Analyzing Gene-Gene Interactions through a Renormalization of the Ising Model

by

Jomar Sastrillo

Undergraduate Thesis supervised by
Dr. David Sivak

Submitted to
Dr. Paul Haljan

Department of Physics
Faculty of Science

LaTeX template modified from the `sfuthesis.cls` (2017) class

Abstract

To study how gene-gene interactions may be controlled, and driven toward particular gene states, an Ising model has been proposed to model genes as binary interacting spins. To determine the effect of 'clamping' the states of particular genes requires accounting for the other interactions in the network through the renormalization scheme proposed in this thesis.

Keywords: gene-gene interactions, stem cell control, renormalization, Ising model

Acknowledgements

Dr. David Sivak has provided the great opportunity to work on this project and given many insights which helped produce the findings in this work.

I also thank Dr. Malcolm Kennett for pointing out the larger classes of Ising models which people have worked on.

Table of Contents

Chapter 1

Introduction

Stem cells are widely known for their ability to transform into different cell types—like nerve and muscle cells, which serve very specialized functions in the body [8]. Their ability to differentiate is attractive from a medical perspective because of the possibility for new treatments to create and replace damaged cells, and by extension, tissues and organs. Thus, research programs have been dedicated to determining how cellular transformations can be controlled in pursuit of these goals. Awards have been given out for ground breaking techniques in cell "reprogramming," one of the most prominent being the 2012 Nobel Prize in Physiology or Medicine to Shinya Yamanaka and Sir John B. Gurdon [1], for showing how mature cells can be turned back into stem cells.

1.1 A Model of Gene-Gene Interactions

Though cells perform different functions, their core genetic code is identical. Their identity is determined in large part by what genes are transcribed and then translated into proteins. This expression or repression of genes is regulated by other genes in a complex web of interactions. These interactions have been modelled as stochastic and dynamical systems [14, 6, 9] which have found cellular identities to be the steady states of gene expression in a stochastic model or stable fixed points of gene expression in a system of differential equations [11]. Experimental measurement of the expression levels of certain genes have allowed these models to be tested.

The data used to study the expression of every gene in a set of cells has been improved by recent developments in measuring techniques that can measure gene expressions in individual cells. In 2009, Tang et al measured the entire gene expression of a single cell [12], and in 2015 the methods for measurement became enhanced significantly when Macosko et al [3] profiled of the expression levels of 'thousands of individual cells' [3, 13]. By essentially taking snapshots of the gene expression states for individual cells, the accumulated data opens up studies in modelling the distribution of gene expression 'states' [13].

Because of their ability to model interactions, Ising models have been adopted in 'interacting many-body systems' [5]. In biology, Ising models have been used to model the transmission of signals in networks of neurons in the brain [2], and they were used to model gene-gene interactions [10].

A probabilistic model of the distribution of gene expression is a probability function $P(\mathbf{g})$ where $\mathbf{g}$ is a gene expression vector whose elements correspond to the level of gene expression g_i for the corresponding gene i. A simple way to apply the Ising model [13] is to let each element of $\mathbf{g}$ be drawn from $\{-1, +1\}$ where -1 represents a lower than average expression of a gene and $+1$ higher than average. The genes are essentially treated as 2-state spins. The hills and valleys of the probability function P reflect the gene-gene interactions described earlier. The probability is described by the Boltzmann distribution

$$P(\mathbf{g}) = \frac{\exp(-H(\mathbf{g}))}{Z} \ , \tag{1.1}$$

where Z is the normalization constant and the energy

$$H(\mathbf{g}) = -\sum_{i,j} g_i J_{ij} g_j - \sum_i h_i g_i, \quad \mathbf{g} = \{g_i\}_{i=1,2,\ldots,N} \ , \tag{1.2}$$

for a genome of N genes. The constants J_{ij} and h_i are coefficients that are to be determined by fitting the function $P(\mathbf{g})$ to the data. Physically, J_{ij} represents an interaction between genes g_i and g_j. If this coupling is positive, then the genes tend to be expressed or repressed together. This is reflected in the lowering of the energy H. The coupling can also be negative and represents an "anti-aligned" interaction where one gene tends to be expressed when the other is repressed. The average expression level of each gene i are determined by h_i.

Thus, this model of a gene system can be translated in terms of a spin system with ferromagnetic and anti-ferromagnetic interactions and local fields at each spin, where spin up ($\uparrow$) and spin down ($\downarrow$) states are represented by $+1$ and -1 respectively.

1.2 Problem of Control

Where the Ising model illuminated the understanding of the bulk behaviour of materials under different temperatures and magnetic fields [7], this model (i.e. given J_{ij} and h_i for a gene system) will hopefully reveal techniques for cell fate control. Sivak and Thomson were among the first to explore strategies in controlling such Ising models, in this context, to control stem cell fate [13].

To 'drive' a stem cell to become a certain differentiated cell requires specifying a particular gene expression profile $g_{t,0}$ for a certain set of "target genes" $\{g_t\}$ (t are the indices of the target) [13]. Currently, it is possible to set or "program" certain genes (called 'control genes' $\{g_c\}$, where c are the indices of the potential controls) in some desired configuration

$g_{c,0}$ [6]. Thus the problem of control is to (1) determine which control genes g_c to clamp, and (2) in what configuration $g_{c,0}$ they must be set so that the marginal probability distribution $P(g_t = g_{t,0}|g_c = g_{c,0})$ of the target profile is maximized, compared to when the cell is uncontrolled, $P(g_t = g_{t,0})$. The marginal probability of the target profile is expected to determine the number of stem cells which differentiate into the desired cells, once the control is set. Some current programming techniques provide very low yield [6], further motivating this examination.

1.3 Model Properties

1.3.1 Structure of the Interactions

Any set of couplings J'_{ij} can be recast, without any change in the probability distribution function, into a new set of couplings J_{ij} so that $J_{ij} = 0$ for $i \geq j$. Since

$$H'(\mathbf{g}) = -\sum_{i,j} g_i J'_{ij} g_j - \sum_i h_i g_i \tag{1.3}$$

$$= -\sum_i J'_{ii} - \sum_{i<j} g_i (J'_{ij} + J'_{ji}) g_j - \sum_i h_i g_i \ . \tag{1.4}$$

The first sum is a constant and will be eliminated in the normalization of the Boltzmann distribution. The new set of couplings is then

$$J_{ij} = \begin{cases} 0, & i \geq j \\ J'_{ij} + J'_{ji}, & i < j \end{cases}, \tag{1.5}$$

so that $H(\mathbf{g}; J'_{ij}) = H(\mathbf{g}; J_{ij}) + \text{const}$. The resulting matrix representation is upper triangular. Moreover, each pair of gene variables g_i and g_j will now only have one coupling coefficient between them that appears in the Hamiltonian.

1.3.2 Gauge Transformation

Any target state g'_t can be changed to be all spin up with appropriate transformations of J_{ij} and h_i. Suppose a target spin g'_t is "desired" to be spin down, then the following gauge transformation preserves the energy spectrum of the system:

$$g_t \to -g'_t, \quad J_{ti} \to -J_{ti} \ \forall i \text{ and likewise for } J_{it}, \quad \text{and} \quad h_t \to -h_t, \tag{1.6}$$

since the only terms which change in the Hamiltonian are:

$$H(\mathbf{g}') \to -\sum_i (-J_{ti})(-g_i) g_t - \sum_i (-h_i)(-g_i) = H(\mathbf{g}) \ . \tag{1.7}$$

The control solution to the new problem (where all targets are spin up) with new couplings and local fields, is the same as the control solution to the original problem.

1.4 Approach

To simplify the study, we first assume that $h_i = 0$ for all i. This helps illuminate how couplings alone should affect our choice and setting of g_c's. Then we examine a few very simple gene systems to determine the relationships between the control and target genes, and what this suggests for choice and setting of the control. Finally, we take into account the various influences examined and combine them through a renormalization approach between genes.

Chapter 2

Elements of Control

2.1 Probability Calculations

Recall that c and t represent the indices of the control and the target genes respectively. Let a represent the indices of the auxiliary genes—neither target nor control. If a target gene is also a control, simply set the control to be spin up, otherwise the marginal conditional probability of the target states is zero. So we assume that the control gene set does not contain any target genes.

The marginal probability distribution of a specified target gene in the absence and presence of control are,

$$P(g_t) = \frac{\sum_{g'_c=\pm 1,\,\forall c}\cdots\sum \;\sum_{g'_a=\pm 1,\,\forall a}\cdots\sum \exp(-H(g'_c, g_t, g'_a))}{\sum_{g'_t=\pm 1,\,\forall t}\cdots\sum \;\sum_{g'_c=\pm 1,\,\forall c}\cdots\sum \;\sum_{g'_a=\pm 1,\,\forall a}\cdots\sum \exp(-H(g'_c, g'_t, g'_a))} \tag{2.1}$$

and

$$P(g_t|g_c) = \frac{P(g_c, g_t)}{P(g_c)} = \frac{\sum_{g'_a=\pm 1,\,\forall a}\cdots\sum \exp(-H(g_c, g_t, g'_a))}{\sum_{g'_t=\pm 1,\,\forall t}\cdots\sum \;\sum_{g'_a=\pm 1,\,\forall a}\cdots\sum \exp(-H(g_c, g'_t, g'_a))}\,, \tag{2.2}$$

respectively, where the energy as a function of $\mathbf{g}$ is naturally a function of particular genes, with the specified indices $H(\mathbf{g}) = H(g_t, g_c, g_a)$. The form of H placed in an exponential and summed over makes evaluating these probabilities analytically difficult except in a few cases. Thus, the following simple cases permit easy calculation of these sums, and allow us to quantify the effectiveness of control analytically.

2.2 The Simple Cases

2.2.1 One Control, One Target Case

Consider a two-gene system with one target and one control gene. The target gene subprofile is $g_t = +1$ (by convention) and the single coupling is $J_{ct} = J$ between the target and control gene. Then, using eq. (2.1) and eq. (2.2), the marginal probabilities are

$$P(g_t|\, g_c) = \frac{e^{Jg_cg_t}}{e^{Jg_c} + e^{-Jg_c}} \tag{2.3}$$

$$= \frac{e^{Jg_cg_t}}{2\cosh J} \tag{2.4}$$

$$P(g_t) = \frac{e^{-Jg_t} + e^{Jg_t}}{2e^{J} + 2e^{-J}} \tag{2.5}$$

$$= \frac{1}{2}\,. \tag{2.6}$$

A plot of the probabilities are shown in fig. 2.1.

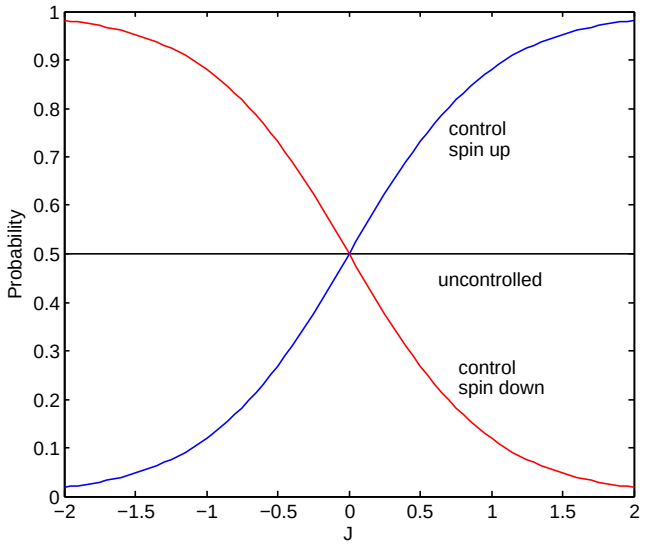

Figure 2.1: Marginal probabilities of the target in the two-gene case.

The qualitative features of the plot are intuitive. If the genes are coupled, then it is favourable to set the control to be spin up. If the genes are anti-coupled, the opposite is true. The control effectiveness increases linearly with the magnitude of the coupling, when the coupling is small. But when the magnitude is large, $J \gtrsim 2$, the control effectiveness levels off. Clearly and notably, if the coupling is small, then there is very little effective control.

When more control genes are present under a single target gene, it is obvious that each control should be set according to the sign of its coupling to the target. But less obvious solutions occur when there are more than one target gene and auxiliary genes.

2.2.2 One Control, Two Target Case

When the system is extended to one control gene g_c with two target genes g_{t_1}, g_{t_2}, the analysis becomes significantly more difficult. We consider various combinations of interactions J_{ct_1}, J_{ct_2} and $J_{t_1t_2}$ between all three genes. The sign of the target-target (t-t) interaction can be positive and negative, so we consider both cases. The plots of the probability improvement through control (normalized by the maximum possible improvement) as a function of the control-target (c-t) interactions is shown in fig. 2.2.

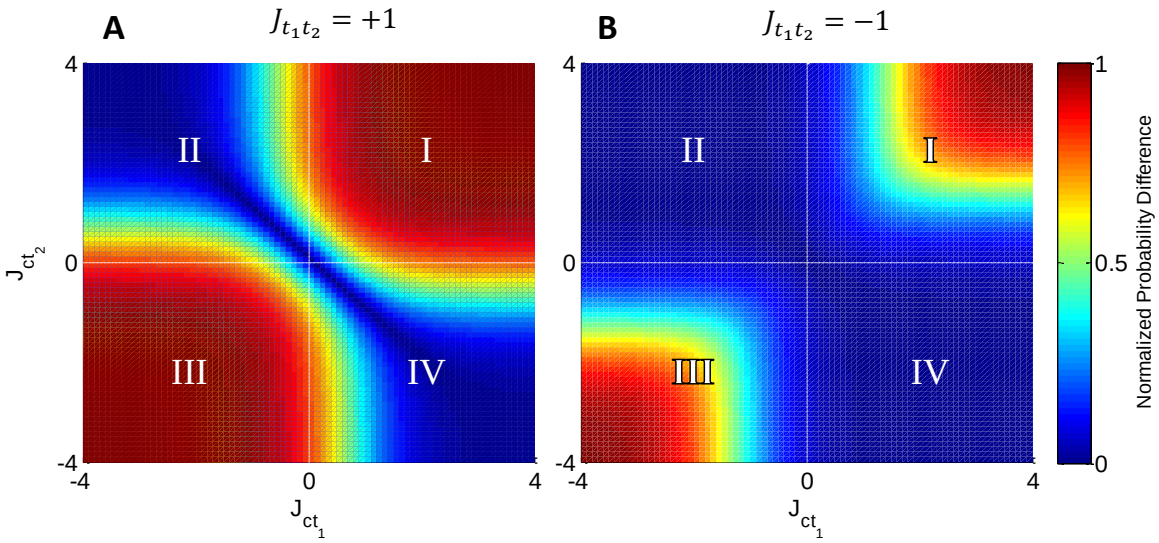

Figure 2.2: The normalized improvement in marginal probability of the target states, upon control. The two target genes are either coupled (left) or anti-coupled (right).

Because we want both target genes in the spin up-configuration, a difference in sign between the two c-t interactions will limit the probability of the desired target configuration, since any setting of the control will tend to make them point in opposing directions.

Nevertheless, the best setting of the single control gene is determined by the stronger of the two c-t interactions. This is concluded from the calculation of the conditional distribution $P(g_{t_1}, g_{t_2}|g_c)$. The results are stark (as illustrated in fig. 2.2): the t-t interactions have a significant influence on the effectiveness of the control.

The control setting is mostly effective when the c-t interactions are of similar sign (which is expected). However, as the red region of the plot shows, it is only guaranteed to be effective if both c-t interactions are sufficiently larger than the t-t interaction. A negative t-t interaction will decrease the probability that the target genes are both spin-up even if both c-t interactions are the same sign. This is illustrated in region I and III of fig. 2.2B. The conclusion from this simple case is that looking at direct c-t interactions are not enough to determining the effectiveness of a control gene.

Chapter 3

Renormalization

At the end of Chapter 2, we saw how interactions other than those directly between the control and target gene can improve or degrade the effectiveness of control. Here, we will examine ways to account for the effect of indirect interactions on the direct interaction between two spins in general.

3.1 The Ising Chain

To aid with the analysis of indirect interactions—that is, interactions between two genes over some chain of interactions—we consider yet another simple case: one target and control gene, and a set of N auxiliary genes, as sketched in fig. 3.1.

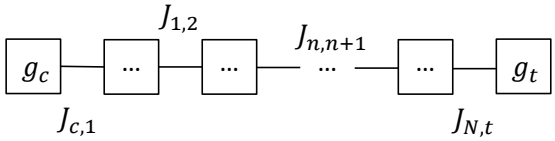

Figure 3.1: One control, one target, and N auxiliary genes, with only nearest-neighbour interactions.

This case is simplified by the assumption that the only interactions in this system are nearest-neighbour interactions along a chain, as shown in fig. 3.1. To determine the conditional probability of the target state, we well-known calculation techniques [7] performed on a another 1D Ising chain model, which, like the system in fig. 3.1, only has nearest-neighbour interactions, but differs in having a boundary condition interaction coefficient between the target and control genes.

3.1.1 Original 1D Ising Chain

Ising-type models are typically used to study the behaviour of bulk systems through the calculation of the partition function of the system. One such system—the 1D Ising chain—is a special case where the function can be computed exactly [7].

The original model consists of N spins, where each spin interacts with its two nearest neighbours through the coupling coefficient $J_{i,i+1}$. Traditionally, the first and last spin of the chain are also interacting, which forms the boundary condition. There is also an assumed global field h applied to each spin, and one can study changes to the partition function as the field is varied (we shall assume this to be zero). The Hamiltonian is

$$H(\mathbf{g}) = \sum_{i=1}^{N} J_{i,i+1} g_i g_{i+1} + \sum_{i=1}^{N} h_i g_i = \sum_{i=1}^{N} J_{i,i+1} g_i g_{i+1} \ , \tag{3.1}$$

where the subscript is implicitly assumed to be mod N, for simplicity of notation.

3.1.2 Transfer Matrix Formulation

In the simple model above, transfer matrices are used to calculate the partition function Z. While Z is usually calculated with temperature and magnetic field terms, we will ignore these since the bulk properties of the system under these variables are not of interest to us. What is of interest, however, is the technique used to calculate sums of spin configurations of $\exp(-H(\mathbf{g}))$, in order to compute the various probabilities of interest in eq. (2.1) and eq. (2.2).

These transfer matrices contain [4], for each interaction coefficient, every configuration of the two spins that the interaction links. The transfer matrix P_{ij} for the interaction term J_{ij} is defined as

$$P_{ij} = \begin{bmatrix} e^{J_{ij}} & e^{-J_{ij}} \\ e^{-J_{ij}} & e^{J_{ij}} \end{bmatrix} \ . \tag{3.2}$$

The element of the transfer matrix in the a-th row and b-th column may be interpreted as

$$(P_{ij})_{ab} = \exp[(g_i)_a J_{ij} (g_j)_b] \ , \tag{3.3}$$

where $(g_i)_a$ denotes the a-th state of the spin (either spin up or down). The indexing of spin states is the same for all spins (i.e. say for example, the b-th state might spin down, but it must be spin down for every spin). The construction of the transfer matrix then encodes all the configurations of the spins it links. The structure of matrix multiplication of two transfer matrices P_{ij} and P_{jk} that share a spin g_j between them is

$$(P_{ij} P_{jk})_{ab} = \sum_{c} \exp[(g_i)_a J_{ij} (g_j)_c] \exp[(g_j)_c J_{jk} (g_k)_b] \ . \tag{3.4}$$

Observe that the states of the spin g_j are summed over in each (a, b)-th element of matrix product. The (a, b)-th element also contains the configuration $(g_i)_a, (g_k)_b$ of the pair of spins g_i and g_k.

Thus, the chain of transfer matrix multiplications linked by shared spins $g_j, g_k \ldots, g_q$,

$$P_{ij} P_{jk} \cdots P_{pq} P_{qr} \ , \tag{3.5}$$

will also sum over all the configurations of the shared spins in each element of the final product. The elements of the final product also encode the configuration of the 'first' and 'last' spins in the chain, namely $(g_i), (g_r)$.

The calculation of Z for the 1D Ising chain with boundary conditions can then come from the transfer matrices for all the interactions in the chain. For the system represented by the Hamiltonian in eq. (3.1) the transfer matrix product is

$$P_{12} P_{23} \cdots P_{N-1\,N} P_{N1} \ . \tag{3.6}$$

Since the index 1 appears on both ends of the chain of transfer matrices, the entries of the final matrix holds the configurations of spin g_1 with itself. Of course, the state of the spin can only be either spin up or down, so the only valid sum of spin configurations in the product in eq. (3.6) are those along the diagonal. Thus the trace of eq. (3.6) is the sum over all spin configurations of the product

$$\prod_{i<j} \exp(g_i J_{ij} g_j) = \exp(-H(\mathbf{g})) \ , \tag{3.7}$$

which is simply the unnormalized Boltzmann distribution of the system. Therefore,

$$Z = \mathrm{Tr}(P_{12} P_{23} \cdots P_{N-1\,N} P_{N1}). \tag{3.8}$$

3.1.3 Clamping in the Original Ising Problem

The formulation used to evaluate Z for *spins* also allows us to evaluate sums involving fixed or pinned *genes*. This is necessary to calculate conditional probabilities, which involve a subset of the sums in the normalization sum Z. This means zeroing offending terms in the sum Z.

For this simple model, the effects of fixing a gene are easily calculable. Since there are only two interactions per gene, the effect of pinning a gene only (directly) affects two other genes. So supposing gene j is fixed to be 'spin up' (biologically in a state of higher than average expression), we must eliminate terms in the matrix where the gene is pointing down. The transfer matrices that accomplish this (and the equivalent operation of pinning

a gene down) are

$$P_\uparrow = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad P_\downarrow = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \ , \tag{3.9}$$

respectively. To fix the gene expression, simply sandwich the appropriate pinning transfer matrix between the transfer matrices that represent the interactions. For example, an ensemble where gene j is pinned in a configuration q will have the pinning matrix in the transfer matrix

$$(P_{ij} P_q P_{jk})_{ab} = \exp[(g_i)_a J_{ij} (g_j)_q] \exp[(g_j)_q J_{jk} (g_k)_b] \ . \tag{3.10}$$

Thus, when all the other matrices are multiplied and the trace is taken, the result is the sum Z, except the terms, where gene j is in a configuration other than q, is zero. The resulting sum is over every gene configuration where gene b is fixed to be q. This would be useful in calculating the marginal probability that gene b is in state q.

These matrices are also special in the sense of pulling out particular elements of a matrix. Thus,

$$P_q (M)_{ab} P_{q'} = M_{qq'} \ , \tag{3.11}$$

where $M_{qq'}$ is the element of M in the q-th row and q'-th column. If M was a chain product of transfer matrices between genes g_i and g_j, the pinning matrices extract the sum of all configurations for genes 'sandwiched' between the matrices, under the configuration $(g_i)_q, (g_j)_{q'}$. This property will be useful in computing the marginal probability of two genes.

3.1.4 One Control, One Target in the Original 1D Ising Chain

We return to the earlier problem of a single target and a single control gene bridged by a chain of length N auxiliary spins. We examine the effect of the auxiliary chain as an indirect interaction between the control and the target.

Using the transfer matrix formalism with pinning matrices, the conditional probability of the target gene state given a control gene state is.

$$P(g_t | g_c) = P(g_t, g_c) / P(g_c) \tag{3.12}$$

$$= \frac{\mathrm{Tr}(P_{c1} P_{12} \cdots P_{Nt} P_t P_{tc} P_c)/Z}{\mathrm{Tr}(P_{c1} P_{12} \cdots P_{Nt} P_{tc} P_c)/Z} \ . \tag{3.13}$$

The pinning matrices P_c and P_t are determined by the variables g_c and g_t according to:

$$P_i = \begin{cases} P_\uparrow, & g_i = 1 \\ P_\downarrow, & g_i = -1 \end{cases} \ . \tag{3.14}$$

Note how we have clamped both the target and control in the numerator but clamped only the control in the denominator, in accordance with Bayes' formula for the conditional probability. Also note that the normalization Z cancels out.

Calculating the matrix products and traces is facilitated by the rather convenient fact that each transfer matrix P_{ij} has a eigenvalue matrix D_{ab} and eigenvector matrix S. The eigenvalue decomposition of the transfer matrix is $P_{ij} = VD_{ij}V^{-1}$ where

$$D_{ij} = \begin{bmatrix} 2\sinh J_{ij} & 0 \\ 0 & 2\cosh J_{ij} \end{bmatrix} , \tag{3.15}$$

and

$$V = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} . \tag{3.16}$$

The inverse of the eigenvector matrix V happens to be $V^{-1} = \frac{1}{2}V$. The eigenvector matrices for the transfer matrix P_{ij} is independent of the interaction coefficient J_{ij}, and therefore constant for all transfer matrices. This conveniently allows us to calculate the product of transfer matrices.

Thus, let

$$\mathcal{P}_{ct} = P_{c1}P_{12}\cdots P_{Nt} . \tag{3.17}$$

This characterize the chain of interactions between the control and the target that pass through other genes (auxiliary genes). Then eq. (3.13) becomes

$$P(g_t|g_c) = \frac{\mathrm{Tr}(\mathcal{P}_{ct}P_tP_{tc}P_c)}{\mathrm{Tr}(\mathcal{P}_{ct}P_{tc}P_c)} . \tag{3.18}$$

Now,

$$\mathcal{P}_{ct} = V^{-1}D_{c1}D_{12}\ldots D_{Nt}V \tag{3.19}$$

$$= \frac{1}{2}V \begin{bmatrix} \prod_{i=c1,12,\ldots,Nt}(2\sinh J_i) & 0 \\ 0 & \prod_{i=c1,12,\ldots,Nt}(2\cosh J_i) \end{bmatrix} V \tag{3.20}$$

$$= \frac{1}{2}\begin{bmatrix} K_C + K_S & K_C - K_S \\ K_C - K_S & K_C + K_S \end{bmatrix} , \tag{3.21}$$

where

$$K_C = \prod_{\substack{i=c1,\\ 12,\ldots,Nt}}(2\cosh J_i) \tag{3.22}$$

$$K_S = \prod_{\substack{i=c1,\\ 12,\ldots,Nt}}(2\sinh J_i) . \tag{3.23}$$

The few remaining matrices may be multiplied and traced out. The denominator of eq. (3.18) becomes,

$$\mathrm{Tr}(\mathcal{P}P_{tc}P_c) = \frac{1}{2}\left(K'_C + K'_S\right) \ , \tag{3.24}$$

where K'_C and K'_S are the following 'extensions' of K_C and K_S respectively:

$$K'_C = (2\cosh J_{ct})K_C \tag{3.25}$$

$$K'_S = (2\sinh J_{ct})K_S \ . \tag{3.26}$$

Although there are two possible choices pinning matrices P_c for the pinning the control gene in the denominator according to g_c, the pinning matrix zeros out one out of the two elements in the diagonal of $\mathcal{P}P_{tc}$. Since the two elements happen to be equal, the trace is independent of the state of the control gene as eq. (3.24) shows.

The trace in the numerator of eq. (3.18) can be simplified to, and summarized as

$$\mathrm{Tr}(\mathcal{P}P_tP_{tc}P_c) = \frac{e^{J_{tc}g_tg_c}}{2}\left(K_C + g_cg_tK_S\right). \tag{3.27}$$

So, the the conditional probability can be calculated and simplified into

$$P(g_t|g_c) = \frac{e^{g_cJ_{ct}g_t}(1 + g_tg_cK_T)}{(1 + K_T\tanh J_{ct})(2\cosh J_{ct})} \ , \tag{3.28}$$

where

$$K_T = \frac{K_S}{K_C} \tag{3.29}$$

$$= \prod_{\substack{i=c1,\\ 12,\ldots,Nt}} \tanh J_i \ . \tag{3.30}$$

In the absence of auxiliary genes, $K_T = 0$, the probability reduces to the one-gene control and one-gene target case (the two-gene case). In fact, if any one of the auxiliary interactions $J_{c1}, J_{12}, \ldots, J_{Nt}$ are severed, the conditional probability again reduces to the two-gene case. Thus, the effectiveness of controlling the target gene is dependent on not only the interaction coefficient J_{ct} between them, but also on the indirect interaction coefficients through the auxiliary genes. To make this point clearer, we recast eq. (3.28) in the equivalent form,

$$P(g_t|g_c) = \frac{1}{2}\frac{(1 + g_tg_c\tanh J_{ct})(1 + g_tg_cK_T)}{1 + K_T\tanh J_{ct}} \ . \tag{3.31}$$

Observe that here, the term K_T is treated on an equal footing with the interaction term $\tanh J_{ct}$. This suggests that K_T behaves as some effective interaction between the target and control. Since K_T is a product of tanhs of interaction couplings (that may be $\pm\infty$),

$|K_T| \leq 1$ so the following effective 'interaction coefficient' would be defined

$$\tilde{J}_{ct} = \tanh^{-1} K_T \tag{3.32}$$

and may also be $\pm\infty$.

3.2 Multiple Independent Auxiliary Strings

3.2.1 Renormalization to One Control, One Target Case

Having established the some notion of an effective interaction along a chain of auxiliary genes, we seek to clarify what this effective coupling means in not just the conditional probability of a target gene state given a control, but on the marginal distribution of the two genes. The interactions between a pair of genes is never just along a single path of interactions through auxiliary genes, but along multiple ones. So to better understand the effect of multiple interaction paths between two genes on the marginal probability distribution of those two genes, we examine the following system illustrated in fig. 3.2.

Consider a control and target gene separated by independent chains of auxiliary genes. The chains are independent: auxiliary genes can be assigned to a chain, and only interact with its 'nearest neighbours' in forming a chain of interactions between the control and the target. Thus, not only can auxiliary genes be assigned to a chain, every interaction coefficient J_{ij} in this system can also be assigned to a chain.

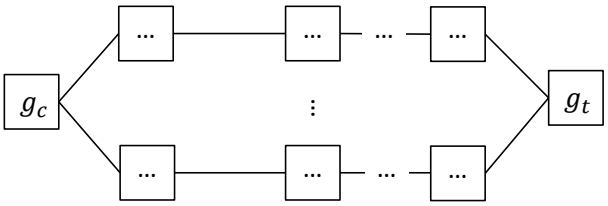

Figure 3.2: A gene system with one control, one target, and multiple auxiliary genes. Each auxiliary gene interacts with only two other genes.

Suppose the system has M independent chains, since each auxiliary gene is independent, the sum over all auxiliary genes can be broken up by chains. Furthermore, since each interaction coefficient J_{ij} also belongs to a chain, the Hamiltonian can be broken up such that each part of the exponential $e^{g_i J_{ij} g_j}$ belongs only to one chain. Thus, the sum over the auxiliaries can be factored out into different chains from 1 to M:

$$\sum_{g'_a=\pm1,\,\forall a} \cdots \sum e^{-H(g_c,g_t,g'_a)} = \left(\sum_{g'_{a_1}=\pm1,\,\forall a_1} \cdots \sum e^{-H(g_c,g_t,g'_{a_1})} \right) \cdots \left(\sum_{g'_{a_M}=\pm1,\,\forall a_M} \cdots \sum e^{-H(g_c,g_t,g'_{a_M})} \right) . \tag{3.33}$$

If we define $\Theta_m(g_c, g_t)$ to represent one of the above factors,

$$\Theta_m(g_c, g_t) = \sum_{g'_{a_m}=\pm 1,\, \forall a_m} \cdots \sum e^{-H(g_c, g_t, g'_{a_m})} \ , \tag{3.34}$$

then the sum over auxiliaries in eq. (3.33) becomes

$$\sum_{g'_{a'}=\pm 1,\, \forall a'} \cdots \sum e^{-H(g_c, g_t, g_{a'})} = \prod_{m=1}^{M} \Theta_m(g_c, g_t) \ . \tag{3.35}$$

Through the transfer matrix formulation, we find that

$$\Theta_m(g_c, g_t) = \frac{1}{2}(1 + g_c g_t K_{T,m}) K_{C,m} \ , \tag{3.36}$$

where $K_{T,m}$ and $K_{C,m}$ function similar to their analogues in the single-chain case, i.e. being the products of hyperbolic tangents and cosines of interactions along chain m:

$$K_{T,m} = \prod_{i \in \mathcal{A}_m} \tanh J_i \tag{3.37}$$

$$K_{C,m} = \prod_{i \in \mathcal{A}_m} 2 \cosh J_i \ , \tag{3.38}$$

where $\mathcal{A}_m = \{ca_{m1}, a_{m1}a_{m2}, \ldots, a_{mn_m}t\}$ is the set of indices that denote all the interactions between genes along the chain m.

Since eq. (3.33) is the traced-out portion of the auxiliary spins, the marginal distribution for just the control and target gene is

$$P(g_c, g_t) = \frac{\prod_{i=1}^{M} \Theta_m(g_c, g_t)}{\sum_{g'_c=\pm 1} \sum_{g_{t'}=\pm 1} \prod_{i=1}^{M} \Theta_m(g'_c, g'_t)} \ . \tag{3.39}$$

Observe that any constant factors on Θ_m cancel each other out in the normalization. Thus, $P(g_c, g_t)$ remains invariant under multiplication of $\Theta_m(g_c, g_t)$ by a constant. This immediately suggests that we can renormalize $\Theta_m(g_c, g_t)$ in the following manner:

$$\Theta_m(g_c, g_t) \propto \frac{1}{2}(1 + g_c g_t K_{T,m}) \tag{3.40}$$

$$\propto \frac{1}{2}(1 + g_c g_t \tanh \tilde{J}_{ct,(m)}) \cosh \tilde{J}_{ct,(m)} \tag{3.41}$$

$$= \exp(g_c \tilde{J}_{ct,(m)} g_t) \ , \tag{3.42}$$

where we define $\tilde{J}_{ct,(m)} = \tanh^{-1} K_{T,m}$. This is the effective interaction coefficient for the mth auxiliary chain between the control and the target, since the marginal distribution eq. (3.39) becomes is equal that of a reduced system with M direct $\tilde{J}_{ct,(m)}$ interactions between the target and control. In other words, the marginal probability eq. (3.39) distribution

takes on the form

$$P(g_c, g_t) = \frac{\exp(-\tilde{H}(g_c, g_t))}{Z} \ , \tag{3.43}$$

where

$$\tilde{H} = -\sum_m g_c \tilde{J}_{ct,(m)} g_t \tag{3.44}$$

$$= -g_c \left(\sum_m \tilde{J}_{ct,(m)} \right) g_t \ . \tag{3.45}$$

This immediately suggests a possible further renormalization by a new Hamiltonian with a single interaction between the target and control gene:

$$\tilde{\tilde{H}}(g_c, g_t) = -g_c \tilde{\tilde{J}}_{ct} g_t, \quad \text{where} \quad \tilde{\tilde{J}}_{ct} = \sum_m \tilde{J}_{ct,(m)} \ . \tag{3.46}$$

The probability distribution of the system which this final Hamiltonian represents has the same marginal distribution as the original multi-chain system.

The significance of this renormalization of the system is profound. We can now determine the overall influence of the control gene on the target gene—indeed, any two genes—provided they are separated by these independent auxiliary chains. Since we have renormalized the system into the marginal distribution between two genes, we can also calculate the reduced conditional marginal probability using Bayes' theorem as $P(g_t|g_c) = P(g_t, g_c)/P(g_c)$. In the case of only one auxiliary chain, eq. (3.39) can be shown to produce the conditional probability in the single-chain case eq. (3.31).

3.2.2 Properties of Renormalization of an Independent Chain

The reduced coupling along an independent chain takes on a special form:

$$K_T = \prod_{i \in \mathcal{A}} \tanh J_i \ , \tag{3.47}$$

where $\mathcal{A}$ is the set of interaction indices in the chain. The reduced interaction $\tilde{J} = \tanh^{-1} K_T$ consequently has some special properties.

To illustrate this, we define a chaining binary operation $\odot$ between two interaction terms that share a spin to be

$$J_{ab} \odot J_{bc} = \tanh^{-1} \left[\tanh(J_{ab}) \tanh(J_{bc}) \right] . \tag{3.48}$$

It is commutative and associative since multiplication is such. It is uniquely defined for any combination of interaction coefficients in $[-\infty, \infty]$.

The operator computes the renormalized interaction $\tilde{J}_{ij}$ along a chain of interactions provided they are independent:

$$\tilde{J}_{ij} = J_{ik} \odot J_{kl} \odot \cdots \odot J_{pj} \ . \tag{3.49}$$

The mathematical properties quantify the physical properties of the renormalized interaction. First, the magnitude of the reduced interaction is no larger than the magnitude of the weakest interaction along the chain. Mathematically, since $|\tanh(x)| \leq 1$,

$$|\tanh J_{ab} \tanh J_{bc}| \leq \min\{|\tanh J_{ab}|, |\tanh_{bc}|\} \ , \tag{3.50}$$

then

$$|J_{ab} \odot J_{bc}| \leq \min\{|J_{ab}|, |J_{bc}|\} \ . \tag{3.51}$$

This quantifies the intuitive idea that information about the state of the control cannot propagate along a link of interactions to the target along a chain in a better way than the weakest interaction in that chain.

The sign of the renormalized interaction represents the type of interaction along the chain. If the chain has an odd number of anti-couplings, the renormalized interaction will also be an anti-coupling. Mathematically,

$$\text{sign}(J_{ab} \odot J_{bc}) = \text{sign}\, J_{ab} \,\text{sign}\, J_{bc} \ . \tag{3.52}$$

The operator $\odot$ can actually be defined between any two interaction coefficients J, but only those that form a chain between spins has a physical interpretation.

3.3 General Renormalization

3.3.1 Scheme

The idea of a general renormalization between any two couplings in an Ising model with general interactions between all spins will be useful in determining the expected effectiveness of control.

The first step is to compute the summing up of all other gene configurations other than two genes of interest: the control g_c and the target g_t. Let

$$\Theta(g_c, g_t) = \sum_{g'_a = \pm 1, \forall a} \cdots \sum \exp(-H(g_c, g_t, g'_a)) \ , \tag{3.53}$$

represent the sum of the unnormalized Boltzmann distribution over all the auxiliary genes. The unnormalized Boltzmann distribution takes on the following form,

$$Zp(\mathbf{g}) = \exp(-H(g_c, g_t, g_{a'})) = \prod_{i<j} \exp(g_i J_{ij} g_j) \ . \tag{3.54}$$

The first trick comes from an earlier observation, writing each of the exponentials as

$$\exp(g_i J_{ij} g_j) = \frac{1}{2}(1 + g_i g_j \tanh J_{ij}) \cosh J_{ij} \ . \tag{3.55}$$

Substituting into eq. (3.54) gives

$$Zp(\mathbf{g}) = \prod_{i<j} \frac{1}{2}(1 + g_i g_j \tanh J_{ij}) \cosh J_{ij} \tag{3.56}$$

The leading constants are not important, as they disappear under normalization of Θ in the final probability calculation. To ease the discussion of the expansion of the product above in eq. (3.56), define

$$\Psi = \prod_{i<j} (1 + g_i g_j \tanh J_{ij}) \tag{3.57}$$

and the following sets of interactions

$$\mathcal{J} = \{J_{i'i^*} : \forall \text{unique pair of gene indices } (i', i^*)\} \ , \tag{3.58}$$

be the set[1] of all interaction coefficients between pairs of genes in a Hamiltonian—where we demand that there is only one interaction coefficient between genes g_i and g_j. The powerset $P(\mathcal{J})$ contains every combination of interaction coefficients. For any such combination $Q \in P(\mathcal{J})$, we may define a renormalization function $R : P(\mathcal{J}) \to [-\infty, \infty]$, according to

$$R(Q) = R\left(\{J_{q'_i q^*_i}\}_{i=1}^{\cdot}\right) \tag{3.59}$$

$$= J_{q'_1 q^*_1} \odot J_{q'_2 q^*_2} \odot \cdots \odot J_{q.'q.^*} \ , \tag{3.60}$$

where we have for convenience, denote the end of a list of interaction indices with a period. Recall that the chaining operation has the following properties

$$\tanh R(Q) = \tanh(J_{q'_1 q^*_1} \odot J_{q'_2 q^*_2} \odot \cdots \odot J_{q.'q.^*}) \tag{3.61}$$

$$= \tanh(J_{q'_1 q^*_1}) \tanh(J_{q'_2 q^*_2}) \cdots \tanh(J_{q.'q.^*}) \ , \tag{3.62}$$

which will become important in the computation of Ψ in eq. (3.57).

[1] We index the set of all interaction coefficients J by i, but to link it to a future discussions, we must identify the gene indices that the interaction coefficient links to. So for example, if J_{12} is the i-th interaction coefficient in the model, then $i' = 1$ and $i^* = 2$.

Now, when the product Ψ is expanded, every combination of interaction coefficients will exist in the expanded sum. This will be encoded by $\tanh R(Q)$, which will be multiplied by each of the gene-pairs $g_{q'_k} g_{q^*_k}$ since the gene-pairs accompany the coefficient $J_{q'_k q^*_k}$ in eq. (3.57). Explicitly,

$$\Psi = 1 + \sum_{Q \in P(\mathcal{J})} g_{q'_1} g_{q^*_1} \cdots g_{q.'} g_{q.^*} \tanh R(Q) \ , \tag{3.63}$$

where the dots in the gene indices $g_{q.'} g_{q.^*}$ also indicate the pair of indices involved the last interaction coefficient $J_{q.'q.^*}$ in the combination Q.

Now, every combination Q of interaction coefficients falls into one of the following cases.

1. The interaction coefficients link genes to form one or multiple loops. If each interaction coefficient in Q belongs in a loop, then let $Q \in \mathcal{L}$.

2. The interaction coefficients can form a path between the control gene and the target gene or can form loops. If every interaction coefficient in Q belongs to a path or a loop, let $Q \in \mathcal{P}_{ct}$. We indicate the indices in this set because it is dependent on the selection of the target and control genes.

3. A wide range of interaction combinations will not belong to any of the above. Let the Q that corresponds to such combinations be an element in $\mathcal{D}_{ct}$, also labeled by ct for the same reason as $\mathcal{P}_{ct}$.

Since $\mathcal{L}, \mathcal{P}_{ct}$, and $\mathcal{D}_{ct}$ partitions $P(\mathcal{J})$. eq. (3.63) may be expanded as

$$\Psi = 1 + \sum_{L \in \mathcal{L}} g_{l'_1} g_{l^*_1} \cdots g_{l.'} g_{l.^*} \tanh R(L) \tag{3.64}$$

$$+ \sum_{P \in \mathcal{P}_{ct}} g_{p'_1} g_{p^*_1} \cdots g_{p.'} g_{p.^*} \tanh R(P) \tag{3.65}$$

$$+ \sum_{D \in \mathcal{D}_{ct}} g_{d'_1} g_{d^*_1} \cdots g_{d.'} g_{d.^*} \tanh R(D) \ . \tag{3.66}$$

Along a loop of interactions, each gene index is repeated an even number of times. Since each gene variable is either ± 1, there will be no gene variables in the sum involving $\mathcal{L}$.

Along a path from g_c to g_t, each gene index is repeated twice except for g_c or g_t which occur an odd number of times. Loops are included in combinations corresponding to $P \in \mathcal{P}_{ct}$ and their corresponding gene variables disappear because they are form part a loop. Thus the only gene variables that appear in the sum involving $\mathcal{P}_{ct}$ are $g_c g_t$.

The gene variables that appear in the sum involving $\mathcal{D}_{ct}$ must contain different combinations of the auxiliary gene indices. Although some will cancel by through repetition as in loops and paths, at least one auxiliary gene variable will not. If there are none, then either the interaction combination corresponding to D forms loops or forms paths between g_c and

g_t or both. If Ψ is summed over all auxiliary gene states, only one leading auxiliary gene variable is sufficient to eliminate the sums involving $\mathcal{D}_{ct}$.

Therefore,

$$\Psi = 1 + \sum_{L\in\mathcal{L}} \tanh R(L) + g_c g_t \sum_{P\in\mathcal{P}_{ct}} \tanh R(P) + \sum_{D\in\mathcal{D}_{ct}} g_{d_1}\, g_{d_1'} \cdots g_{d_v}\, g_{d_v'} \tanh R(D)\ , \quad (3.67)$$

and once a sum over all auxiliary genes is taken, eq. (3.53) becomes

$$\begin{aligned}\Theta_m(g_c, g_t) &= \sum_{g_a'=\pm 1,\,\forall a} \cdots \sum \exp[-H(g_c, g_t, g_a')] \\ &\propto \sum_{g_a'=\pm 1,\,\forall a} \cdots \sum \Psi \qquad (3.68)\\ &= C\left(1 + \sum_{L\in\mathcal{L}} \tanh R(L) + g_c g_t \sum_{P\in\mathcal{P}_{ct}} \tanh R(P)\right)\ , \qquad (3.69)\end{aligned}$$

where C is some constant.

Thus, in summing the exponential over all the configurations of the auxiliary genes, we have found an expression eq. (3.69) that involves renormalized interactions of loops L and of paths P from the control to the target.

We demand that the renormalized coupling coefficient $\tilde{J}_{ct}$ quantifies all the interactions between the control and the target by requiring that the marginal probability distribution of the gene system be equal to that of the renormalized two-gene system with coupling $\tilde{J}_{ct}$

The next trick requires the observation that in a two-gene system (with coupling $\tilde{J}_{ct}$), the probability of two genes being aligned $\tilde{P}(A)$ is related to the probability of being unaligned $\tilde{P}(\bar{A})$ by

$$\begin{aligned}\tilde{P}(A) - \tilde{P}(\bar{A}) &= \frac{2e^{\tilde{J}_{ct}}}{2e^{\tilde{J}_{ct}} + 2e^{-\tilde{J}_{ct}}} - \frac{2e^{-\tilde{J}_{ct}}}{2e^{\tilde{J}_{ct}} + 2e^{-\tilde{J}_{ct}}} \qquad (3.70)\\ &= \tanh \tilde{J}_{ct}\ . \qquad (3.71)\end{aligned}$$

Now, we compute $P(A) - P(\bar{A})$ in the original (unnormalized) system to be

$$P(A) - P(\bar{A}) = \frac{\Theta(1,1) + \Theta(-1,-1) - \Theta(-1,1) - \Theta(1,-1)}{\Theta(1,1) + \Theta(-1,-1) + \Theta(-1,1) + \Theta(1,-1)}\ , \quad (3.72)$$

since $\Theta(g_c, g_t)$ is proportional to the unnormalized Boltzmann distribution. The denominator is clearly just Z. Using the expression for Θ derived in eq. (3.69), we find that

$$P(A) - P(\bar{A}) = \frac{\sum_{P\in\mathcal{P}} \tanh P}{1 + \sum_{L\in\mathcal{L}} \tanh L}\ . \quad (3.73)$$

Therefore, if

$$\boxed{\tanh \tilde{J}_{ct} = \frac{\sum_{P \in \mathcal{P}_{ct}} \tanh R(P)}{1 + \sum_{L \in \mathcal{L}} \tanh R(L)} \ ,} \tag{3.74}$$

then

$$P(A) - P(\bar{A}) = \tilde{P}(A) - \tilde{P}(\bar{A}) \ . \tag{3.75}$$

Since

$$P(A) + P(\bar{A}) = \tilde{P}(A) + \tilde{P}(\bar{A}) = 1 \ , \tag{3.76}$$

it follows that $P(A) = \tilde{P}(A)$ and $P(\bar{A}) = \tilde{P}(\bar{A})$. Thus, if we demand that the normalized system obey the renormalization condition eq. (3.74), then the probability of the control and target being aligned in the unrenormalized system is the same as corresponding probability in the normalized system. Then, the same result holds for the probability of being unaligned.

The following observation completes the renormalization: every gene microstate $\mathbf{g}$ will have a corresponding microstate $-\mathbf{g}$ with the same energy. This is guaranteed by the Hamiltonian:

$$H(\mathbf{g}) \to H(-\mathbf{g}) = -\sum_{i<j} (-g_i) J_{ij} (-g_j) \tag{3.77}$$

$$= -\sum_{i<j} g_i J_{ij} g_j \tag{3.78}$$

$$= H(\mathbf{g}) \ . \tag{3.79}$$

Hence, for each microstate where the control and target genes are both spin up, there is a corresponding microstate of equal energy where they are both spin-down. Therefore the two aligned configurations occur with equal probability, i.e. $P(1,1) = P(-1,-1)$, and the same is true for the two unaligned configurations, i.e. $P(-1,1) = P(1,-1)$. This is true for the unrenormalized system and the two-gene system.

Since the probability of being aligned (and being unaligned) are equal between the two systems, it follows that

$$P(g_c, g_t) = \tilde{P}(g_c, g_t) \ , \tag{3.80}$$

which completes the proof for the renormalization. The marginal probability between the control and target gene is then

$$P(g_c, g_t) = \frac{e^{g_c \tilde{J}_{ct} g_t}}{Z} \ , \tag{3.81}$$

where $\tilde{J}_{ct}$ is given by eq. (3.74). Since the target and controls may refer to any gene, this renormalization works for all pairs of genes.

If $\tilde{J}_{ij}$ has been determined for every gene pair g_i, g_j, then for each target gene, we can identify the one control gene which, when pinned, will maximize the probability that

the target is in a particular state. Once the control gene is pinned, however, the new system will likely have a new set of renormalization coefficients which reduce the marginal probabilities to the two-gene case for the remaining genes. It may even be the case that the renormalization coefficients will not exist.

3.3.2 Consistency Considerations

The renormalization coefficient in eq. (3.74) is only defined only if

$$\left| \frac{\sum_{P \in \mathcal{P}_{ct}} \tanh R(P)}{1 + \sum_{L \in \mathcal{L}} \tanh R(L)} \right| \leq 1 \tag{3.82}$$

since the hyperbolic arctangent function is only defined in $[-1, 1]$. At the moment, this condition is unproven if $\mathcal{P}_{ct}$ and $\mathcal{L}$ are simply given. However, since the formula comes from computation of the difference in the marginal probabilities of being aligned and being unaligned (a number necessarily confined in $[-1, 1]$) in the unrenormalized system, it would not be surprising that the condition holds.

The renormalization formula in eq. (3.74) produces the same result for the multiple, independent auxiliary chain system examined in subsection 3.2.1. Recall that in this system there are M paths of interaction coefficients between the control and target gene. Therefore, the renormalization of the path, which is composed of a set of interactions $P \in \mathcal{P}_{ct}$ is denoted by $R(P_m)$. The proof for showing that the renormalization formula produces the result relies on a messy expansion of

$$\tanh \left[\sum_m R(P_m) \right] \ , \tag{3.83}$$

by applying the formula

$$\tanh(a + b) = \frac{\tanh(a) + \tanh(b)}{1 + \tanh a \tanh b} \ , \tag{3.84}$$

repeatedly. For two chains, the expansion gives the correct terms in the sum of paths and loops, since

$$\tanh(R(P_1) + R(P_2)) = \frac{\tanh R(P_1) + \tanh R(P_2)}{1 + \tanh R(P_1) \tanh R(P_2)} \ , \tag{3.85}$$

and $\tanh R(P_1) \tanh R(P_2) = \tanh R(P_1 \cup P_2)$, where $P_1 \cup P_2$ is the loop through P_1 and P_2. Subsequent addition of paths will produce the necessary paths and loops.

3.3.3 Aspects of Computation

The combinatorics needed to perform the renormalization (that is, finding unique, non-repeating paths and loops) might be challenging.

At the moment, my attempts to produce exact calculations of the renormalization coefficient have been spent on transfer matrices, the multiplication of which preserves paths. To illustrate, let $(J)_{ij}$ represent the entries of a matrix J whose diagonals are zero and whose ij-th element is $\tanh J_{ij}$. The matrix is symmetric to encode the information that the interaction J_{ij} is the same on a path from i to j and vice versa.

In an attempt to find the product of tangents required in the renormalization formula, we note that the matrix products

$$(J^{n+1})_{ij} = \sum_{k_1,k_2,\ldots,k_n} \tanh J_{ik_1} \tanh J_{k_1k_2} \cdots \tanh J_{k_nj} \tag{3.86}$$

is a sum of chains of length $n+1$ from gene i gene j. The diagonals of the matrix product will contain paths that start and end at the same gene, i.e. loops. So the renormalization coefficient $\tilde{J}_{ij}$ was hypothesized to have the form of

$$(\tilde{J})_{ij} = \frac{(M)_{ij}}{1 + \mathrm{Tr}[(M)_{ij}]} \tag{3.87}$$

where

$$M = \sum_{n=1}^{N} (J^n)_{ij}. \tag{3.88}$$

where N is the total number of interactions in the system, and therefore, the maximum path length allowed by the renormalization formula. The sum accounts for paths of lengths n through J^n.

Unfortunately, the paths included in such a matrix M (1) have interaction terms that repeat in the same path and (2) the loops that appear in the diagonal are repeated in other places of the diagonal, since loops start and end on any spin it travels through and (3) loops that count multiple times in the entry in the diagonal because the trace matrix multiplication considers different directions in a loop as distinct contributions. These are just a few examples of how the transfer matrices produce paths that are not in the renormalization scheme. Manipulations to avoid these problems have yielded other problems as well. For example, attempting to correct for the appearance of repeated interaction coefficients along a path will eliminate certain combinations of paths that are required in the renormalization. Thus, at the moment, there does not seem to be a way to handle all the constraints required for choosing which paths are added in the renormalization.

The problems in trying to making exact calculations have extended to finding approximations by transfer matrix methods. Because of the inclusion or exclusion of paths in the sums of the matrix multiplication, the approximations have not been shown to converge to the renormalization formula in eq. (3.74).

3.4 Final Remarks on Renormalization

This study has started on trying to find the optimal control choice and setting to maximize the probability of a target gene profile. In some ways this process was hindered by not being able to take into account the effects of indirect interactions, which the renormalization scheme attempts to resolve. By reducing the interaction between any two genes by summing over paths, we showed that it is theoretically possible to account for the influence of an interaction network between two genes through Ising model, under some simplifying assumptions. The renormalization through a single chain showed how effective control can be degraded by the weakest interaction in the chain. Examining the renormalization of multiple, independent chains showed how the renormalized interactions added up and therefore, how control can be enhanced or reduced. Finally, the general renormalization showed how the addition of independent interactions changed when they become interconnected, by incorporating the loops that are prevalent in an interconnected system like the general Ising model.

An examination of computational barriers and of selection strategies will necessarily follow from this renormalization formulation. The renormalization requires finding paths to sum over, which is a combinatorics problem that becomes exaggerated when the gene system has hundreds of thousands of genes. Transfer matrices seem to provide a lower computation cost to summing over paths but from early examinations, these set of paths do not reflect what is required in the renormalization and this hampers the ability to provide approximations that can be shown to converge to the renormalization answer.

Finally, we have also spent little time examining the strategic consequences for the renormalization. The renormalized interactions do not necessarily say something about how a control affects a block of targets as a whole. And although we can identify the largest effective interactions in the system, these will be expected to change when genes start being pinned. This is anticipated to create mean fields and 'block' certain interaction paths in the system. Whether this will require new renormalization techniques or more clever analytical devices is a question for future study. A possible starting point for the renormalization for an Ising model with a non-zero fields is the incorporation of such fields at the start of the derivation of the renormalization scheme in this thesis.

Bibliography

[1] Nobel Media AB. The nobel prize in physiology or medicine – 2012 press release.

[2] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[3] Evan Z. Macosko, Anindita Basu, Rahul Satija, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161:1202–1214, 2015.

[4] Marc Mezard and Andrea Montanari. *Information, Physics and Computation*. Oxford University Press, 2009.

[5] Hidetoshi Nishimori. *Statistical Physics of Spin Glass and Information Processing*. Oxford University Press, 2001.

[6] Dmitri Papatsenko, Huilei Xu, Avi Ma'ayan, and Ihor Lemischka. *Quantitative Approaches to Model Pluripotency and Differentiation in Stem Cells*, pages 59–74. Springer New York, New York, NY, 2013.

[7] Michael Plischke. *Equilibrium Statistical Physics*. World Scientific Publishing, 3rd edition, 2006.

[8] Stewart Sell, editor. *Stem Cells Handbook*. Springer New York, New York, 2013.

[9] Cameron Sokolik, Yanxia Liu, David Bauer, Jade McPherson, Michael Broeker, Graham Heimberg, Lei S. Qi, David A. Sivak, and Matt Thomson. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Systems*, 1(2):117–129, 2015.

[10] Richard R. Stein, Debora S. Marks, and Chris Sander. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLOS Computational Biology*, 11(7):1–22, 07 2015.

[11] Steven Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley Publishing, 1994.

[12] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, and et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, Jun 2009.

[13] Matt Thomson and David A. Sivak. Learning predictive network models from noisy high-throughput single-cell data. unpublished grant proposal, 2016.

[14] Bin Zhang and Peter G. Wolynes. Stem cell differentiation as a many-body problem. *Proceedings of the National Academy of Sciences*, 111(28):10185–10190, 2014.