, 2013.

[31] P. Creixell, E. M Schoof, J.T. Erler, and R. Linding. Navigating cancer network attractors for tumor-specific therapy. *Nature Biotechnol.*, 30(9):842–848, 2012.

[32] Péter Csermely, Vilmos Ágoston, and Sandor Pongor. The efficiency of multi-target drugs: the network approach might help drug design. *Trends in Pharmacological Sciences*, 26(4):178–182, 2005.

[33] Simon Cutting, Michelle Anderson, Elena Lysenko, Anthony Page, Toshifumi Tomoyasu, Kenji Tatematsu, Takashi Tatsuta, Lee Kroos, and Teru Ogura. spovm, a small protein essential to development in bacillus subtilis, interacts with the atp-dependent protease ftsh. *Journal of Bacteriology*, 179(17):5534–5542, 1997.

[34] Josephine T Daub, Isabelle Dupanloup, Marc Robinson-Rechavi, and Laurent Excoffier. Inference of evolutionary forces acting on human biological pathways. *Genome Biology and Evolution*, page evv083, 2015.

[35] G. De Falco, E. Leucci, D. Lenze, P.P. Piccaluga, P.P. Claudio, A. Onnis, G. Cerino, J. Nyagol, W. Mwanda, C. Bellan, et al. Gene-expression analysis identifies novel rbl2/p130 target genes in endemic burkitt lymphoma cell lines and primary tumors. *Blood*, 110(4):1301–1307, 2007.

[36] R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nature Rev. Microbiol.*, 8(10):717–729, 2010.

[37] Paramvir Dehal and Jeffrey L Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10):1700, 2005.

[38] Glynn Dennis Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, Richard A Lempicki, et al. David: database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3, 2003.

[39] B Derrida and H Flyvbjerg. The random map model: a disordered model with deterministic dynamics. *Journal de Physique*, 48(6):971–978, 1987.

[40] B. Derrida, E. Gardner, and A. Zippelius. An exactly solvable asymmetric neural network model. *Europhys. Lett.)*, 4(2):167, 1987.

[41] Bernard Derrida and Henrik Flyvbjerg. Multivalley structure in kauffman's model: Analogy with spin glasses. *Journal of Physics A: Mathematical and General*, 19(16):L1003, 1986.

[42] Bernard Derrida and Yves Pomeau. Random networks of automata: a simple annealed approximation. *EPL (Europhysics Letters)*, 1(2):45, 1986.

[43] A. Diaz-Alderete, A. Doval, F. Camacho, L. Verde, P. Sabin, R. Arranz-Saez, C. Bellas, C. Corbacho, J. Gil, M. Perez-Martin, et al. Frequency of bcl2 and bcl6 translocations in follicular lymphoma: relation with histological and clinical features. *Leukemia Lymphoma*, 49(1):95–101, 2008.

[44] Daniel Dominguez, Yi-Hsuan Tsai, Nicholas Gomez, Deepak Kumar Jha, Ian Davis, and Zefeng Wang. A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Research*, 2016.

[45] Federal Aviation Administration (2016). Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports for Calendar Year 2015. Retrieved on 22 Aug 2016 from http://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger.

[46] K. Eppert, K. Takenaka, E.R. Lechman, L. Waldron, B. Nilsson, P. van Galen, K.H. Metzeler, A. Poeppl, V. Ling, J. Beyene, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature Med.*, 17(9):1086–1093, 2011.

[47] Adam Ertel and Aydin Tozeren. Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC genomics*, 9(1):1, 2008.

[48] Philipp Eser, Carina Demel, Kerstin C Maier, Björn Schwalb, Nicole Pirkl, Dietmar E Martin, Patrick Cramer, and Achim Tresch. Periodic mrna synthesis and degradation co-operate during cell cycle gene expression. *Molecular systems biology*, 10(1):717, 2014.

[49] S. Fabris, L. Mosca, G. Cutrona, M. Lionetti, L. Agnelli, G. Ciceri, M. Barbieri, F. Maura, S. Matis, M. Colombo, et al. Chromosome 2p gain in monoclonal b-cell lymphocytosis and in early stage chronic lymphocytic leukemia. *Am. J. Hemat.*, 88(1):24–31, 2013.

[50] G. Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76:026107, Aug 2007.

[51] Jacob D Feala, Jorge Cortes, Phillip M Duxbury, Carlo Piermarocchi, Andrew D McCulloch, and Giovanni Paternostro. Systems approaches and algorithms for discovery of combinatorial therapies. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(2):181–193, 2010.

[52] J.D. Feala, J. Cortes, P.M. Duxbury, A.D. McCulloch, C. Piermarocchi, and G. Paternostro. Statistical properties and robustness of biological controller-target networks. *PLoS ONE*, 7(1):e29374, 2012.

[53] Hunter B Fraser, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752, 2002.

[54] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[55] Pauline A. Fujita, Brooke Rhead, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Melissa S. Cline, Mary Goldman, Galt P. Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R. Dreszer, Belinda M. Giardine, Rachel A. Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M. Kuhn, Katrina Learned, Chin H. Li, Laurence R. Meyer, Andy Pohl, Brian J. Raney, Kate R. Rosenbloom, Kayla E. Smith, David Haussler, and W. James Kent. The ucsc genome browser database: update 2011. *Nucleic Acids Research*, 2010.

[56] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Nat. Acad. Sci. USA*, 99(12):7821–7826, 2002.

[57] S. Z. Glud, A. B. Sörensen, M. Andrulis, B. Wang, E. Kondo, R. Jessen, L. Krenacs, E. Stelkovics, M. Wabl, E. Serfling, et al. A tumor-suppressor function for nfatc3 in t-cell lymphomagenesis by murine leukemia virus. *Blood*, 106(10):3546–3552, 2005.

[58] Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[59] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[60] R. Hamid and S.J. Brandt. Transforming growth-interacting factor tgif regulates proliferation and differentiation of human myeloid leukemia cells. *Mol. Oncol.*, 3(5):451–463, 2009.

[61] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[62] C.P. Hans, D.D. Weisenburger, T.C. Greiner, R.D. Gascoyne, J. Delabie, G. Ott, H.K. Müller-Hermelink, E. Campo, R.M. Braziel, E. S. Jaffe, et al. Confirmation of the molecular classification of diffuse large b-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1):275–282, 2004.

[63] Sandra L Harris and Arnold J Levine. The p53 pathway: positive and negative feedback loops. *Oncogene*, 24(17):2899–2908, 2005.

[64] A.J. Hartemink. Reverse engineering gene regulatory networks. *Nature Biotechnol.*, 23(5):554–555, 2005.

[65] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1, 2012.

[66] James Hays and Alexei A Efros. Scene completion using millions of photographs. *Communications of the ACM*, 51(10):87–94, 2008.

[67] John Hertz, Anders Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

[68] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA*, 79(8):2554–2558, 1982.

[69] S. Huang, G. Eichler, Y. Bar-Yam, and D.E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.*, 94:128701, Apr 2005.

[70] Mustafa Hussain, Mahadev Rao, Ashley E Humphries, Julie A Hong, Fang Liu, Maocheng Yang, Diana Caragacianu, and David S Schrump. Tobacco smoke induces polycomb-mediated repression of dickkopf-1 in lung cancer cells. *Cancer research*, 69(8):3570–3578, 2009.

[71] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[72] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[73] Stuart Kauffman, Carsten Peterson, Björn Samuelsson, and Carl Troein. Genetic networks with canalyzing boolean rules are always stable. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17102–17107, 2004.

[74] Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

[75] Haseong Kim and Erol Gelenbe. Stochastic gene expression modeling with hill function for switch-like gene responses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):973–979, 2012.

[76] Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384, 1978.

[77] William S. Klug, Michael R. Cummings, and Charlotte A. Spencer. *Concepts of Genetics, 8th Edition*. Pearson Education, 2006.

[78] Eugene V Koonin. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2):127–136, 2003.

[79] Kirill S Korolev, Joao B Xavier, and Jeff Gore. Turning ecology and evolution against cancer. *Nature Reviews Cancer*, 14(5):371–380, 2014.

[80] Johannes Krause, Qiaomei Fu, Jeffrey M Good, Bence Viola, Michael V Shunkov, Anatoli P Derevianko, and Svante Pääbo. The complete mitochondrial dna genome of an unknown hominin from southern siberia. *Nature*, 464(7290):894–897, 2010.

[81] Dmitry Krotov, Julien O. Dubuis, Thomas Gregor, and William Bialek. Morphogenesis at criticality. *Proceedings of the National Academy of Sciences*, 2014.

[82] Sudhir Kumar, Michael P Suleski, Glenn J Markov, Simon Lawrence, Antonio Marco, and Alan J Filipski. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Research*, 19(9):1562–1569, 2009.

[83] K.E. Kürten. Correspondence between neural threshold networks and kauffman boolean cellular automata. *J. Phys. A*, 21(11):L615, 1988.

[84] K.E. Kürten. Critical phenomena in model neural networks. *Phys. Lett. A*, 129(3):157 – 160, 1988.

[85] A. H. Lang, H. Li, J. J. Collins, and P. Mehta. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *ArXiv e-prints*, page arXiv:1211.3133v3, November 2012.

[86] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 2014.

[87] Lori Layne, Elena Dimitrova, and Matthew Macauley. Nested canalyzing depth and network stability. *Bulletin of mathematical biology*, 74(2):422–433, 2012.

[88] Insuk Lee, U Martin Blom, Peggy I Wang, Jung Eun Shim, and Edward M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121, 2011.

[89] C. Lefebvre, P. Rajbhandari, M.J. Alvarez, P. Bandaru, W.K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B.C Bisikirska, et al. A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.*, 6(1), 2010.

[90] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[91] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International Conference on World Wide Web*, pages 695–704. ACM, 2008.

[92] T.N. Libório, E. N. Ferreira, F. C. Aquino Xavier, D. M. Carraro, L. P. Kowalski, F. A. Soares, and F.D. Nunes. Tgif1 splicing variant 8 is overexpressed in oral squamous cell carcinoma and is related to pathologic and clinical behavior. *Oral Surg. Oral Med.*, 116(5):614–625, 2013.

[93] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536, 2008.

[94] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[95] Camilla Looman, Magnus Åbrink, Charlotta Mark, and Lars Hellman. Krab zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Molecular Biology and Evolution*, 19(12):2118–2130, 2002.

[96] N.M. Luscombe, M.M. Babu, H. Yu, M. Snyder, S.A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, 2004.

[97] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[98] Lesley T MacNeil and Albertha JM Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research*, 21(5):645–657, 2011.

[99] Gerard Manning, Gregory D Plowman, Tony Hunter, and Sucha Sudarsanam. Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, 27(10):514–520, 2002.

[100] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, et al. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, 2003.

[101] Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, Olga V Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

[102] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[103] M.-L. Mo, Z. Chen, H.-M. Zhou, H. Li, T. Hirata, D.M. Jablons, and B. He. Detection of e2a-pbx1 fusion transcripts in human non-small-cell lung cancer. *J. Exp. Clin. Canc. Res.*, 32(1):29, 2013.

[104] C. Montagut and J. Settleman. Targeting the raf–mek–erk pathway in cancer therapy. *Canc. Lett.*, 283(2):125–134, 2009.

[105] Aslaug A Muggerud, Henrik Edgren, Maija Wolf, Kristine Kleivi, Emelyne Dejeux, Jörg Tost, Therese Sørlie, and Olli Kallioniemi. Data integration from two microarray platforms identifies bi-allelic genetic inactivation of ric8a in a breast cancer cell line. *BMC medical genomics*, 2(1):26, 2009.

[106] Arne Müller, Robert M MacCallum, and Michael JE Sternberg. Structural characterization of the human proteome. *Genome research*, 12(11):1625–1641, 2002.

[107] Patricia AJ Muller and Karen H Vousden. p53 mutations in cancer. *Nature cell biology*, 15(1):2–8, 2013.

[108] Javier Munoz, Teck Y Low, Yee J Kok, Angela Chin, Christian K Frese, Vanessa Ding, Andre Choo, and Albert JR Heck. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular systems biology*, 7(1), 2011.

[109] Katy A Muzikar, Nicholas G Nickols, and Peter B Dervan. Repression of dna-binding dependent glucocorticoid receptor-mediated gene expression. *Proceedings of the National Academy of Sciences*, 106(39):16598–16603, 2009.

[110] Mark Newman. *Networks: An Introduction*. Oxford university press, 2010.

[111] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[112] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[113] Yoshihito Niimura. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. *Current Genomics*, 13(2):103–114, 2012.

[114] Hidetoshi Nishimori, Tota Nakamura, and Masatoshi Shiino. Retrieval of spatio-temporal sequence in asynchronous neural network. *Physical Review A*, 41(6):3346, 1990.

[115] Eric N Olson. Gene regulatory networks in the evolution and development of the heart. *Science*, 313(5795):1922–1927, 2006.

[116] Edison Ong, Anthony Szedlak, Yunyi Kang, Peyton Smith, Nicholas Smith, Madison McBride, Darren Finlay, Kristiina Vuori, James Mason, Edward D Ball, et al. A scalable method for molecular network reconstruction identifies properties of targets and mutations in acute myeloid leukemia. *Journal of Computational Biology*, 22(4):266–288, 2015.

[117] Christiane A Opitz, Michael Kulke, Mark C Leake, Ciprian Neagoe, Horst Hinssen, Roger J Hajjar, and Wolfgang A Linke. Damped elastic recoil of the titin spring in myofibrils of human myocardium. *Proceedings of the National Academy of Sciences*, 100(22):12688–12693, 2003.

[118] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[119] Gergely Palla, Illes J Farkas, Peter Pollner, Imre Derenyi, and Tamás Vicsek. Directed network modules. *New journal of physics*, 9(6):186, 2007.

[120] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[121] Ute Paul, Viktor Kaufman, and Barbara Drossel. Properties of attractors of canalyzing random boolean networks. *Physical Review E*, 73(2):026118, 2006.

[122] Mihaela Pertea and Steven L Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11(5):1, 2010.

[123] P.P. Piccaluga, G. De Falco, M. Kustagi, A. Gazzola, C. Agostinelli, C. Tripodo, E. Leucci, A. Onnis, A. Astolfi, M. R. Sapienza, et al. Gene expression analysis uncovers similarity and differences among burkitt lymphoma subtypes. *Blood*, 117(13):3596–3608, 2011.

[124] M. D. Plummer and L. Lovász. *Matching theory*. Elsevier, 1986.

[125] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[126] T. Rohlf and S. Bornholdt. Self-organized criticality and adaptation in discrete dynamical networks. In *Adaptive Networks*, pages 73–106. Springer, 2009.

[127] A. Rosenwald, G. Wright, W.C. Chan, J. M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J. M. Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New Engl. J. Med.*, 346(25):1937–1947, 2002.

[128] S.I. Rothschild, O. Gautschi, E.B. Haura, and F.M. Johnson. Src inhibitors in lung cancer: current status and future directions. *Clin. Lung Canc.*, 11(4):238–242, 2010.

[129] Maureen A Sartor, Vasudeva Mahavisno, Venkateshwar G Keshamouni, James Cavalcoli, Zachary Wright, Alla Karnovsky, Rork Kuick, HV Jagadish, Barbara Mirel, Terry Weymouth, et al. Conceptgen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics*, 26(4):456–463, 2010.

[130] Tatjana Sauka-Spengler, Daniel Meulemans, Matthew Jones, and Marianne Bronner-Fraser. Ancient evolutionary origin of the neural crest gene regulatory network. *Developmental Cell*, 13(3):405–420, 2007.

[131] Martin H Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E Wanker, and Miguel A Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS One*, 7(2):e31826, 2012.

[132] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467, 1995.

[133] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[134] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genet.*, 31(1):64–68, 2002.

[135] Ilya Shmulevich and Stuart A Kauffman. Activities and sensitivities in boolean network models. *Physical review letters*, 93(4):048701, 2004.

[136] Sitabhra Sinha and Swarup Poria. Multiple dynamical time-scales in networks with hierarchically nested modular organization. *arXiv preprint arXiv:1110.2906*, 2011.

[137] E.D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer, 1998.

[138] Thomas A Steitz and Peter B Moore. Rna, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends in Biochemical Sciences*, 28(8):411–418, 2003.

[139] Susanna Stinson, Mark R Lackner, Alex T Adai, Nancy Yu, Hyo-Jin Kim, Carol O'Brien, Jill Spoerke, Suchit Jhunjhunwala, Zachary Boyd, Thomas Januario, et al. Trps1 targeting by mir-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. *Science Signaling*, 4(177):ra41, 2011.

[140] H. Sui, I. Ernberg, and S. Kauffman. Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Sem. Cell Dev. Biol.*, 20(7):869 – 876, 2009.

[141] Anthony Szedlak, Giovanni Paternostro, and Carlo Piermarocchi. Control of asymmetric hopfield networks and application to cancer attractors. *PloS one*, 9(8):e105842, 2014.

[142] Anthony Szedlak, Nicholas Smith, Li Liu, Giovanni Paternostro, and Carlo Piermarocchi. Evolutionary and topological properties of genes and community structures in human gene regulatory networks. *PLOS Computational Biology*, 12(6):e1005009, 2016.

[143] T. Takahashi, M.M. Nau, I. Chiba, M.J. Birrer, R.K. Rosenberg, M. Vinocour, M. Levitt, H. Pass, A.F. Gazdar, and J.D. Minna. p53: a frequent target for genetic abnormalities in lung cancer. *Science*, 246(4929):491–494, 1989.

[144] E TAUB, FLOYD, JAMES M DeLEO, and E BRAD THOMPSON. Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated rnas. *DNA*, 2(4):309–327, 1983.

[145] Diethard Tautz. Evolution of transcriptional regulation. *Current Opinion in Genetics & Development*, 10(5):575–579, 2000.

[146] Sarah A Teichmann and M Madan Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36(5):492–496, 2004.

[147] V.S. Tompkins, S.-S. Han, A. Olivier, S. Syrbu, T. Bair, A. Button, Laura Jacobus, Zebin Wang, Samuel Lifton, Pradip Raychaudhuri, et al. Identification of candidate b-lymphoma genes by cross-species gene expression profiling. *PLoS ONE*, 8(10):e76889, 2013.

[148] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[149] Sonal Tuteja. Enhancement in weighted pagerank algorithm using vol. *IOSR Journal of Computer Engineering (IOSR-JCE), ISSN*, pages 2278–0661, 2013.

[150] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

[151] B.C. Valdez, A.R. Zander, G. Song, D. Murray, Y. Nieto, Y. Li, R.E. Champlin, and B.S. Andersson. Synergistic cytotoxicity of gemcitabine, clofarabine and edelfosine in lymphoma cell lines. *Blood Canc. J.*, 4(1):e171, 2014.

[152] Vera Van Noort, Berend Snel, and Martijn A Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, 5(3):280–284, 2004.

[153] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[154] Klaus W Wagner, Elizabeth A Punnoose, Thomas Januario, David A Lawrence, Robert M Pitti, Kate Lancaster, Dori Lee, Melissa von Goetz, Sharon Fong Yee, Klara Totpal, et al. Death-receptor o-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand apo2l/trail. *Nature medicine*, 13(9):1070–1077, 2007.

[155] J. Walczynski, S. Lyons, N. Jones, and W. Breitwieser. Sensitisation of c-myc-induced b-lymphoma cells to apoptosis by atf2. *Oncogene*, 33:1027–1036, 2013.

[156] L. Wang, S. Pal, and S. Sif. Protein arginine methyltransferase 5 suppresses the transcription of the rb family of tumor suppressors in leukemia and lymphoma cells. *Mol. Cell. Biol.*, 28(20):6262–6277, 2008.

[157] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[158] Eugene P Wigner. The unreasonable effectiveness of mathematics in the natural sciences. richard courant lecture in mathematical sciences delivered at new york university, may 11, 1959. *Communications on pure and applied mathematics*, 13(1):1–14, 1960.

[159] J.N. Winter, E.A. Weller, S.J. Horning, M. Krajewska, D. Variakojis, T.M. Habermann, R.I. Fisher, P.J. Kurtin, W.R. Macon, M. Chhanabhai, et al. Prognostic significance of bcl-6 protein expression in dlbcl treated with chop or r-chop: a prospective correlative study. *Blood*, 107(11):4207–4213, 2006.

[160] Gregory A Wray, Matthew W Hahn, Ehab Abouheif, James P Balhoff, Margaret Pizer, Matthew V Rockman, and Laura A Romano. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, 20(9):1377–1419, 2003.

[161] C.-Y. Yang, C.-H. Chang, Y.-L. Yu, T.-C. E. Lin, S.-A. Lee, C.-C. Yen, J.-M. Yang, J.-M. Lai, Y.-R. Hong, T.-L. Tseng, K.-M. Chao, and C.-Y. F. Huang. Phosphopoint: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, 24(16):i14–i20, 2008.

[162] Hong-Mei Zhang, Hu Chen, Wei Liu, Hui Liu, Jing Gong, Huili Wang, and An-Yuan Guo. Animaltfdb: a comprehensive animal transcription factor database. *Nucleic acids research*, 40(D1):D144–D149, 2012.

[163] Y. Zhang, C. Duan, C. Bian, Y. Xiong, and J. Zhang. Steroid receptor coactivator-1: A versatile regulator and promising therapeutic target for breast cancer. *J. Steroid Biochem.*, 138:17, 2013.