On Mathematical Aspects of Genomic Analysis

by

Vera Cherepinsky

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

September 2003

Bud Mishra

(Dissertation Advisor)

To the memory of my maternal grandparents, Evgenia and Alexandr Emerel, who did not get a chance to watch me grow up.

Acknowledgments

I would like to take this opportunity to thank all the people in my life who encouraged, supported, and motivated me.

First, I'd like to thank my advisor, Bud Mishra, without whom this work would not have been possible. He has been a driving force, a source of ideas, and a sounding board, and I feel privileged to have had the opportunity to work with him for the last several years. I sincerely hope that our collaborations continue in the future.

More thanks go to collaborators who had a big part in moving forward the work presented here: Jiawu Feng and Marc Rejali, together with Bud, worked on the shrinkage metric project with me; Ghazala Hashmi and Michael Seul spent a lot of time with us, working on the hybridization model project, taking the results of simulations directly to the lab, and vice versa.

I would like to express a deep appreciation to the members of my defense committee: Joel Spencer, Charlie Peskin, Jack Schwartz, and Michael Seul.

I am deeply grateful to all the wonderful teachers who made learning fun; there was no shortage of them both at Polytechnic University and at New York University.

Everyone in the NYU Bioinformatics Group: Thank you for being my colleagues and my friends; graduate school wouldn't have been the same without you. You listened, advised, bounced ideas back and forth, and in general, created an atmo-

sphere that made all the difference. Thank you, Will Casey, Raoul Daruwala, Archi Rudra, Marco Antoniotti, Violet Chang, Jiawu Feng, Gilad Lerner, Joe McQuown, Toto Paxia, Marc Rejali, Naomi Silver, Marina Spivak, Bing Sun, Nadia Ugel, and Joey Zhou. For all that, I am very grateful.

I want to thank Tamar Arnon for always being there, helping with whatever needed help, making life as a graduate student easier, and most importantly, being my friend.

Finally, I would like to thank my wonderful family for instilling in me a love for learning, for their support, and their love. My grandparents Maia and Mikhail; my parents Tatiana and Alex; my brother Igor and his family (Lena and Jenny); my new family, Tatiana, Michael, and Stella; and most of all, my husband Igor. I know that I would never have gotten this far if it weren't for you, and I want to acknowledge this as our success.

Thank you.

Doctoral Dissertation Abstract

This thesis focuses on three problems that reveal themselves at different stages of genomic analysis: gene expression analysis, analysis of errors in microarray experiments due to unintended probe-target interaction in a multiplexed setup, and the design of an optimal microarray hybridization experiment for genotyping.

The problem of clustering gene expression vectors is known to be significantly dependent on the choice of similarity metric. A mathematically rigorous correlation coefficient of two gene expression vectors, based on James-Stein shrinkage estimators, is derived; the improvement in accuracy due to shrinkage is evaluated by conducting *in silico* experiments and comparing similarity metrics on a biological example. The relative merits of clustering algorithms based on different statistical correlation coefficients as well as the sensitivity of the clustering algorithm to small perturbations in the correlation coefficients are studied.

A detailed physical model of hybridization is presented as a means of understanding probe interactions in a multiplexed reaction. The model is formulated as a system of ordinary differential equations describing kinetic mass action, with conservation-of-mass equations completing the system. Pairwise probe interactions are examined in detail; a model of "competition" between the probes for the target, especially when target is in short supply, is presented. These effects are shown to be predictable from the affinity constants for each of the four probe sequences involved, namely, the match and mismatch for both probes. Simulations based on the competitive hybridization model explain the observed variability in the signal of a

given probe when measured in parallel with different groupings of other probes or individually. These simulation results are used for experiment design and pooling strategies.

The problem of genotyping is examined on the example of HLA typing, which has many biological implications; particularly, knowing the correct allele is essential to ensure the compatibility of the donor organ with the recipient. Most of the contemporary techniques are time-consuming and lack optimality. Here, a graph model on the set of potential probes is presented, the HLA typing problem is formulated mathematically as an optimization problem on the graph model, and an algorithm for solving the optimization problem is described.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Background

The success of the Human Genome Project has revolutionized the biological sciences. Biological experiments continue to produce ever increasing amounts of data which require a greater number of researchers and tools to help find the hidden answers that this data contains. While the scientific, clinical, and commercial implications are enormous, progress in this area requires efficient computational and numerical tools.

In the past 5–10 years there has been a new kind of information revolution in biology, facilitated by the availability of inexpensive microarray technology to the research labs. This technology allows hundreds of thousands of experiments to be done in parallel on a single chip, generating an unprecedented abundance of data. To date, microarrays have been used to answer a variety of questions, ranging from sequence analysis to understanding how gene expression patterns vary under different conditions. The data generated leads to a better understanding of

genetic diseases, mechanisms by which cells process information and communicate with each other, and pharmaceutical applications, such as rational drug design. Types of microarrays currently in use include spotted arrays ([43]), high density oligonucleotide arrays ([30]), gene-chips ([24]), and cDNA microarrays ([42]). We now see protein chips, beginning to bypass the intermediate steps of transcription of DNA into RNA and translation of RNA into the amino acid sequences defining proteins, allowing us to see more directly what is going on inside cells.

While the generated data lends itself to traditional analysis, it is hard to see the forest for the proverbial trees; new, more sophisticated tools are needed to extract meaning from the data. Traditional statistics and data analysis techniques continue to be useful in many applications. However, their lack of mathematical sophistication makes them fall short in the face of attributes such as small sample sizes and high dimensionality that often characterize microarray data. As a result, traditional techniques fail to reveal all that the data has to offer.

The unifying theme explored throughout this thesis is correcting the lack of careful attention to the design of microarray experiments in the past, and providing new mathematical tools to better analyze the results of these experiments. This manuscript is an attempt to remedy the situation and, as such, addresses some of the core issues in the proper design of experiments and the analysis of experimental results. The three problems investigated in this thesis, which appear quite disparate at first glance, reveal themselves at different stages of genomic analysis and follow this underlying thread.

In the problem of gene expression analysis, one has to deal with data generated from a small number of experiments conducted on a large number of genes; the

presence of small-sample data necessitates the use of non-conventional algorithms in statistics. There are many sources of error in microarray data, including (give examples here); one of the least explored such sources stems from the presence of large numbers of probe sequences on the chip, resulting in unintended probe-target interactions in multiplexed reactions. Here, competitive hybridization models are presented and analyzed in an effort to understand and, eventually, compensate for the effect of these unintended interactions on the experimental results. Finally, the problem of microarray experiment design is addressed most directly on the example of HLA typing—a biological problem where a given DNA string must be classified into one of the known existing types or identified as a representative of a new type.

1.2 Thesis Outline

This thesis is organized as follows. Chapter 2 deals with gene expression analysis. It presents the details of the derivation and analysis of a metric designed to assess the similarity of a pair of gene expression vectors. The metric is based on James-Stein shrinkage estimators. The improvement in accuracy of cluster analysis due to the use of the shrinkage-based metric is evaluated via *in silico* experiments and a comparison with two of the currently used similarity metrics on a biological example. The work described in chapter 2 has been published, co-authored with Jiawu Feng, Marc Rejali, and Bud Mishra, in its entirety as a technical report ([12]), as well as in short form as a journal paper ([13]). Chapter 3 deals with the analysis of "errors" due to unintended interactions among targets and probes in a multiplexed hybridization experiment. While this phenomenon has been observed

by experimentalists, it has not been adequately explained. A detailed physical model of the probe-target hybridization process is presented, and pairwise probe interactions are examined in detail as a means of understanding the "competitive hybridization" phenomenon. A heuristic based on the thermodynamic parameters of the hybridization process is presented as well. This heuristic serves to predict the extent of the competition effects. Simulations based on the models described explain the observed variation in the signal from a given probe when measured individually or in parallel with groups of other probes. The work discussed in chapter 3 was previously introduced as an oral presentation at the 2003 Cold Spring Harbor Lab Genome Informatics conference and is based on joint research with Michael Seul, Ghazala Hashmi, and Bud Mishra. Chapter 4 focuses on the process of designing a microarray hybridization experiment for the biological problem of HLA typing. Discerning the correct HLA type of a given DNA sequence is essential for determining the compatibility of a donor organ or bone marrow with that of the recipient, and plays a role in many other biological applications. While contemporary methods occasionally employ microarray approaches, they lack optimality. Here, a graph model on the space of potential probes is presented. The problem of designing the "best" microarray, given a set of known HLA sequences, is then formulated mathematically as an optimization problem on the graph model. An algorithm for obtaining a "best" independent set of at most a specified size that solves the optimization problem is described. The spatial arrangement of the selected probe set on the chip surface is also discussed. Finally, chapter 5 summarizes this thesis and suggests directions for future work.

It is suspected that individual chapters of this thesis will appeal to different

readers, based on their fields of interest. With this in mind, each main chapter of this thesis was written to be self-contained, with its own abstract and appendix, so that each can be read separately from the others.

Chapter 2

Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data

ABSTRACT

The current standard correlation coefficient used in the analysis of microarray data, including gene expression arrays, was introduced in [21]. Its formulation is rather arbitrary. We give a mathematically rigorous derivation of the correlation coefficient of two gene expression vectors based on James-Stein Shrinkage estimators. We use the background assumptions described in [21], also taking into account the fact that the data can be treated as transformed into normal distributions. While [21] uses zero as an estimator for the expression vector mean μ, we start with the assumption that for each gene, μ is itself a zero-mean normal random

variable (with *a priori* distribution $\mathcal{N}(0, \tau^2)$), and use Bayesian analysis to update that belief, to obtain *a posteriori* distribution of μ in terms of the data. The estimator for μ, obtained after shrinkage towards zero, differs from the mean of the data vectors and ultimately leads to a statistically robust estimator for correlation coefficients.

To evaluate the effectiveness of shrinkage, we conducted *in silico* experiments and also compared similarity metrics on a biological example using the data set from [21]. For the latter, we classified genes involved in the regulation of yeast cell-cycle functions by computing clusters based on various definitions of correlation coefficients, including the one using shrinkage, and contrasting them against clusters based on the activators known in the literature. In addition, we conducted an extensive computational analysis of the data from [21], empirically testing the performance of different values of the shrinkage factor γ and comparing them to the values of γ corresponding to the three metrics addressed here, namely, $\gamma = 0$ for the Eisen metric, $\gamma = 1$ for the Pearson correlation coefficient, and γ computed from the data for the Shrinkage metric.

The estimated "false-positives" and "false-negatives" from this study indicate the relative merits of clustering algorithms based on different statistical correlation coefficients as well as the sensitivity of the clustering algorithm to small perturbations in the correlation coefficients. These results indicate that using the shrinkage metric improves the accuracy of the analysis.

All derivation steps are described in detail; all mathematical asser-

tions used in the derivation are proven in the appendix.

[**21**] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. **(1998),** ***PNAS USA* 95, 14863–14868.**

2.1 Background

Traditionally, biology has proceeded as an observational science. Robert Hooke, whose work "Micrographia" of 1665 included the first identification of biological cells through his microscopical investigations, had said, "The truth is, the science of Nature has already been too long made only a work of the brain and the fancy. It is now high time that it should return to the plainness and soundness of observations on material and obvious things." Recently, we have seen an unprecedented progress in our observational and experimental abilities, allowing us to understand the structure of a largely unobservable transparent cell. The most prominent step in this direction has been through microarray-based gene expression analysis, providing us with the ability to quantify the transcriptional states of cells.

The most interesting insight can be obtained from transcriptome abundance data within a single cell under different experimental conditions. In the absence of technology to provide one with such a detailed picture, we have to make do with mRNA collected from a small population of cells, even when individual cells within the population may not be completely synchronized. Furthermore, these mRNAs will only give a partial picture, supported only by those genes that we are

already familiar with and possibly missing many crucial undiscovered genes. Of course, without the proteomic data, transcriptomes tell less than half the story. Nonetheless, it goes without saying that microarrays have already revolutionized our understanding of biology even though they only provide occasional, noisy, unreliable, partial, and occluded snapshots of the transcriptional states of cells.

If one hypothesizes that the number of potential genes involved in cellular processes is relatively large compared to the regulatory elements and their effective combinations responsible for controlling these genes, then the transcriptional state-space should be rather low-dimensional compared to its apparent dimension. As a result, understanding this structure accurately from transcriptome data has many non-trivial implications to functional understanding of the cell. Partitioning genes into closely related groups has thus become the key mathematical first step in practically all statistical analyses of microarray data.

Traditionally, algorithms for cluster analysis of genome-wide expression data from DNA microarray hybridization are based upon statistical properties of gene expressions and result in organizing genes according to similarity in pattern of gene expression. These algorithms display the output graphically, often in a binary tree form, conveying the clustering and the underlying expression data simultaneously. If two genes belong to a cluster (or, equivalently, if they belong to a subtree of small depth) then one may infer a common regulatory mechanism for the two genes or interpret this information as an indication of the status of cellular processes. Furthermore, coexpression of genes of known function with novel genes may lead to a discovery process for characterizing unknown or poorly characterized genes. In general, since false-negatives (where two coexpressed genes are assigned to distinct

clusters) may cause the discovery process to ignore useful information for certain novel genes, and false-positives (where two independent genes are assigned to the same cluster) may result in noise in the information provided to the subsequent algorithms used in analyzing regulatory patterns, it is important that the statistical algorithms for clustering be reasonably robust. Unfortunately, as the microarray experiments that can be carried out in an academic laboratory for a reasonable cost are small in number and suffer from experimental noise, often a statistician must resort to unconventional algorithms to deal with small-sample data.

A popular and one of the earliest clustering algorithms reported in the literature was introduced in [21]. In this paper, the gene-expression data were collected on spotted DNA microarrays [43] and were based upon gene expression in the budding yeast *Saccharomyces cerevisiae* during the diauxic shift [18], the mitotic cell division cycle [47], sporulation [14], and temperature and reducing shocks. In all experiments, RNA from experimental samples (taken at selected times during the process) was labeled during reverse transcription with the red-fluorescent dye Cy5 and was mixed with a reference sample labeled in parallel with the green-fluorescent dye Cy3. After hybridization and appropriate washing steps, separate images were acquired for each fluorophore, and fluorescence intensity ratios were obtained for all target elements. The experimental data were given in an $M \times N$ matrix structure, in which the M rows represented all genes for which data had been collected, the N columns represented individual array experiments (e.g., single time points or conditions), and each entry represented the measured Cy5/Cy3 fluorescence ratio at the corresponding target element on the appropriate array. All ratio values were log transformed to treat inductions and repressions of identical

magnitude as numerically equal but opposite in sign. It was assumed that the raw ratio values followed log-normal distributions and hence, the log-transformed data followed normal distributions. While our mathematical derivations will rely on this assumption for the sake of simplicity, we note that our approach can be generalized in a straightforward manner to deal with other situations where this assumption is violated.

The gene similarity metric employed was a form of correlation coefficient. Let G_i be the (log-transformed) primary data for gene G in condition i. For any two genes X and Y observed over a series of N conditions, the classical similarity score based upon Pearson correlation coefficient is:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right),$$

where

$${\Phi_G}^2 = \sum_{i=1}^{N}\frac{(G_i - G_{offset})^2}{N}$$

and G_{offset} is the estimated mean of the observations, i.e.,

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{i=1}^{N} G_i.$$

Note that Φ_G is simply the (rescaled) estimated standard deviation of the observations. In the analysis presented in [21], "values of G_{offset} which are not the average over observations on G *were* used when there was an assumed unchanged or reference state represented by the value of G_{offset}, against which changes were to be analyzed; in all of the examples presented there, G_{offset} was set to 0, corresponding to a fluorescence ratio of 1.0." To distinguish this modified correlation coefficient from the classical Pearson correlation coefficient, we shall refer to it as Eisen correlation coefficient. Our main innovation is in suggesting a different value for G_{offset},

namely $G_{offset} = \gamma \bar{G}$, where γ is allowed to take a value between 0.0 and 1.0. Note that when $\gamma = 1.0$, we have the classical Pearson correlation coefficient and when $\gamma = 0.0$, we have replaced it by Eisen correlation coefficient. For a non-unit value of γ, the estimator for $G_{offset} = \gamma \bar{G}$ can be thought of as the unbiased estimator $\bar{G}$ being shrunk towards the believed value for $G_{offset} = 0.0$. We address the following questions: What is the optimal value for the shrinkage parameter γ from a Bayesian point of view? How do the gene expression data cluster as the correlation coefficient is modified with this optimal shrinkage parameter?

In order to achieve a consistent comparison, we leave the rest of the algorithms undisturbed. Namely, once the similarity measure has been assumed, we cluster the genes using the same hierarchical clustering algorithm as the one used by Eisen *et al.* Their hierarchical clustering algorithm is based on the centroid-linkage method (referred to as "average-linkage method" of Sokal and Michener [45] in [21]) and computes a binary tree (dendrogram) that assembles all the genes at the leaves of the tree, with each internal node representing possible clusters at different levels. For any set of M genes, an upper-triangular similarity matrix is computed by using a similarity metric of the type described above, which contains similarity scores for all pairs of genes. A node is created joining the most similar pair of genes, and a gene expression profile is computed for the node by averaging observations for the joined genes. The similarity matrix is updated with this new node replacing the two joined elements, and the process is repeated $(M - 1)$ times until only a single element remains. The modified algorithm has been implemented by the authors within the "NYUMAD" microarray database system and can be freely downloaded from: `http://bioinformatics.cat.nyu.edu/nyumad/clustering/`. As each in-

ternal node can be labeled by a value representing the similarity between its two children nodes (i.e., the two elements that were combined to create the internal node), one can create a set of clusters by simply breaking the tree into subtrees by eliminating all the internal nodes with labels below a certain predetermined threshold value. The clusters created in this manner were used to compare the effects of choosing differing similarity measures.

2.2 Model

Recall that a family of correlation coefficients parametrized by $0 \leq \gamma \leq 1$ may be defined as follows:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \tag{2.1}$$

where

$$\begin{aligned} \Phi_G &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G_i - G_{offset}\right)^2} \quad \text{and} \\ G_{offset} &= \gamma\bar{G} \quad \text{for} \quad G \in \{X,Y\} \end{aligned} \tag{2.2}$$

- *Pearson Correlation Coefficient* uses

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{j=1}^{N} G_i \quad \text{for every gene } G, \text{ or} \quad \gamma = 1.$$

- *Eisen et al.* (in [21]) use

$$G_{offset} = 0 \quad \text{for every gene } G, \text{ or} \quad \gamma = 0.$$

- We propose using the general form of equation (2.1) to derive a similarity metric which is dictated by the data and reduces the occurrence of false-positives (relative to the Eisen metric) and false-negatives (relative to the Pearson correlation coefficient).

2.2.1 Motivation and Setup

As mentioned above, the metric used by Eisen *et al.* in [21] had the form of equation (2.1) with G_{offset} set to 0 for every gene G (as a reference state against which to measure the data). Here, we rigorously examine the mathematical validity of setting G_{offset} to 0 arbitrarily. Even if it is initially assumed that each gene G has zero mean, that assumption must be updated when data becomes available. To this end, we derive a correlation coefficient formula which is dictated by the data, and can be justified by a Bayesian argument.

The microarray data is given in the form of the levels of M genes expressed under N experimental conditions. The data can be viewed as

$$\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$$

where $M \gg N$ and $\{X_{ij}\}_{i=1}^{N}$ is the data vector for gene j.

2.2.2 Derivation

We begin by rewriting S in our notation:

$$\begin{aligned} & S(X_j, X_k) \\ = \; & \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k}\right), \end{aligned} \tag{2.3}$$

$$\Phi_j{}^2 = \frac{1}{N}\sum_i \Big(X_{ij} - (X_j)_{\mathit{offset}}\Big)^2$$

In the most general setting, we can make the following assumptions on the data distribution: let all values X_{ij} for gene j have a Normal distribution with mean θ_j and standard deviation β_j (variance $\beta_j{}^2$); i.e.,

$$X_{ij} \sim \mathcal{N}(\theta_j, \beta_j{}^2) \quad \text{for} \quad i = 1, \ldots, N$$

with j fixed ($1 \leq j \leq M$), where θ_j is an unknown parameter (taking different values for different j). To estimate θ_j, it is convenient to assume that θ_j is itself a random variable taking values close to zero:

$$\theta_j \sim \mathcal{N}(0, \tau^2).$$

The assumed distribution aids us in obtaining the estimate of θ_j given in (2.14).

For convenience, let us also assume that the data are range-normalized, so that $\beta_j{}^2 = \beta^2$ for every j. If this assumption does not hold on the given data set, it is easily corrected by scaling each gene vector appropriately. Following common practice, we adjusted the range to scale to an interval of unit length, i.e., its maximum and minimum values differ by 1. Thus,

$$X_{ij} \sim \mathcal{N}(\theta_j, \beta^2) \quad \text{and} \quad \theta_j \sim \mathcal{N}(0, \tau^2).$$

Replacing $(X_j)_{\mathit{offset}}$ in (2.3) by the exact value of the mean θ_j yields a *Clairvoyant* correlation coefficient of X_j and X_k. In reality, since θ_j is itself a random variable, it must be estimated from the data. Therefore, to get an explicit formula for $S(X_j, X_k)$ we must derive estimators $\widehat{\theta_j}$ for all j.

In Pearson correlation coefficient, θ_j is estimated by the vector mean $\overline{X}_{\cdot j}$; Eisen correlation coefficient corresponds to replacing θ_j by 0 for every j, which is equiv-

alent to assuming $\theta_j \sim \mathcal{N}(0,0)$ (i.e., $\tau^2 = 0$.) We propose to find an estimate of θ_j (call it $\widehat{\theta_j}$) that takes into account both the prior assumption and the data.

2.2.3 Estimation of θ_j

First, let us obtain the posterior distribution of θ_j from the prior $\mathcal{N}(0,\tau^2)$ and the data. This derivation can be done either from the Bayesian considerations, or via the James-Stein Shrinkage estimators (see [25], or [23] for a recent review). Here, we discuss the former method.

$N = 1$

Assume initially that $N = 1$, i.e., we have one data point for each gene, and denote the variance by σ^2 for the moment:

$$X_j \sim \mathcal{N}(\theta_j, \sigma^2) \tag{2.4}$$

$$\theta_j \sim \mathcal{N}(0, \tau^2) \tag{2.5}$$

For clarity, we denote the probability density function (pdf) of θ_j by $\pi(\cdot)$ and the pdf of X_j by $f(\cdot)$. It is immediate from (2.4) and (2.5) that

$$\begin{aligned} \pi(\theta_j) &= \frac{1}{\sqrt{2\pi}\tau} \exp\left(-{\theta_j}^2/2\tau^2\right), \\ f(X_j|\theta_j) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(X_j-\theta_j)^2/2\sigma^2\right). \end{aligned}$$

By Bayes' Rule, the joint pdf of X_j and θ_j is given by

$$\begin{aligned} f(X_j,\theta_j) &= f(X_j|\theta_j)\ \pi(\theta_j) \\ &= \frac{1}{2\pi\sigma\tau} \exp\left(-\left[\frac{{\theta_j}^2}{2\tau^2} + \frac{(X_j-\theta_j)^2}{2\sigma^2}\right]\right) \end{aligned} \tag{2.6}$$

Then $f(X_j)$, the marginal pdf of X_j alone is

$$\begin{aligned} f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{\theta=-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left(-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}\right), \end{aligned} \tag{2.7}$$

where the equality in equation (2.7) is written out in Appendix A.2. It follows that the posterior distribution of θ_j, again by Bayes' Theorem, is given by

$$\begin{aligned} \pi(\theta_j|X_j) &= \frac{f(X_j,\theta_j)}{f(X_j)} \\ &= \frac{f(X_j|\theta_j)\ \pi(\theta_j)}{f(X_j)} \quad \text{by (2.6)} \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}} \exp\left[-\frac{\left(\theta_j - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right]. \end{aligned} \tag{2.8}$$

(See Appendix A.3 for derivation of (2.8).)

Since this has Normal form, we can read off the mean and variance

$$\begin{aligned} \mathbf{E}(\theta_j|X_j) &= \frac{\tau^2}{\sigma^2+\tau^2}X_j \\ &= \left(1-\frac{\sigma^2}{\sigma^2+\tau^2}\right)X_j, \\ Var(\theta_j|X_j) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \end{aligned} \tag{2.9}$$

We can estimate θ_j by its mean.

N arbitrary

Now, if $N > 1$ is arbitrary, X_j becomes a vector $X_{\cdot j}$. It can be easily shown by using likelihood functions that the vector of values $\{X_{ij}\}_{i=1}^N$, with $X_{ij} \sim \mathcal{N}(\theta_j, \beta^2)$, can be treated as a single data point $Y_j = \overline{X}_{\cdot j} = \sum_{i=1}^N X_{ij}/N$ from the distribution $\mathcal{N}(\theta_j, \beta^2/N)$ (see Appendix A.4).

Thus, following the above derivation with $\sigma^2 = \beta^2/N$, we have a Bayesian estimator for θ_j given by $\mathbf{E}(\theta_j|X_{\cdot j})$:

$$\widehat{\theta_j} = \left(1 - \frac{\beta^2/N}{\beta^2/N + \tau^2}\right) Y_j. \tag{2.10}$$

Unfortunately, (2.10) cannot be used in (2.3) directly, because τ^2 and β^2 are unknown, so must be estimated from the data.

Estimation of $1/\left(\beta^2/N + \tau^2\right)$

Let

$$W = \frac{M-2}{\sum_{j=1}^{M} {Y_j}^2}. \tag{2.11}$$

The form of W comes from James-Stein estimation ([25]), but its derivation will not be discussed here; instead we treat it as an educated guess and verify that it is indeed an appropriate estimator for $1/\left(\beta^2/N + \tau^2\right)$.

$$\begin{aligned} Y_j &\sim \theta_j + \frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \tau^2\mathcal{N}(0,1) + \frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \left(\frac{\beta^2}{N} + \tau^2\right)\mathcal{N}(0,1) \sim \mathcal{N}\left(0, \frac{\beta^2}{N} + \tau^2\right) \end{aligned} \tag{2.12}$$

The transition in (2.12) is justified in Appendix A.5. Let $\alpha^2 = \beta^2/N + \tau^2$. Then from (2.12) it follows that

$$\frac{Y_j}{\sqrt{\alpha^2}} = \frac{Y_j}{\alpha} \sim \mathcal{N}(0,1),$$

and hence

$$\sum_{j=1}^{M} {Y_j}^2 = \alpha^2 \sum_{j=1}^{M} \left(\frac{Y_j}{\alpha}\right)^2 = \alpha^2 \chi^2_M,$$

where χ^2_M is a Chi-square random variable with M degrees of freedom. By properties of the Chi-square distribution and the linearity of expectation,

$$\begin{aligned}\mathbf{E}\left(\frac{\alpha^2}{\sum {Y_j}^2}\right) &= \frac{1}{M-2} \quad \text{(see Appendix A.6)} \\ \mathbf{E}(W) &= \mathbf{E}\left(\frac{M-2}{\sum {Y_j}^2}\right) = \frac{1}{\alpha^2} = \frac{1}{\frac{\beta^2}{N}+\tau^2}\end{aligned}$$

Thus, W is an unbiased estimator of $1/\left(\beta^2/N+\tau^2\right)$, and can be used to replace $1/\left(\beta^2/N+\tau^2\right)$ in (2.10).

Estimation of β^2

It can be shown (see Appendix A.7) that

$$ {S_j}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_{ij}-Y_j\right)^2 $$

is an unbiased estimator for β^2 based solely on data from gene j, and that $\frac{N-1}{\beta^2}{S_j}^2$ has Chi-square distribution with $(N-1)$ degrees of freedom. Since this holds for every j, we can get a more accurate estimate for β^2 by pooling all available data, i.e., by averaging the estimates for each j:

$$\begin{aligned}\widehat{\beta^2} &= \frac{1}{M}\sum_{j=1}^{M}{S_j}^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij}-Y_j)^2\right) \\ &= \frac{1}{M(N-1)}\sum_{j=1}^{M}\sum_{i=1}^{N}(X_{ij}-Y_j)^2. \qquad (2.13)\end{aligned}$$

$\widehat{\beta^2}$ is an unbiased estimator for β^2, since

$$\begin{aligned}\mathbf{E}(\widehat{\beta^2}) &= \mathbf{E}\left(\frac{1}{M}\sum_{j=1}^{M}{S_j}^2\right) \\ &= \frac{1}{M}\sum_{j=1}^{M}\mathbf{E}({S_j}^2) = \frac{1}{M}\sum_{j=1}^{M}\beta^2 = \beta^2.\end{aligned}$$

Substituting the estimates (2.11) and (2.13) into (2.10), we obtain the explicit estimate for θ_j:

$$\begin{aligned}
&\widehat{\theta_j} \\
&= \left(1 - \widehat{\frac{1}{\frac{\beta^2}{N} + \tau^2}} \frac{\widehat{\beta^2}}{N}\right) Y_j \\
&= \left(1 - W \cdot \frac{\widehat{\beta^2}}{N}\right) Y_j \\
&= \left(1 - \left(\frac{M-2}{\sum_{k=1}^{M} {Y_k}^2}\right) \cdot \frac{1}{N} \cdot \frac{1}{M(N-1)} \sum_{k=1}^{M} \sum_{i=1}^{N} (X_{ik} - Y_k)^2\right) Y_j \\
&= \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M} \sum_{i=1}^{N} (X_{ik} - Y_k)^2}{\sum_{k=1}^{M} {Y_k}^2}\right)}_{\gamma} Y_j \qquad (2.14) \\
&= \gamma \overline{X}_{\cdot j}
\end{aligned}$$

Finally, we can substitute $\widehat{\theta_j}$ from equation (2.14) into the correlation coefficient in (2.3) wherever $(X_j)_{\mathit{offset}}$ appears to obtain an explicit formula for $S(X_{\cdot j}, X_{\cdot k})$.

2.3 Algorithm & Implementation

The implementation of hierarchical clustering proceeds in a greedy manner, always choosing the most similar pair of elements (starting with genes at the bottom-most level) and combining them to create a new element. The "expression vector" for the new element is simply the weighted average of the expression vectors of the two most similar elements that were combined. This structure of repeated pair-wise combinations is conveniently represented in a binary tree, whose leaves are the set

of genes and internal nodes are the elements constructed from the two children nodes. The algorithm is described below in pseudocode.

2.3.1 Hierarchical clustering pseudocode

Given $\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$:

`Switch:`

`Pearson:` $\gamma = 1$;

`Eisen:` $\gamma = 0$;

`Shrinkage:` {

`Compute` $W = (M-2) \Big/ \sum_{j=1}^{M} \overline{X}_{\cdot j}{}^2$

`Compute` $\widehat{\beta^2} = \sum_{j=1}^{M} \sum_{i=1}^{N} \left(X_{ij} - \overline{X}_{\cdot j}\right)^2 \Big/ \left(M(N-1)\right)$

$\gamma = 1 - W \cdot \widehat{\beta^2}/N$

}

`While (# clusters > 1) do`

- ◇ `Compute similarity table:`

  $S(G_j, G_k) = \frac{\sum_i \left(G_{ij} - (G_j)_{offset}\right)\left(G_{ik} - (G_k)_{offset}\right)}{\sqrt{\sum_i \left(G_{ij} - (G_j)_{offset}\right)^2 \cdot \sum_i \left(G_{ik} - (G_k)_{offset}\right)^2}}$,

  `where` $(G_\ell)_{offset} = \gamma \overline{G_\ell}$.

- ◇ `Find` (j^*, k^*) :

  $S(G_{j^*}, G_{k^*}) \geq S(G_j, G_k)$ $\forall$ `clusters` j, k

- ◇ `Create new cluster` $N_{j^*k^*}$

  = `weighted average of` G_{j^*} `and` G_{k^*}.

◇ Take out clusters j^* and k^*.

The implementation of generalized hierarchical clustering with options to choose different similarity measures has been incorporated into NYUMAD (NYU MicroArray Database), an integrated system to maintain and analyze biological abundance data along with associated experimental conditions and protocols. While the initial goal was to provide a system to manage microarray data, the system has been designed to store any type of abundance data, including protein levels. This system uses a relational database management system for the storage of data and has a flexible database schema that stores abundance data along with general research data such as experimental conditions and protocols. The database schema is defined using standard SQL (Structured Query Language) and is therefore portable to any SQL database platform. To enable widespread utility, NYUMAD supports the MAGE-ML standard ([46]) for the exchange of gene expression data, defined by the Microarray Gene Expression Data Group (MGED)—web site at `http://www.mged.org/`.

There are several ways to access the system: using the NYUMAD Java application, through web pages, or through custom applications (for details, see `http://bioinformatics.cat.nyu.edu/nyumad/`). Data transfer is affected using the world wide web (WWW) with the HTTP protocol. The use of the WWW for communication ensures accessibility from any location.

The graphical user interface (GUI) provided by the Java application facilitates easy data submission, retrieval, and analysis. The Java application presents data in a logical manner and allows easy navigation through the data. The GUI also

allows straightforward updating of existing data and insertion of new data.

NYUMAD supports collaborative research efforts by allowing groups to submit data from any location (via HTTP) and to view, retrieve, or analyze each other's data immediately. Groups can share protocols and divide a large project covering a wide range of experimental conditions into sub-projects performed by individual groups.

NYUMAD is a secure repository for both public and private data. Users can control the visibility of their data so that initially the data might be private but after the publication of the results, the data can be marked public and made visible to the larger research community. Public users can log in with a general login ID without the need for a password and view and retrieve any of the public data.

The system provides a wide range of data analysis and interpretation tools and algorithms that help in identifying patterns and relationships. A general feature of NYUMAD is the flexibility for users to build their own queries and utilize their own parameters, data transformations, and filters where appropriate. Users can retrieve queried data for input to their own tools or use other tools within NYUMAD—for example, perform a clustering of their microarray data or determine the statistical significance of differential expression values for a specific set of genes. Data analysis tools are supplemented with visualization tools.

2.4 Results

2.4.1 Mathematical Simulation

To compare the performance of these algorithms, we started with a relatively simple *in silico* experiment. In such an experiment, one can create two genes X and Y and simulate N (about 100) experiments as follows:

$$\begin{aligned} X_i &= \theta_X + \sigma_X(\alpha_i(X,Y) + \mathcal{N}(0,1)), \text{ and} \\ Y_i &= \theta_Y + \sigma_Y(\alpha_i(X,Y) + \mathcal{N}(0,1)), \end{aligned}$$

where α_i, chosen from a uniform distribution over a range $[L,H]$ ($\mathcal{U}(L,H)$), is a "bias term" introducing a correlation (or none if all α's are zero) between X and Y. $\theta_X \sim \mathcal{N}(0,\tau^2)$ and $\theta_Y \sim \mathcal{N}(0,\tau^2)$ are the means of X and Y, respectively. Similarly, σ_X and σ_Y are the standard deviations for X and Y, respectively.

Note that, with this model

$$\begin{aligned} S(X,Y) &= \frac{1}{N}\sum_{i=1}^{N} \frac{(X_i-\theta_X)}{\sigma_X}\frac{(Y_i-\theta_Y)}{\sigma_Y} \\ &\sim \frac{1}{N}\sum_{i=1}^{N}(\alpha_i+\mathcal{N}(0,1))(\alpha_i+\mathcal{N}(0,1)) \\ &\sim \frac{1}{N}\left[\left(\sum_{i=1}^{N}\alpha_i^2\right)+\chi^2_N+2\mathcal{N}(0,1)\sum_{i=1}^{N}\alpha_i\right] \end{aligned}$$

if the exact values of the mean and variance are used.

We denote the distribution of S by $\mathcal{F}(\mu,\delta)$, where μ is the mean and δ is the standard deviation.

The model was implemented in Mathematica [48]; the following parameters were used in the simulation: $N = 100$, $\tau \in \{0.1, 10.0\}$ (representing very low or

high variability among the genes), $\sigma_X = \sigma_Y = 10.0$, and $\alpha = 0$ representing no correlation between the genes or $\alpha \sim \mathcal{U}(0,1)$ representing some correlation between the genes. Once the parameters were fixed for a particular *in silico* experiment, the gene-expression vectors for X and Y were generated many thousand times, and for each pair of vectors $S_c(X,Y)$, $S_p(X,Y)$, $S_e(X,Y)$, and $S_s(X,Y)$ were estimated by four different algorithms and further examined to see how the estimators of S varied over these trials. These four different algorithms estimated S according to equations (2.1), (2.2) as follows: *Clairvoyant* estimated S_c using the true values of θ_X, θ_Y, σ_X, and σ_Y; *Pearson* estimated S_p using the unbiased estimators $\bar{X}$ and $\bar{Y}$ of θ_X and θ_Y (for X_{offset} and Y_{offset}), respectively; *Eisen* estimated S_e using the value 0.0 as the estimator of both θ_X and θ_Y; and *Shrinkage* estimated S_s using the shrunk biased estimators $\widehat{\theta}_X$ and $\widehat{\theta}_Y$ of θ_X and θ_Y, respectively. In the latter three, the standard deviation was estimated as in (2.2). The histograms corresponding to these *in silico* experiments can be found in Figure 2.1. Our observations can be summarized as follows:

- When X and Y are not correlated and the noise in the input is low ($N = 100$, $\tau = 0.1$, and $\alpha = 0$), Pearson does just as well as Eisen, Shrinkage, or Clairvoyant:
  $S_c \sim \mathcal{F}(-0.000297, 0.0996)$, $S_p \sim \mathcal{F}(-0.000269, 0.0999)$,
  $S_e \sim \mathcal{F}(-0.000254, 0.0994)$, and $S_s \sim \mathcal{F}(-0.000254, 0.0994)$.

- When X and Y are not correlated but the noise in the input is high ($N = 100$, $\tau = 10.0$, and $\alpha = 0$), Pearson does just as well as Shrinkage or Clairvoyant, but Eisen introduces far too many false-positives:

$S_c \sim \mathcal{F}(-0.000971, 0.0994)$, $S_p \sim \mathcal{F}(-0.000939, 0.100)$,
$S_e \sim \mathcal{F}(-0.00119, 0.354)$, and $S_s \sim \mathcal{F}(-0.000939, 0.100)$.

- When X and Y are correlated and the noise in the input is low ($N = 100$, $\tau = 0.1$, and $\alpha \sim \mathcal{U}(0, 1)$), Pearson does much more poorly compared to Eisen, Shrinkage, or Clairvoyant—these three doing equally well; Pearson introduces too many false-negatives:
  $S_c \sim \mathcal{F}(0.331, 0.132)$, $S_p \sim \mathcal{F}(0.0755, 0.0992)$,
  $S_e \sim \mathcal{F}(0.248, 0.0915)$, and $S_s \sim \mathcal{F}(0.245, 0.0915)$.

- Finally, when X and Y are correlated and the noise in the input is high, the signal-to-noise ratio becomes extremely poor and all the algorithms fail, i.e., introduce errors:
  $S_c \sim \mathcal{F}(0.333, 0.133)$, $S_p \sim \mathcal{F}(0.0762, 0.100)$,
  $S_e \sim \mathcal{F}(0.117, 0.368)$, and $S_s \sim \mathcal{F}(0.0762, 0.0999)$.

In summary, one can conclude that for the same clustering algorithm, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage. Shrinkage, on the other hand, reduces these errors by combining the good properties of both algorithms.

2.4.2 Biological Example

We then proceeded to test the algorithms on a biological example. We chose a biologically well-characterized system, and analyzed the clusters of genes involved in the yeast cell cycle. These clusters were computed using the hierarchical clustering algorithm with the underlying similarity measure chosen from the following three:

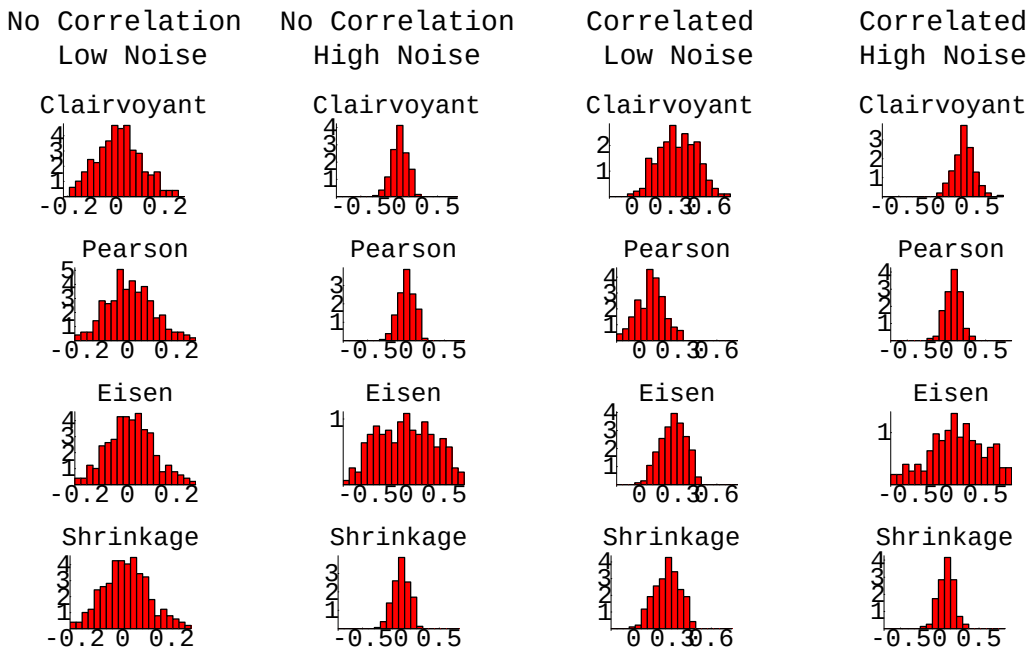


Figure 2.1: Histograms.

Pearson, Eisen, or Shrinkage. As a reference, the computed clusters were compared to the ones implied by the common cell-cycle functions and regulatory systems inferred from the roles of various transcriptional activators (see Figure 2.2).

Note that our experimental analysis is based on the assumption that the groupings suggested by the ChIP (Chromatin ImmunoPrecipitation) analysis are, in fact, correct and thus, provide a direct approach to compare various correlation coefficients. It is quite likely that the ChIP-based groupings themselves contain many false relations (both positives and negatives) and corrupt our inference in some unknown manner. Nonetheless, we observe that the trends of reduced false positives and negatives in shrinkage analysis with these biological data are consistent with the analysis based on mathematical simulation and hence, reassuring.

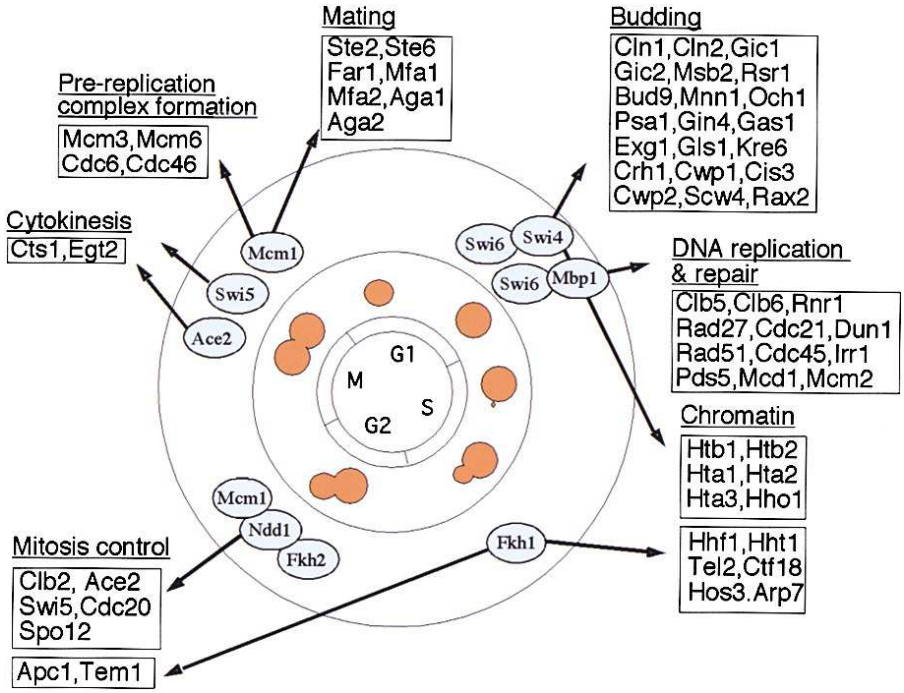


Figure 2.2: Regulation of cell cycle functions by the activators. [Reproduced with permission from [44] (Copyright 2001, Elsevier)].

In the work of Simon *et al.* ([44]), genome-wide location analysis was used to determine how the yeast cell cycle gene expression program is regulated by each of the nine known cell cycle transcriptional activators: Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Ndd1, Swi4, Swi5, and Swi6. It was also found that cell cycle transcriptional activators which function during one stage of the cell cycle regulate transcriptional activators that function during the next stage. This serial regulation of transcriptional activators together with various functional properties suggests a simple way of partitioning some selected cell cycle genes into nine clusters, each one characterized by a group of transcriptional activators working together and their functions

(see Table 2.1): for instance, Group 1 is characterized by the activators Swi4 and Swi6 and the function of budding; Group 2 is characterized by the activators Swi6 and Mbp1 and the function involving DNA replication and repair at the juncture of G1 and S phases, etc.

Table 2.1: Genes in our data set, grouped by transcriptional activators and cell-cycle functions.

	Activators	Genes	Functions
1	Swi4, Swi6	Cln1, Cln2, Gic1, Gic2, Msb2, Rsr1, Bud9, Mnn1, Och1, Exg1, Kre6, Cwp1	Budding
2	Swi6, Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2	DNA replication and repair
3	Swi4, Swi6	Htb1, Htb2, Hta1, Hta2, Hta3, Hho1	Chromatin
4	Fkh1	Hhf1, Hht1, Tel2, Arp7	Chromatin
5	Fkh1	Tem1	Mitosis Control
6	Ndd1, Fkh2, Mcm1	Clb2, Ace2, Swi5, Cdc20	Mitosis Control
7	Ace2, Swi5	Cts1, Egt2	Cytokinesis
8	Mcm1	Mcm3, Mcm6, Cdc6, Cdc46	Pre-replication complex formation
9	Mcm1	Ste2, Far1	Mating

Our initial hypothesis can be summarized as follows: *Genes expressed during the same cell cycle stage, and regulated by the same transcriptional activators should be in the same cluster.* Below we list some of the deviations from the hypothesis observed in the raw data.

Possible False-Positives:

- Bud9 (Group 1: Budding) and {Cts1, Egt2} (Group 7: Cytokinesis) are placed in the same cluster by all three metrics: P49 = S82 $\simeq$ E47; however, the Eisen metric also places Exg1 (Group 1) and Cdc6 (Group 8: Pre-replication complex formation) in the same cluster.

- Mcm2 (Group 2: DNA replication and repair) and Mcm3 (Group 8) are placed in the same cluster by all three metrics: P10 = S20 $\simeq$ E73; however, the Eisen metric places several more genes from different groups in the same cluster: {Rnr1, Rad27, Cdc21, Dun1, Cdc45} (Group 2), Hta3 (Group 3: Chromatin), and Mcm6 (Group 8) are also placed in cluster E73.

Possible False-Negatives:

- Group 1: Budding (Table 2.1) is split into four clusters by the Eisen metric: {Cln1, Cln2, Gic2, Rsr1, Mnn1} $\in$ Cluster *a* (E39), Gic2 $\in$ Cluster *b* (E62), {Bud9, Exg1} $\in$ Cluster *c* (E47), and {Kre6, Cwp1} $\in$ Cluster *d* (E66);
  and into six clusters by both the Shrinkage and Pearson metrics:
  {Cln1, Cln2, Gic2, Rsr1, Mnn1} $\in$ Cluster *a* (S3=P66), {Gic1, Kre6} $\in$ Cluster *b* (S39=P17), Msb2 $\in$ Cluster *c* (S24=P71), Bud9 $\in$ Cluster *d* (S82=P49), Exg1 $\in$ Cluster *e* (S48=P78), and Cwp1 $\in$ Cluster *f* (S8=P4).

Table 2.1 contains those genes from Figure 2.2 that were present in our data set. The following tables contain these genes grouped into clusters by a hierarchical clustering algorithm using the three metrics (Eisen in Table 2.2, Pearson in Table 2.3, and Shrinkage in Table 2.4) thresholded at a correlation coefficient value

of 0.60. The choice of the threshold parameter is discussed further in section 2.5. Genes that have not been grouped with any others at a similarity of 0.60 or higher are absent from the tables; in the subsequent analysis they are treated as *singleton* clusters.

The value $\gamma \simeq 0.89$ estimated from the raw yeast data was surprisingly high, contrary to the suggestion in [21] that the value $\gamma = 0$ performed better than $\gamma = 1$. It also did not yield as great an improvement in the yeast data clusters as the simulations indicated. This suggested that the true value of γ is closer to 0. Upon closer examination of the data, we observed that the data in its raw "pre-normalized" form is inconsistent with the assumptions used in deriving γ:

1. The gene vectors are not range-normalized, so ${\beta_j}^2 \neq \beta^2$ for every j, and
2. The N experiments are not necessarily independent.

2.4.3 Corrections

We attempted to remedy the first flaw by normalizing all gene vectors with respect to range (dividing each entry in gene X by $(X_{\max} - X_{\min})$), recomputing the estimated γ value, and repeating the clustering process. As normalized gene expression data yielded the estimate $\gamma \simeq 0.91$, still too high a value, we conducted an extensive computational experiment to determine the best empirical γ value by also clustering with the shrinkage factors of 0.2, 0.4, 0.6, and 0.8. The clusters taken at the correlation factor cut-off of 0.60, as above, are presented in Tables 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, and 2.11.

To compare the resulting sets of clusters, we introduced the following notation.

Write each cluster set as follows:

$$\Big\{x \to \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\Big\}_{x=1}^{\#\text{ of groups}}$$

where x denotes the group number (as described in Table 2.1), n_x is the number of clusters group x appears in, and for each cluster $j \in \{1, \ldots, n_x\}$ there are y_j genes from group x and z_j genes from other groups in Table 2.1. A value of "$*$" for z_j denotes that cluster j contains additional genes, although none of them are cell cycle genes; in subsequent computations, this value is treated as 0.

This notation naturally lends itself to a scoring function for measuring the number of false-positives, number of false-negatives, and total error score, which aids in the comparison of cluster sets.

$$\text{FP}(\gamma) = \frac{1}{2}\sum_{x}\sum_{j=1}^{n_x} y_j \cdot z_j \tag{2.15}$$

$$\text{FN}(\gamma) = \sum_{x}\sum_{1 \le j < k \le n_x} y_j \cdot y_k \tag{2.16}$$

$$\text{Error\_score}(\gamma) = \text{FP}(\gamma) + \text{FN}(\gamma) \tag{2.17}$$

In this notation, the cluster sets with their error scores can be listed as follows:

$$\gamma = 0.0(E) \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{3,*\},\{2,13\},\{1,*\},\{1,*\},\\
&\qquad \{1,*\},\{1,4\},\{1,0\},\{1,0\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,7\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,14\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,14\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,3\}\},\\
6 \;&\rightarrow\; \{\{3,1\},\{1,14\}\},\\
7 \;&\rightarrow\; \{\{2,3\}\},\\
8 \;&\rightarrow\; \{\{2,13\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{2,3\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.0) = 97 + 88 = 185$$

$$\gamma = 0.2 \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,7\},\{1,*\},\{1,*\},\\
&\qquad \{1,1\},\{1,2\},\{1,0\},\{1,0\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{7,1\},\{1,5\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,5\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,5\},\{1,1\}\},\\
5 \;&\rightarrow\; \{\{1,3\}\},\\
6 \;&\rightarrow\; \{\{3,1\},\{1,5\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,4\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.2) = 38 + 94 = 132$$

$$\gamma = 0.4 \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\qquad \{2,*\},\{1,2\},\{1,0\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,6\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,13\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,3\}\},\\
6 \;&\rightarrow\; \{\{3,1\},\{1,13\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,12\},\{1,*\},\{1,1\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\}
\end{aligned}$$

$$\text{Error\_score}(0.4) = 78 + 86 = 164$$

$$\gamma = 0.6 \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\qquad \{2,*\},\{1,2\},\{1,0\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,6\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,13\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,0\}\},\\
6 \;&\rightarrow\; \{\{3,*\},\{1,13\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.6) = 75 + 86 = 161$$

$$\gamma = 0.8 \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\qquad \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,6\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,13\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,0\}\},\\
6 \;&\rightarrow\; \{\{3,*\},\{1,13\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.8) = 75 + 86 = 161$$

$$\gamma = 0.91(S) \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,13\}\{1,*\},\{1,*\},\\
&\qquad \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,6\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,13\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,0\}\},\\
6 \;&\rightarrow\; \{\{3,*\},\{1,13\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.91) = 75 + 86 = 161$$

$$\gamma = 1.0(P) \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\qquad \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\
2 \;&\rightarrow\; \{\{8,6\},\{1,1\}\},\\
3 \;&\rightarrow\; \{\{5,2\},\{1,13\}\},\\
4 \;&\rightarrow\; \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 \;&\rightarrow\; \{\{1,0\}\},\\
6 \;&\rightarrow\; \{\{3,*\},\{1,13\}\},\\
7 \;&\rightarrow\; \{\{2,1\}\},\\
8 \;&\rightarrow\; \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(1.0) = 75 + 86 = 161$$

Clearly, in this notation, γ values of 0.8, 0.91, and 1.0 give identical cluster groupings, and the best error score is attained at $\gamma = 0.2$.

To improve the estimated value of γ, we proceeded to correct the second flaw due to the statistical dependence among the experiments. We sought to reduce the effective number of experiments by subsampling from the set of all (possibly correlated) experiments—the candidates were chosen via clustering all the experiments, i.e., columns of the data matrix, and then selecting one representative experiment from each cluster of experiments. We then clustered the subsampled data, once again using the cut-off correlation value of 0.60. The resulting cluster sets under

the Eisen, Shrinkage, and Pearson metrics are given in Tables 2.12, 2.13, and 2.14, respectively.

The subsampled data yielded the lower estimated value $\gamma \simeq 0.66$. In our set notation, the resulting clusters with the corresponding error scores can be written as follows:

$$\gamma = 0.0(E) \Longrightarrow$$

$$\begin{aligned}
\{1 &\rightarrow \{\{6,23\},\{2,*\},\{2,5\},\{1,*\},\{1,*\}\},\\
2 &\rightarrow \{\{7,22\},\{2,5\}\},\\
3 &\rightarrow \{\{5,24\},\{1,6\}\},\\
4 &\rightarrow \{\{3,26\},\{1,*\}\},\\
5 &\rightarrow \{\{1,28\}\},\\
6 &\rightarrow \{\{3,26\},\{1,6\}\},\\
7 &\rightarrow \{\{1,*\},\{1,28\}\},\\
8 &\rightarrow \{\{3,26\},\{1,6\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.0) = 370 + 79 = 449$$

$$\gamma = 0.66(S) \Longrightarrow$$

$$\begin{aligned}
\{1 &\rightarrow \{\{6,6\},\{3,2\},\{2,5\},\{1,*\}\},\\
2 &\rightarrow \{\{6,6\},\{2,5\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,*\}\},\\
4 &\rightarrow \{\{2,5\},\{1,3\},\{1,6\}\},\\
5 &\rightarrow \{\{1,*\}\},\\
6 &\rightarrow \{\{3,1\},\{1,6\}\},\\
7 &\rightarrow \{\{1,*\},\{1,4\}\},\\
8 &\rightarrow \{\{1,*\},\{1,1\},\{1,4\},\{1,6\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.66) = 76 + 88 = 164$$

$$\gamma = 1.0(P) \Longrightarrow$$

$$\begin{aligned}
\{1 \;&\rightarrow\; \{\{3,6\},\{2,*\},\{2,1\},\{1,*\},\\
&\qquad \{1,*\},\{1,*\},\{1,5\},\{1,5\}\},\\
2 \;&\rightarrow\; \{\{5,4\},\{2,4\},\{1,2\},\{1,7\}\},\\
3 \;&\rightarrow\; \{\{5,3\},\{1,5\}\},\\
4 \;&\rightarrow\; \{\{2,6\},\{1,*\},\{1,1\}\},\\
5 \;&\rightarrow\; \{\{1,*\}\},\\
6 \;&\rightarrow\; \{\{3,3\},\{1,5\}\},\\
7 \;&\rightarrow\; \{\{1,*\},\{1,5\}\},\\
8 \;&\rightarrow\; \{\{1,1\},\{1,5\},\{1,5\},\{1,8\}\},\\
9 \;&\rightarrow\; \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(1.0) = 69 + 107 = 176$$

From the tables for the range-normalized, subsampled yeast data, as well as by comparing the error scores, one can conclude that for the same clustering algorithm and threshold value, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage, as Shrinkage reduces these errors by combining the good properties of both algorithms. This observation is consistent with our mathematical analysis and the simulation presented in section 2.4.1.

2.5 Discussion

Microarray-based genomic analysis and other similar high-throughput methods have begun to occupy an increasingly important role in biology, as they have helped to create a visual image of the state-space trajectories at the core of the cellular processes. This analysis will address directly to the observational nature of the "new" biology. As a result, we need to develop our ability to "see," accurately and reproducibly, the information in the massive amount of quantitative measurements produced by these approaches—or be able to ascertain when what we "see" is unreliable and forms a poor basis for proposing novel hypotheses. Our investigation demonstrates the fragility of many of these analysis algorithms when used in the context of a small number of experiments. In particular, we see that a small perturbation of, or a small error in, the estimation of a parameter (the shrinkage parameter) has a significant effect on the overall conclusion. The errors in the estimators manifest themselves by missing certain biological relations between two genes (false-negatives) or by proposing phantom relations between two otherwise unrelated genes (false-positives).

A global picture of these interactions can be seen in Figure 2.3, the Receiver Operator Characteristic (ROC) figure, with each curve parametrized by the cut-off threshold in the range of $[-1, 1]$. An ROC curve ([20]) for a given metric plots sensitivity against (1−specificity), where

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)},$$

Specificity = fraction of negatives detected by a metric

$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)},$$

and $\mathrm{TP}(\gamma)$, $\mathrm{FN}(\gamma)$, $\mathrm{FP}(\gamma)$, and $\mathrm{TN}(\gamma)$ denote the number of True Positives, False Negatives, False Positives, and True Negatives, respectively, arising from a metric associated with a given γ. (Recall that γ is 0.0 for Eisen, 1.0 for Pearson, and is computed according to (2.14) for Shrinkage, which yields 0.66 on this data set.) For each pair of genes, $\{j, k\}$, we define these events using our hypothesis (see section 2.4.2) as a measure of truth:

TP: $\{j, k\}$ are in the same group (see Table 2.1) and $\{j, k\}$ are placed in the same cluster;

FP: $\{j, k\}$ are in different groups, but $\{j, k\}$ are placed in the same cluster;

TN: $\{j, k\}$ are in different groups and $\{j, k\}$ are placed in different clusters; and

FN: $\{j, k\}$ are in the same group, but $\{j, k\}$ are placed in different clusters.

$\mathrm{FP}(\gamma)$ and $\mathrm{FN}(\gamma)$ were already defined in equations (2.15) and (2.16), respectively, and we define

$$\mathrm{TP}(\gamma) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2} \tag{2.18}$$

and

$$\mathrm{TN}(\gamma) = \mathrm{Total} - (\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma)) \tag{2.19}$$

where Total= $\binom{44}{2} = 946$ is the total # of gene pairs $\{j, k\}$ in Table 2.1.

The ROC figure suggests the best threshold to use for each metric, and can also be used to select the best metric to use for a particular sensitivity.

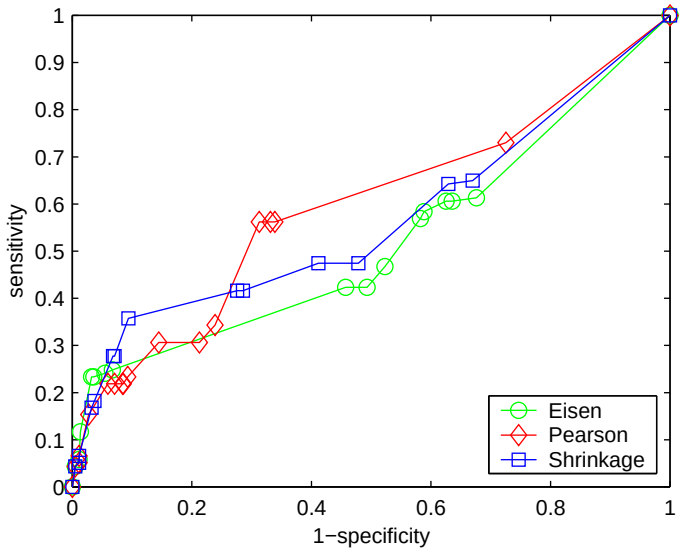


Figure 2.3: Receiver Operator Characteristic curves. Each curve is parametrized by the cut-off value $\theta \in \{1.0, 0.95, \ldots, -1.0\}$

The dependence of the error scores on the threshold can be more clearly seen from Figure 2.4. It shows that the conclusions we draw in section 2.4.3 hold for a wide range of threshold values, and hence a threshold value of 0.60 is a reasonable representative value.

As a result, in order to study the clustering algorithms and their effectiveness, one may ask the following questions. If one must err, is it better to err on the side of more false-positives or more false-negatives? What are the relative costs of these two kinds of errors? In general, since false-negatives may cause the inference process to ignore useful information for certain novel genes, and since false-positives may result in noise in the information provided to the algorithms used in analyzing regulatory patterns, intelligent answers to our questions depend crucially on how

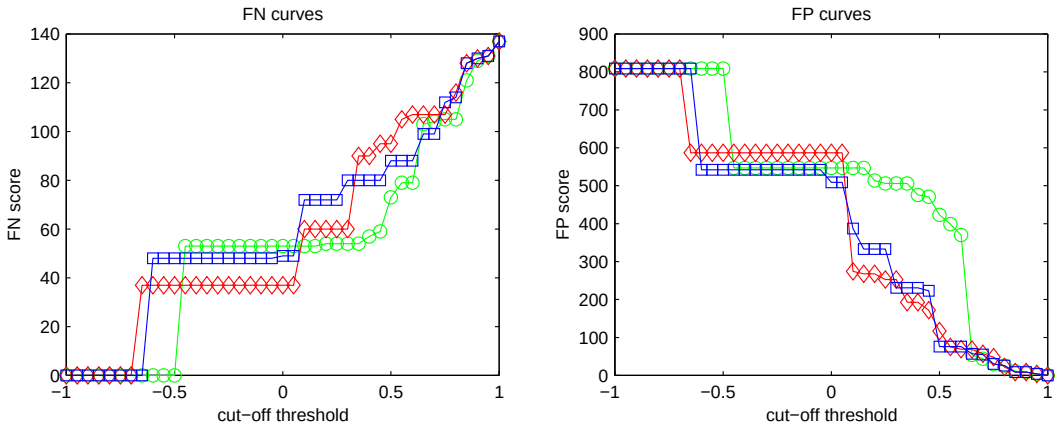


Figure 2.4: FN and FP curves, plotted as functions of θ.

the cluster information is used in the subsequent discovery processes.

Table 2.2: Eisen Clusters

E39	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
E62	Swi4/Swi6	Gic1
E47	Swi4/Swi6	Bud9, Exg1
	Ace2/Swi5	Cts1, Egt2
	Mcm1	Cdc6
E66	Swi4/Swi6	Kre6, Cwp1
E71	Swi6/Mbp1	Clb5, Clb6, Rad51
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
E73	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Cdc45, Mcm2
	Swi4/Swi6	Hta3
	Mcm1	Mcm3, Mcm6
E63	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
E32	Fkh1	Arp7
E38	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
E51	Mcm1	Ste2, Far1

Table 2.3: Pearson Clusters

P66	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
P17	Swi4/Swi6	Gic1, Kre6
P71	Swi4/Swi6	Msb2
P49	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
P78	Swi4/Swi6	Exg1
P4	Swi4/Swi6	Cwp1
P12	Swi6/Mbp1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
P10	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
P54	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P37	Fkh1	Arp7
P16	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P50	Mcm1	Ste2, Far1

Table 2.4: Shrinkage Clusters

S3	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
S39	Swi4/Swi6	Gic1, Kre6
S24	Swi4/Swi6	Msb2
S82	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S48	Swi4/Swi6	Exg1
S8	Swi4/Swi6	Cwp1
S14	Swi6/Mbp1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S20	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S4	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S13	Swi4/Swi6	Hta3
S63	Fkh1	Arp7
S22	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S83	Mcm1	Ste2, Far1

Table 2.5: RN Data, $\gamma = 0.0$ (Eisen Clusters)

E8	Swi4/Swi6	Cln1, Msb2, Mnn1
E71	Swi4/Swi6	Cln2, Rsr1
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
E14	Swi4/Swi6	Gic1
E17	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
	Mcm1	Ste2, Far1
E16	Swi4/Swi6	Exg1
E59	Swi4/Swi6	Kre6
E18	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
E86	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
E10	Fkh1	Arp7
E19	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
E11	Mcm1	Cdc6

Table 2.6: Range-normalized data, $\gamma = 0.2$

$S_{0.2}59$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.2}26$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
$S_{0.2}23$	Swi4/Swi6	Gic1
$S_{0.2}58$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.2}57$	Swi4/Swi6	Exg1
	Fkh1	Arp7
$S_{0.2}61$	Swi4/Swi6	Kre6
$S_{0.2}18$	Swi6/Mbp1	Clb5
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.2}28$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.2}25$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.2}29$	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.2}4$	Mcm1	Ste2
$S_{0.2}55$	Mcm1	Far1

Table 2.7: Range-normalized data, $\gamma = 0.4$

$S_{0.4}64$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.4}13$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.4}44$	Swi4/Swi6	Gic1, Kre6
$S_{0.4}27$	Swi4/Swi6	Msb2
$S_{0.4}46$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.4}73$	Swi4/Swi6	Exg1
$S_{0.4}2$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.4}48$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.4}26$	Fkh1	Arp7
$S_{0.4}25$	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.4}16$	Mcm1	Cdc6
$S_{0.4}47$	Mcm1	Ste2
$S_{0.4}58$	Mcm1	Far1

Table 2.8: Range-normalized data, $\gamma = 0.6$

$S_{0.6}34$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.6}77$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.6}35$	Swi4/Swi6	Gic1, Kre6
$S_{0.6}47$	Swi4/Swi6	Msb2
$S_{0.6}62$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.6}20$	Swi4/Swi6	Exg1
$S_{0.6}73$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.6}91$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.6}48$	Fkh1	Arp7
$S_{0.6}37$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.6}64$	Mcm1	Ste2
$S_{0.6}63$	Mcm1	Far1

Table 2.9: Range-normalized data, $\gamma = 0.8$

$S_{0.8}51$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.8}7$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.8}64$	Swi4/Swi6	Gic1, Kre6
$S_{0.8}90$	Swi4/Swi6	Msb2
$S_{0.8}31$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.8}43$	Swi4/Swi6	Exg1
$S_{0.8}65$	Swi4/Swi6	Cwp1
$S_{0.8}13$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.8}17$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.8}76$	Fkh1	Arp7
$S_{0.8}74$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.8}33$	Mcm1	Ste2
$S_{0.8}32$	Mcm1	Far1

Table 2.10: RN Data, $\gamma = 0.91$ (Shrinkage Clusters)

S49	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
S73	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S45	Swi4/Swi6	Gic1, Kre6
S15	Swi4/Swi6	Msb2
S90	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S56	Swi4/Swi6	Exg1
S46	Swi4/Swi6	Cwp1
S71	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S61	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S37	Fkh1	Arp7
S7	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S91	Mcm1	Ste2
S92	Mcm1	Far1

Table 2.11: RN Data, $\gamma = 1.0$ (Pearson Clusters)

P10	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
P68	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
P1	Swi4/Swi6	Gic1, Kre6
P39	Swi4/Swi6	Msb2
P66	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
P20	Swi4/Swi6	Exg1
P2	Swi4/Swi6	Cwp1
P72	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
P53	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P12	Fkh1	Arp7
P46	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P64	Mcm1	Ste2
P65	Mcm1	Far1

Table 2.12: RN Subsampled Data, $\gamma = 0.0$ (Eisen)

E58	Swi4/Swi6	Cln1, Och1
E68	Swi4/Swi6	Cln2, Msb2, Rsr1, Bud9, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1, Arp7
	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Mcm3, Mcm6, Cdc6
E29	Swi4/Swi6	Gic1
E64	Swi4/Swi6	Gic2
E33	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
E73	Fkh1	Tel2
E23	Ace2/Swi5	Cts1
E43	Mcm1	Ste2
E66	Mcm1	Far1

Table 2.13: RN Subsampled Data, $\gamma = 0.66$ (Shrinkage)

S49	Swi4/Swi6	Cln1, Bud9, Och1
	Ace2/Swi5	Egt2
	Mcm1	Cdc6
S6	Swi4/Swi6	Cln2, Gic2, Msb2, Rsr1, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
S32	Swi4/Swi6	Gic1
S65	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
S15	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S11	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S60	Swi4/Swi6	Hta3
S30	Fkh1	Arp7
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S62	Fkh1	Tem1
S53	Ace2/Swi5	Cts1
S14	Mcm1	Mcm6
S35	Mcm1	Ste2
S36	Mcm1	Far1

Table 2.14: RN Subsampled Data, $\gamma = 1.0$ (Pearson)

P1	Swi4/Swi6	Cln1, Och1
P15	Swi4/Swi6	Cln2, Rsr1, Mnn1
	Swi6/Mbp1	Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Mcm1	Mcm3
P29	Swi4/Swi6	Gic1
P2	Swi4/Swi6	Gic2
P3	Swi4/Swi6	Msb2, Exg1
	Swi6/Mbp1	Rnr1
P51	Swi4/Swi6	Bud9
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Cdc6
P11	Swi4/Swi6	Kre6
P62	Swi4/Swi6	Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
P49	Swi6/Mbp1	Rad27
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P10	Fkh1	Tel2
	Mcm1	Mcm6
P23	Fkh1	Arp7
P50	Fkh1	Tem1
P69	Ace2/Swi5	Cts1
P42	Mcm1	Ste2
P13	Mcm1	Far1

Chapter 3

Hybridization Models

ABSTRACT

Microarray technology, in its simplest form, allows one to gather abundance data for target DNA molecules, associated with genomes or gene-expressions, and relies on hybridizing the target to many short probe oligonucleotides arrayed on a surface. While for such multiplexed reactions conditions are optimized to make the most of each individual probe-target interaction, subsequent analysis of these experiments is based on the implicit assumption that a given experiment gives the same result regardless of whether it was conducted in isolation or in parallel with many others. It has been discussed in the literature that this assumption is frequently false, and its validity depends on the types of probes and their interactions with each other. We present a detailed physical model of hybridization as a means of understanding probe interactions in a multiplexed reaction. The model is formulated as a system

of ordinary differential equations (ODE's) describing kinetic mass action and conservation-of-mass equations completing the system.

We examine pair-wise probe interactions in detail and present a model of "competition" between the probes for the target—especially, when target is in short supply. These effects are shown to be predictable from the affinity constants for each of the four probe sequences involved, namely, the match and mismatch for both probes. These affinity constants are calculated from the thermodynamic parameters such as the free energy of hybridization, which are in turn computed according to the nearest neighbor (NN) model for each probe and target sequence.

Simulations based on the competitive hybridization model explain the observed variability in the signal of a given probe when measured in parallel with different groupings of other probes or individually. The results of the simulations are used for experiment design and pooling strategies, based on which probes have been shown to have a strong effect on each other's signal in the *in silico* experiment. These results are aimed at better design of multiplexed reactions on arrays used in genotyping (e.g., HLA typing) and mutation analysis (e.g., cystic fibrosis).

3.1 Preliminary

Recognition of a target nucleic acid and analysis of its composition can be carried out by hybridization based on complementary base pairing with a suitably designed

much shorter probe oligonucleotide. In essence, the presence of one of several possible known "messages" in the target is detected by checking if a population of identical targets in solution binds, under suitable thermodynamic conditions, to the probe molecules encoding a sequence, designed to be complementary to a message. Furthermore, a more precise quantitative answer can be obtained if other "control" probes are also mixed in with the designed probe in a well-controlled proportion and sharing similar thermodynamic properties.

Many recent advances in genome analysis, detection of polymorphisms, molecular karyotyping, and gene-expression analysis have relied on our abilities to conduct high-throughput multiplexed hybridization involving thousands or millions of probes on a surface (e.g., gene-chips and microarrays) and then, interpret the resulting assay readings. Thus, the reliability of the final computational interpretation of the data depends on understanding the errors due to unintended interactions among targets and probes, as probes and targets are multiplexed.

In particular, we focus on a mathematical analysis of "competitive hybridization," a phenomenon that has been observed in experimental data, but not adequately explained. In the following simple example of this phenomenon, a target consisting of possibly two distinct messages m_A and m_B can be characterized by separately hybridizing the target with either a mixture of specific probes pm_A and control probes mm_A or a mixture of specific probes pm_B and control probes mm_B, respectively. In either case the ratio of specific signal to the control signal, obtained from each separate experiment, indicates how often either message is present. On the other hand, contrary to one's expectations, if the two messages were queried by ratios of the respective signals in a multiplexed experiment consisting of all four

probes pm_A, mm_A, pm_B, and mm_B, one finds these ratios to differ from their values in the earlier experiments and by amounts that cannot simply be explained by the statistical noise. In particular, if one of the ratio values decreases severely, the resulting false negative errors will yield a catastrophic failure of the entire multiplexed assay. Clearly, the situation worsens precipitously as the number of multiplexed probes is increased to any realistic number. Furthermore, it becomes important to ask whether such a multiplexed assay can be rescued by judicious choice of the selected probes and the thermodynamic parameters.

3.2 Setup

More specifically, we consider the following experimental setup: Probes are bound to encoded microparticles (e.g., "beads") whose sizes are relatively large compared to the size of the probes. We assume that there are thousands of copies of the same probe attached to a single bead, and that the beads are spaced on a planar surface far enough apart in order to ensure that a single target strand may only hybridize to probes on a single bead. Thus, for all intends and purposes, this assumption implies that the only possible complexes involve one target and one probe. The targets are obtained from a longer DNA, by PCR amplification with two primers to select clones of a region that are subjected to further characterization.

Let T be a target with a single region perfectly complementary to probe P_{11} and another region perfectly complementary to probe P_{12}.

Let P_{01} differ from P_{11} in one base (i.e, the Hamming distance between P_{01} and

P_{11} equals to 1, $H(P_{01}, P_{11}) = 1$). If P_{11} and P_{01} are the only probes present, we can expect that when we compare the concentration of the P_{11} probes bound to T (denoted $[TP_{11}]$) to the concentration of the P_{01} probes bound to T (denoted $[TP_{01}]$) the resulting ratio to be large, i.e.,

$$\frac{[TP_{11}]}{[TP_{01}]} \gg 1,$$

since their free energies are chosen to satisfy $\Delta G(P_{01}) < \Delta G(P_{11})$. P_{01} clearly "competes" with P_{11} for the target T.

Consider yet another probe, P_{02}, that differs from P_{11} in one base as well ($H(P_{11}, P_{02}) = 1$), but at a location different from the one in P_{01} ($H(P_{01}, P_{02}) = 2$). Then P_{02} also competes with P_{11}, but not as much with P_{01}, since $H(P_{01}, P_{02}) = 2$. Thus, in the presence of P_{02}, we expect $\frac{[TP_{11}]}{[TP_{01}]}$ to decrease, since $[TP_{01}]$ does not decrease much, but $[TP_{11}]$ does. However, in the presence of all four probes P_{11}, P_{01}, P_{12}, and P_{02}, the analysis of the resulting "mutual competitions" poses a non-trivial problem.

3.3 Dynamics

A mathematical model to analyze the dynamics involved in a setup like the earlier one is described below. As before, we assume that the steric effects prevent multiple probes from hybridizing to a single target strand (as probes are bound to large beads).

3.3.1 Full Model

We may observe a target strand T in one of the following nine *possible states*:

(1) T (Target is unbound.)

(2) TP_{11}^{1}, (3) TP_{01}^{1}, (4) TP_{12}^{2}, (5) TP_{02}^{2}

(Target is bound by "specific" hybridization.)

(8) TP_{11}^{2}, (9) TP_{01}^{2}, (6) TP_{12}^{1}, (7) TP_{02}^{1}

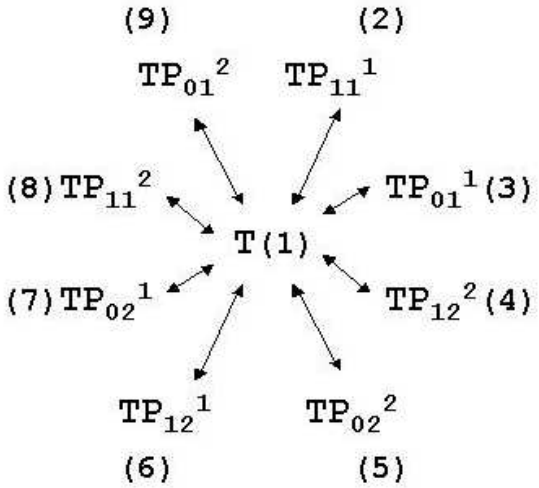
(Target is bound by "non-specific" hybridization.)

Bound target states have form TP_{ij}^{k}, where $j \in \{1, 2\}$ is the probe index,

$$i = \begin{cases} 1 & \text{for matched probe,} \\ 0 & \text{for mismatch probe,} \end{cases}$$

and $k \in \{1, 2\}$ is the binding site. States within each category are numbered "left-to-right" w.r.t. location on the target.

State Transition Diagram



The set of reversible reactions operating between unbound and bound states can be written as shown below, where the forward and backward reaction rates are indicated with $k_{i,j}$ and $k_{j,i}$, respectively. While the reaction rates themselves are difficult to compute, the ratios (affinity constants, $K_i^j = k_{i,j}/k_{j,i}$) may be computed from purely thermodynamic considerations, and are sufficient for the "equilibrium analysis."

$$
\begin{array}{llll}
T + P_{11} \underset{k_{2,1}}{\overset{k_{1,2}}{\rightleftarrows}} TP_{11}^{\ 1} & & T + P_{12} \underset{k_{6,1}}{\overset{k_{1,6}}{\rightleftarrows}} TP_{12}^{\ 1} & \\
T + P_{01} \underset{k_{3,1}}{\overset{k_{1,3}}{\rightleftarrows}} TP_{01}^{\ 1} & & T + P_{02} \underset{k_{7,1}}{\overset{k_{1,7}}{\rightleftarrows}} TP_{02}^{\ 1} & \\
T + P_{12} \underset{k_{4,1}}{\overset{k_{1,4}}{\rightleftarrows}} TP_{12}^{\ 2} & & T + P_{11} \underset{k_{8,1}}{\overset{k_{1,8}}{\rightleftarrows}} TP_{11}^{\ 2} & \\
T + P_{02} \underset{k_{5,1}}{\overset{k_{1,5}}{\rightleftarrows}} TP_{02}^{\ 2} & & T + P_{01} \underset{k_{9,1}}{\overset{k_{1,9}}{\rightleftarrows}} TP_{01}^{\ 2} &
\end{array}
$$

We wish to perform a stationary analysis, where these reactions are allowed to run to *equilibrium*. We begin by assuming that all complexes can be distinguished and writing down the ODE's (ordinary differential equations) describing the dynamics of the system as follows.

$$
\begin{aligned}
\frac{d[T]}{dt} \;=\; & k_{2,1}[TP_{11}^{\ 1}] + k_{3,1}[TP_{01}^{\ 1}] + k_{4,1}[TP_{12}^{\ 2}] + k_{5,1}[TP_{02}^{\ 2}] \\
& + k_{6,1}[TP_{12}^{\ 1}] + k_{7,1}[TP_{02}^{\ 1}] + k_{8,1}[TP_{11}^{\ 2}] + k_{9,1}[TP_{01}^{\ 2}] \\
& - k_{1,2}[T][P_{11}] - k_{1,3}[T][P_{01}] - k_{1,4}[T][P_{12}] - k_{1,5}[T][P_{02}] \\
& - k_{1,6}[T][P_{12}] - k_{1,7}[T][P_{02}] - k_{1,8}[T][P_{11}] - k_{1,9}[T][P_{01}] \qquad (3.1)
\end{aligned}
$$

$$\frac{d[TP_{11}{}^1]}{dt} = k_{1,2}[T][P_{11}] - k_{2,1}[TP_{11}{}^1] \tag{3.2}$$
$$\frac{d[TP_{01}{}^1]}{dt} = k_{1,3}[T][P_{01}] - k_{3,1}[TP_{01}{}^1] \tag{3.3}$$

$$\frac{d[TP_{12}{}^2]}{dt} = k_{1,4}[T][P_{12}] - k_{4,1}[TP_{12}{}^2] \tag{3.4}$$
$$\frac{d[TP_{02}{}^2]}{dt} = k_{1,5}[T][P_{02}] - k_{5,1}[TP_{02}{}^2] \tag{3.5}$$

$$\frac{d[TP_{12}{}^1]}{dt} = k_{1,6}[T][P_{12}] - k_{6,1}[TP_{12}{}^1] \tag{3.6}$$
$$\frac{d[TP_{02}{}^1]}{dt} = k_{1,7}[T][P_{02}] - k_{7,1}[TP_{02}{}^1] \tag{3.7}$$

$$\frac{d[TP_{11}{}^2]}{dt} = k_{1,8}[T][P_{11}] - k_{8,1}[TP_{11}{}^2] \tag{3.8}$$
$$\frac{d[TP_{01}{}^2]}{dt} = k_{1,9}[T][P_{01}] - k_{9,1}[TP_{01}{}^2] \tag{3.9}$$

Let

$$\begin{aligned} \vec{X} &= (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9)^T \\ &= \Big([T], \\ &\quad [TP_{11}{}^1], [TP_{01}{}^1], [TP_{12}{}^2], [TP_{02}{}^2], \\ &\quad [TP_{12}{}^1], [TP_{02}{}^1], [TP_{11}{}^2], [TP_{01}{}^2] \Big)^T \end{aligned} \tag{3.10}$$

Note that at equilibrium,

$$\frac{d\vec{X}}{dt} = \vec{0}. \tag{3.11}$$

Applying (3.11) to equations (3.2) – (3.9) yields

$$\begin{aligned} k_{1,2}[T][P_{11}] &= k_{2,1}[TP_{11}{}^{1}] \\ \implies K_1^2 \equiv \frac{k_{1,2}}{k_{2,1}} &= \frac{[TP_{11}{}^{1}]}{[T][P_{11}]} \end{aligned} \tag{3.12}$$

$$\begin{aligned} k_{1,3}[T][P_{01}] &= k_{3,1}[TP_{01}{}^{1}] \\ \implies K_1^3 \equiv \frac{k_{1,3}}{k_{3,1}} &= \frac{[TP_{01}{}^{1}]}{[T][P_{01}]} \end{aligned} \tag{3.13}$$

$$\begin{aligned} k_{1,4}[T][P_{12}] &= k_{4,1}[TP_{12}{}^{2}] \\ \implies K_1^4 \equiv \frac{k_{1,4}}{k_{4,1}} &= \frac{[TP_{12}{}^{2}]}{[T][P_{12}]} \end{aligned} \tag{3.14}$$

$$\begin{aligned} k_{1,5}[T][P_{02}] &= k_{5,1}[TP_{02}{}^{2}] \\ \implies K_1^5 \equiv \frac{k_{1,5}}{k_{5,1}} &= \frac{[TP_{02}{}^{2}]}{[T][P_{02}]} \end{aligned} \tag{3.15}$$

$$\begin{aligned} k_{1,6}[T][P_{12}] &= k_{6,1}[TP_{12}{}^{1}] \\ \implies K_1^6 \equiv \frac{k_{1,6}}{k_{6,1}} &= \frac{[TP_{12}{}^{1}]}{[T][P_{12}]} \end{aligned} \tag{3.16}$$

$$\begin{aligned} k_{1,7}[T][P_{02}] &= k_{7,1}[TP_{02}{}^{1}] \\ \implies K_1^7 \equiv \frac{k_{1,7}}{k_{7,1}} &= \frac{[TP_{02}{}^{1}]}{[T][P_{02}]} \end{aligned} \tag{3.17}$$

$$\begin{aligned} k_{1,8}[T][P_{11}] &= k_{8,1}[TP_{11}{}^{2}] \\ \implies K_1^8 \equiv \frac{k_{1,8}}{k_{8,1}} &= \frac{[TP_{11}{}^{2}]}{[T][P_{11}]} \end{aligned} \tag{3.18}$$

$$\begin{aligned} k_{1,9}[T][P_{01}] &= k_{9,1}[TP_{01}{}^{1}] \\ \implies K_1^9 \equiv \frac{k_{1,9}}{k_{9,1}} &= \frac{[TP_{01}{}^{1}]}{[T][P_{01}]} \end{aligned} \tag{3.19}$$

and applying it to equation (3.1) yields

$$\begin{aligned} & k_{2,1}[TP_{11}{}^{1}] + k_{3,1}[TP_{01}{}^{1}] + k_{4,1}[TP_{12}{}^{2}] + k_{5,1}[TP_{02}{}^{2}] \\ & \quad + k_{6,1}[TP_{12}{}^{1}] + k_{7,1}[TP_{02}{}^{1}] + k_{8,1}[TP_{11}{}^{2}] + k_{9,1}[TP_{01}{}^{2}] \\ = \; & [T]\big(k_{1,2}[P_{11}] + k_{1,3}[P_{01}] + k_{1,4}[P_{12}] + k_{1,5}[P_{02}] \\ & \quad + k_{1,6}[P_{12}] + k_{1,7}[P_{02}] + k_{1,8}[P_{11}] + k_{1,9}[P_{01}]\big) \end{aligned} \tag{3.20}$$

Equation (3.20) is a linear combination of (3.12), ..., (3.19), and hence provides no additional information. Observe that

$$\begin{aligned} \frac{d[T]}{dt} &= -\frac{d}{dt}\big\{[TP_{11}{}^{1}] + [TP_{01}{}^{1}] + [TP_{12}{}^{2}] + [TP_{02}{}^{2}] \\ &\qquad +[TP_{12}{}^{1}] + [TP_{02}{}^{1}] + [TP_{11}{}^{2}] + [TP_{01}{}^{2}]\big\} \\ \text{or (3.1)} &= -\sum_{j=(3.2)}^{(3.9)} \{\text{equation } j\} \end{aligned}$$

The constants K_1^j for $j \in \{2, \ldots, 9\}$, appearing in equations (3.12)–(3.19), can be computed from probe sequence data. For each j,

$$\Delta G_{\text{total}} = -RT \ln K_1^j,$$

where R is the gas constant and T is the temperature (in degrees Kelvin). Thus, we have

$$K_1^j = \exp\left[-\Delta G_{\text{total}}/RT\right], \tag{3.21}$$

where

$$\Delta G_{\text{total}} = -(\underbrace{\Delta g_i}_{\text{initiation}} + \underbrace{\Delta g_{\text{symm}}}_{\text{symmetry}}) + \sum_x \underbrace{\Delta g_x}_{\text{sequence data}} .$$

This notation and form follows [9]. Since a more recent paper by SantaLucia [40] presents the calculation of ΔG_{total} in a slightly different format (see equation (3.131)), both versions are available in the implementation of our model.

The described model can now be used to predict equilibrium concentrations of complexes TP_{ij} $\{i \in \{0,1\}, j \in \{1,2\}\}$:

- K_1^j can be computed from (3.21), where ΔG_{total} is computed based on sequence information.
- The following conservation rules must hold:

$$[P_{11}]_0 = [P_{11}] + [TP_{11}^{\ 1}] + [TP_{11}^{\ 2}] \tag{3.22}$$

$$[P_{01}]_0 = [P_{01}] + [TP_{01}^{\ 1}] + [TP_{01}^{\ 2}] \tag{3.23}$$

$$[P_{12}]_0 = [P_{12}] + [TP_{12}^{\ 1}] + [TP_{12}^{\ 2}] \tag{3.24}$$

$$[P_{02}]_0 = [P_{02}] + [TP_{02}^{\ 1}] + [TP_{02}^{\ 2}] \tag{3.25}$$

$$\begin{aligned} [T]_0 &= [T] + [TP_{11}^{\ 1}] + [TP_{01}^{\ 1}] + [TP_{12}^{\ 2}] + [TP_{02}^{\ 2}] \\ &\quad + [TP_{11}^{\ 2}] + [TP_{01}^{\ 2}] + [TP_{12}^{\ 1}] + [TP_{02}^{\ 1}] \\ &= [T] + ([P_{11}]_0 - [P_{11}]) + ([P_{01}]_0 - [P_{01}]) \\ &\quad + ([P_{12}]_0 - [P_{12}]) + ([P_{02}]_0 - [P_{02}]) \end{aligned} \tag{3.26}$$

Note that in these expressions $[X]_0$ denotes initial concentration of X, which is a free parameter, and $[X]$ denotes its equilibrium concentration.

Consider the system consisting of equations (3.12)–(3.19) and the conservation rule equations (3.22)–(3.26). We have a system of

- 13 polynomial equations (some quadratic, others linear) in
- 13 unknowns: $X_1, \ldots, X_9$ (see (3.10)) and $[P_{11}]$, $[P_{01}]$, $[P_{12}]$, $[P_{02}]$, with
- 5 free parameters: $[P_{11}]_0$, $[P_{01}]_0$, $[P_{12}]_0$, $[P_{02}]_0$, and $[T]_0$.

Therefore, this algebraic system, when solved, yields the equilibrium concentrations. From these computed concentrations, we can evaluate the "match-to-mismatch ratio" for each probe:

$$\left(\frac{[TP_{11}^{\ 1}] + [TP_{11}^{\ 2}]}{[TP_{01}^{\ 1}] + [TP_{01}^{\ 2}]}\right)_{\text{full model}}$$

and

$$\left(\frac{[TP_{12}^{\ 2}] + [TP_{12}^{\ 1}]}{[TP_{02}^{\ 2}] + [TP_{02}^{\ 1}]}\right)_{\text{full model}}$$

In order to examine the effects of competition between probes P_{11} and P_{12} on the signals for each of them, we should now compare this situation with the one where only P_{11} and P_{01} are present without P_{12} or P_{02}, and vice versa. In the rest of the paper, we will refer to the model introduced in this section as the *Full Model* and will compare its performance with the other two partial models, one consisting of P_{11}, P_{01}, and T only (referred to as *Model I*) and the other consisting of P_{12}, P_{02}, and T only (referred to as *Model II*).

3.3.2 Partial Model — Model I

This model consists of two probes P_{11}, P_{01}, and the target T only. We proceed as before by solving the algebraic system of equations to evaluate:

$$\left(\frac{[TP_{11}{}^{1}] + [TP_{11}{}^{2}]}{[TP_{01}{}^{1}] + [TP_{01}{}^{2}]}\right)_{\mathrm{I}}$$

Possible States

We consider the following states:

(1) T

(Target is unbound.)

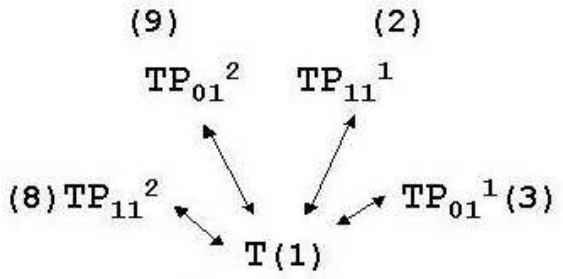
(2) $TP_{11}{}^{1}$, **(3)** $TP_{01}{}^{1}$

(Target is bound by "specific" hybridization.)

(8) $TP_{11}{}^{2}$, **(9)** $TP_{01}{}^{2}$

(Target is bound by "non-specific" hybridization.)

State Transition Diagram



The set of reversible reactions operating between unbound and bound states can be written as shown below.

$$T + P_{11} \underset{k_{2,1}}{\overset{k_{1,2}}{\rightleftarrows}} TP_{11}{}^{1} \qquad T + P_{11} \underset{k_{8,1}}{\overset{k_{1,8}}{\rightleftarrows}} TP_{11}{}^{2}$$

$$T + P_{01} \underset{k_{3,1}}{\overset{k_{1,3}}{\rightleftarrows}} TP_{01}{}^{1} \qquad T + P_{01} \underset{k_{9,1}}{\overset{k_{1,9}}{\rightleftarrows}} TP_{01}{}^{2}$$

Dynamics

The following are the ODE's describing the dynamics of the system.

$$\begin{aligned}
\frac{d[T]}{dt} &= k_{2,1}[TP_{11}{}^{1}] - k_{1,2}[T][P_{11}] \\
&\quad + k_{3,1}[TP_{01}{}^{1}] - k_{1,3}[T][P_{01}] \\
&\quad + k_{8,1}[TP_{11}{}^{2}] - k_{1,8}[T][P_{11}] \\
&\quad + k_{9,1}[TP_{01}{}^{2}] - k_{1,9}[T][P_{01}] \qquad (3.27) \\
\frac{d[TP_{11}{}^{1}]}{dt} &= k_{1,2}[T][P_{11}] - k_{2,1}[TP_{11}{}^{1}] \qquad (3.28) \\
\frac{d[TP_{01}{}^{1}]}{dt} &= k_{1,3}[T][P_{01}] - k_{3,1}[TP_{01}{}^{1}] \qquad (3.29) \\
\frac{d[TP_{11}{}^{2}]}{dt} &= k_{1,8}[T][P_{11}] - k_{8,1}[TP_{11}{}^{2}] \qquad (3.30) \\
\frac{d[TP_{01}{}^{2}]}{dt} &= k_{1,9}[T][P_{01}] - k_{9,1}[TP_{01}{}^{2}] \qquad (3.31)
\end{aligned}$$

Note that equations (3.28)–(3.31) are the same as equations (3.2), (3.3), (3.8), and (3.9) in the original system in section 3.3.1, while equation (3.27) differs from (3.1), since it now involves only the states with probes P_{11} and P_{01}.

At equilibrium, $\frac{d[\cdot]}{dt} = 0$ for all substances, i.e., T, $TP_{11}{}^{1}$, $TP_{11}{}^{2}$, $TP_{01}{}^{1}$, and $TP_{01}{}^{2}$,

yielding:

$$K_1^2 = \frac{[TP_{11}{}^1]}{[T][P_{11}]} \tag{3.32}$$

$$K_1^3 = \frac{[TP_{01}{}^1]}{[T][P_{01}]} \tag{3.33}$$

$$K_1^8 = \frac{[TP_{11}{}^2]}{[T][P_{11}]} \tag{3.34}$$

$$K_1^9 = \frac{[TP_{01}{}^1]}{[T][P_{01}]} \tag{3.35}$$

Since nothing else has changed in the thermodynamics, K_1^j computed from (3.21) are the same as before for $j \in \{2, 3, 8, 9\}$, and we have the following conservation rules:

$$[P_{11}]_0 = [P_{11}] + [TP_{11}{}^1] + [TP_{11}{}^2] \tag{3.36}$$

$$[P_{01}]_0 = [P_{01}] + [TP_{01}{}^1] + [TP_{01}{}^2] \tag{3.37}$$

$$\begin{aligned} [T]_0 &= [T] + [TP_{11}{}^1] + [TP_{01}{}^1] + [TP_{11}{}^2] + [TP_{01}{}^2] \\ &= [T] + ([P_{11}]_0 - [P_{11}]) + ([P_{01}]_0 - [P_{01}]) \end{aligned} \tag{3.38}$$

In this case, we have

- 7 variables (unknowns): $[TP_{11}{}^1]$, $[TP_{11}{}^2]$, $[TP_{01}{}^1]$, $[TP_{01}{}^2]$, $[P_{11}]$, $[P_{01}]$, and $[T]$; and
- 7 polynomial equations: (3.32)–(3.35), (3.36), (3.37), and (3.38), with
- 3 free parameters $[P_{11}]_0$, $[P_{01}]_0$, and $[T]_0$.

Note that, for comparison with full model, the free parameters will need to be scaled to retain the same initial target-to-probe ratio.

3.3.3 Partial Model — Model II

This model consists of two probes P_{12}, P_{02}, and the target T only. We proceed as before by solving the algebraic system of equations to evaluate:

$$\left(\frac{[TP_{12}^{\ 2}] + [TP_{12}^{\ 1}]}{[TP_{02}^{\ 2}] + [TP_{02}^{\ 1}]}\right)_{\text{II}}$$

Possible States

We consider the following states:

(1) T

(Target is unbound.)

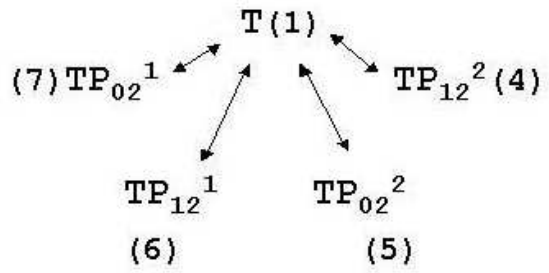
(4) $TP_{12}^{\ 2}$, **(5)** $TP_{02}^{\ 2}$

(Target is bound by "specific" hybridization.)

(6) $TP_{12}^{\ 1}$, **(7)** $TP_{02}^{\ 1}$

(Target is bound by "non-specific" hybridization.)

State Transition Diagram



The set of reversible reactions operating between unbound and bound states can be written as shown below.

$$T + P_{12} \underset{k_{4,1}}{\overset{k_{1,4}}{\rightleftarrows}} TP_{12}^{\,2} \qquad T + P_{12} \underset{k_{6,1}}{\overset{k_{1,6}}{\rightleftarrows}} TP_{12}^{\,1}$$

$$T + P_{02} \underset{k_{5,1}}{\overset{k_{1,5}}{\rightleftarrows}} TP_{02}^{\,2} \qquad T + P_{02} \underset{k_{7,1}}{\overset{k_{1,7}}{\rightleftarrows}} TP_{02}^{\,1}$$

Dynamics

The following are the ODE's describing the dynamics of the system.

$$\begin{aligned}
\frac{d[T]}{dt} &= k_{4,1}[TP_{12}^{\,2}] - k_{1,4}[T][P_{12}] \\
&\quad + k_{5,1}[TP_{02}^{\,2}] - k_{1,5}[T][P_{02}] \\
&\quad + k_{6,1}[TP_{12}^{\,1}] - k_{1,6}[T][P_{12}] \\
&\quad + k_{7,1}[TP_{02}^{\,1}] - k_{1,7}[T][P_{02}] \qquad (3.39) \\
\frac{d[TP_{12}^{\,2}]}{dt} &= k_{1,4}[T][P_{12}] - k_{4,1}[TP_{12}^{\,2}] \qquad (3.40) \\
\frac{d[TP_{02}^{\,2}]}{dt} &= k_{1,5}[T][P_{02}] - k_{5,1}[TP_{02}^{\,2}] \qquad (3.41) \\
\frac{d[TP_{12}^{\,1}]}{dt} &= k_{1,6}[T][P_{12}] - k_{6,1}[TP_{12}^{\,1}] \qquad (3.42) \\
\frac{d[TP_{02}^{\,1}]}{dt} &= k_{1,7}[T][P_{02}] - k_{7,1}[TP_{02}^{\,1}] \qquad (3.43) \\
& \qquad (3.44)
\end{aligned}$$

Note that equations (3.40)–(3.43) are the same as equations (3.4), (3.5), (3.6), and (3.7) in the original system in section 3.3.1, while equation (3.39) differs from (3.1), since it now involves only the states with probes P_{12} and P_{02}.

At equilibrium, $\frac{d[\cdot]}{dt} = 0$ for all substances, i.e., T, $TP_{12}^{\,2}$, $TP_{12}^{\,1}$, $TP_{02}^{\,2}$, and $TP_{02}^{\,1}$,

yielding:

$$K_1^4 = \frac{[TP_{12}^{\ 2}]}{[T][P_{12}]} \tag{3.45}$$

$$K_1^5 = \frac{[TP_{02}^{\ 2}]}{[T][P_{02}]} \tag{3.46}$$

$$K_1^6 = \frac{[TP_{12}^{\ 1}]}{[T][P_{12}]} \tag{3.47}$$

$$K_1^7 = \frac{[TP_{02}^{\ 1}]}{[T][P_{02}]} \tag{3.48}$$

Again, since nothing else has changed in the thermodynamics, K_1^j computed from (3.21) are the same as before for $j \in \{4, 5, 6, 7\}$, and we have the following conservation rules:

$$[P_{12}]_0 = [P_{12}] + [TP_{12}^{\ 2}] + [TP_{12}^{\ 1}] \tag{3.49}$$

$$[P_{02}]_0 = [P_{02}] + [TP_{02}^{\ 2}] + [TP_{02}^{\ 1}] \tag{3.50}$$

$$\begin{aligned} [T]_0 &= [T] + [TP_{12}^{\ 2}] + [TP_{02}^{\ 2}] + [TP_{12}^{\ 1}] + [TP_{02}^{\ 1}] \\ &= [T] + ([P_{12}]_0 - [P_{12}]) + ([P_{02}]_0 - [P_{02}]) \end{aligned} \tag{3.51}$$

In this case, we also have

- 7 variables: $[TP_{12}^{\ 2}]$, $[TP_{12}^{\ 1}]$, $[TP_{02}^{\ 2}]$, $[TP_{02}^{\ 1}]$, $[P_{12}]$, $[P_{02}]$, and $[T]$; and
- 7 equations: (3.45)–(3.48), (3.49), (3.50), and (3.51), with
- 3 free parameters $[P_{12}]_0$, $[P_{02}]_0$, and $[T]_0$.

As above (section 3.3.2), the parameters will need to be scaled.

In practice, once the *exact nucleotide sequences* of T, P_{11}, P_{01}, P_{12}, and P_{02} are determined from the needs of the biological assay, we can compute K_1^j explicitly, and then solve for the unknowns in all three setups:

- Full Model,
- Model I, and
- Model II.

With these computed ratio values, we are ready to evaluate and compare the models in order to discern the effects of competition:

$$\left(\frac{P_{11}}{P_{01}}\right)_{\text{full}} \quad \text{vs.} \quad \left(\frac{P_{11}}{P_{01}}\right)_{\text{I}}$$

and

$$\left(\frac{P_{12}}{P_{02}}\right)_{\text{full}} \quad \text{vs.} \quad \left(\frac{P_{12}}{P_{02}}\right)_{\text{II}}$$

3.4 Change of Variables

3.4.1 Full Model

In order to simplify the algebraic system of equation, we rename the unknown variables as follows:

$$\begin{aligned}
&X_1 = [T] \\
&X_2 = [TP_{11}{}^1] \qquad X_6 = [TP_{12}{}^1] \qquad Y_1 = [P_{11}] \\
&X_3 = [TP_{01}{}^1] \qquad X_7 = [TP_{02}{}^1] \qquad Y_2 = [P_{01}] \\
&X_4 = [TP_{12}{}^2] \qquad X_8 = [TP_{11}{}^2] \qquad Y_3 = [P_{12}] \\
&X_5 = [TP_{02}{}^2] \qquad X_9 = [TP_{01}{}^2] \qquad Y_4 = [P_{02}]
\end{aligned}$$

The constant parameters in the system are initially left in their symbolic forms.

$$
\begin{aligned}
&K_1^2, \quad K_1^3, \quad K_1^4, \quad K_1^5, \quad K_1^6, \quad K_1^7, \quad K_1^8, \quad K_1^9, \\
&a_0 = [P_{11}]_0, \quad b_0 = [P_{01}]_0, \\
&c_0 = [P_{12}]_0, \quad d_0 = [P_{02}]_0, \\
&e_0 = [T]_0.
\end{aligned}
$$

Equations (3.12)–(3.19) and (3.22)–(3.26) can now be rewritten in terms of $\{X_i, Y_j\}$ as follows.

$$
\left.
\begin{array}{rcl@{\quad\Longrightarrow\quad}rcl}
[TP_{11}^{\ 1}] &=& K_1^2[T][P_{11}] & X_2 &=& K_1^2 X_1 Y_1 \\
[TP_{01}^{\ 1}] &=& K_1^3[T][P_{01}] & X_3 &=& K_1^3 X_1 Y_2 \\
[TP_{12}^{\ 2}] &=& K_1^4[T][P_{12}] & X_4 &=& K_1^4 X_1 Y_3 \\
[TP_{02}^{\ 2}] &=& K_1^5[T][P_{02}] & X_5 &=& K_1^5 X_1 Y_4 \\
[TP_{12}^{\ 1}] &=& K_1^6[T][P_{12}] & X_6 &=& K_1^6 X_1 Y_3 \\
[TP_{02}^{\ 1}] &=& K_1^7[T][P_{02}] & X_7 &=& K_1^7 X_1 Y_4 \\
[TP_{11}^{\ 2}] &=& K_1^8[T][P_{11}] & X_8 &=& K_1^8 X_1 Y_1 \\
[TP_{01}^{\ 2}] &=& K_1^9[T][P_{01}] & X_9 &=& K_1^9 X_1 Y_2 \\
[P_{11}]_0 &=& [P_{11}] + [TP_{11}^{\ 1}] + [TP_{11}^{\ 2}] & a_0 &=& X_2 + X_8 + Y_1 \\
[P_{01}]_0 &=& [P_{01}] + [TP_{01}^{\ 1}] + [TP_{01}^{\ 2}] & b_0 &=& X_3 + X_9 + Y_2 \\
[P_{12}]_0 &=& [P_{12}] + [TP_{12}^{\ 1}] + [TP_{12}^{\ 2}] & c_0 &=& X_4 + X_6 + Y_3 \\
[P_{02}]_0 &=& [P_{02}] + [TP_{02}^{\ 1}] + [TP_{02}^{\ 2}] & d_0 &=& X_5 + X_7 + Y_4 \\
[T]_0 &=& [T] + [TP_{11}^{\ 1}] + [TP_{01}^{\ 1}] & e_0 &=& X_1 + X_2 + X_3 \\
&& \quad + [TP_{12}^{\ 2}] + [TP_{02}^{\ 2}] & && \quad + X_4 + X_5 \\
&& \quad + [TP_{11}^{\ 2}] + [TP_{01}^{\ 2}] & && \quad + X_6 + X_7 \\
&& \quad + [TP_{12}^{\ 1}] + [TP_{02}^{\ 1}] & && \quad + X_8 + X_9
\end{array}
\right\} \tag{3.52}
$$

3.4.2 Model I

Now, we consider a system of algebraic equations representing the concentrations at equilibrium and involving unknown variables X_1, X_2, X_3, X_8, X_9, Y_1, and Y_2, and constant parameters K_1^2, K_1^3, K_1^8, K_1^9, a_0, b_0, and e_0. Thus, in a manner analogous to that derived for the full model in the previous section, we may rewrite the equations (3.32), (3.33), (3.34), (3.35), (3.36), (3.37), and (3.38) in terms of $\{X_i, Y_j\}$, as shown below.

$$\left.\begin{aligned} X_2 &= K_1^2 X_1 Y_1 \\ X_3 &= K_1^3 X_1 Y_2 \\ X_8 &= K_1^8 X_1 Y_1 \\ X_9 &= K_1^9 X_1 Y_2 \\ a_0 &= [P_{11}]_0 = X_2 + X_8 + Y_1 \\ b_0 &= [P_{01}]_0 = X_3 + X_9 + Y_2 \\ e_0 &= [T]_0 = X_1 + X_2 + X_3 + X_8 + X_9 \end{aligned}\right\} \tag{3.53}$$

3.4.3 Model II

Next, we consider a system of algebraic equations representing the concentrations at equilibrium and involving unknown variables X_1, X_4, X_5, X_6, X_7, Y_3, and Y_4, and constant parameters K_1^4, K_1^5, K_1^6, K_1^7, c_0, d_0, and e_0. Once again we may rewrite the equations (3.45), (3.46), (3.47), (3.48), (3.49), (3.50), and (3.51) in terms of

$\{X_i, Y_j\}$, as shown below.

$$\left.\begin{aligned} X_4 &= K_1^4 X_1 Y_3 \\ X_5 &= K_1^5 X_1 Y_4 \\ X_6 &= K_1^6 X_1 Y_3 \\ X_7 &= K_1^7 X_1 Y_4 \\ c_0 &= [P_{12}]_0 = X_4 + X_6 + Y_3 \\ d_0 &= [P_{02}]_0 = X_5 + X_7 + Y_4 \\ e_0 &= [T]_0 = X_1 + X_4 + X_5 + X_6 + X_7 \end{aligned}\right\} \tag{3.54}$$

Note that with the exception of the conservation rules for $[T]$ (i.e., the last equations in (3.52), (3.53), and (3.54)) under the different models, we have

$$(3.52) = (3.53) \cup (3.54).$$

3.5 System Reduction

3.5.1 Model I

Starting with (3.53), we may obtain the following linear equalities:

$$Y_1 = a_0 - X_2 - X_8 \tag{3.55}$$

$$Y_2 = b_0 - X_3 - X_9 \tag{3.56}$$

Furthermore, since

$$\frac{X_2}{X_8} = \frac{K_1^2 X_1 Y_1}{K_1^8 X_1 Y_1} = \frac{K_1^2}{K_1^8} \quad \Longrightarrow \quad X_8 = \frac{K_1^8}{K_1^2} X_2 \tag{3.57}$$

$$\frac{X_3}{X_9} = \frac{K_1^3 X_1 Y_2}{K_1^9 X_1 Y_2} = \frac{K_1^3}{K_1^9} \quad \Longrightarrow \quad X_9 = \frac{K_1^9}{K_1^3} X_3 \tag{3.58}$$

we may simplify to obtain

$$\begin{aligned}
X_2 &= K_1^2 X_1 Y_1 = K_1^2 X_1 (a_0 - X_2 - X_8) \\
&= K_1^2 X_1 \left(a_0 - X_2 - \frac{K_1^8}{K_1^2} X_2 \right) \\
&= K_1^2 X_1 \left(a_0 - X_2 \left[1 + \frac{K_1^8}{K_1^2} \right] \right) \\
&= K_1^2 X_1 a_0 - K_1^2 X_1 X_2 \left[1 + \frac{K_1^8}{K_1^2} \right] \\
&= K_1^2 X_1 a_0 - X_1 X_2 \left[K_1^2 + K_1^8 \right] \\
X_2 &+ X_1 X_2 \left[K_1^2 + K_1^8 \right] = a_0 K_1^2 X_1 \\
&\therefore \quad \boxed{X_2 = \frac{a_0 K_1^2 X_1}{1 + X_1 \left(K_1^2 + K_1^8 \right)}}
\end{aligned} \tag{3.59}$$

and

$$\begin{aligned}
X_3 &= K_1^3 X_1 Y_2 = K_1^3 X_1 (b_0 - X_3 - X_9) \\
&= K_1^3 X_1 \left(b_0 - X_3 - \frac{K_1^9}{K_1^3} X_3 \right) \\
&= K_1^3 X_1 \left(b_0 - X_3 \left[1 + \frac{K_1^9}{K_1^3} \right] \right) \\
&= K_1^3 X_1 b_0 - K_1^3 X_1 X_3 \left[1 + \frac{K_1^9}{K_1^3} \right] \\
&= K_1^3 X_1 b_0 - X_1 X_3 \left[K_1^3 + K_1^9 \right] \\
X_3 &+ X_1 X_3 \left[K_1^3 + K_1^9 \right] = b_0 K_1^3 X_1 \\
&\therefore \quad \boxed{X_3 = \frac{b_0 K_1^3 X_1}{1 + X_1 \left(K_1^3 + K_1^9 \right)}}
\end{aligned} \tag{3.60}$$

We also obtain, from (3.57),

$$\begin{aligned}
X_8 &= \frac{K_1^8}{K_1^2} X_2 = \frac{K_1^8}{K_1^2} \, \frac{a_0 K_1^2 X_1}{1 + X_1 \left(K_1^2 + K_1^8 \right)} \\
&= \boxed{\frac{a_0 K_1^8 X_1}{1 + X_1 \left(K_1^2 + K_1^8 \right)} = X_8}
\end{aligned} \tag{3.61}$$

and from (3.58),

$$\begin{aligned} X_9 &= \frac{K_1^9}{K_1^3} X_3 = \frac{K_1^9}{K_1^3} \frac{b_0 K_1^3 X_1}{1 + X_1 \left(K_1^3 + K_1^9\right)} \\ &= \boxed{\frac{b_0 K_1^9 X_1}{1 + X_1 \left(K_1^3 + K_1^9\right)} = X_9} \end{aligned} \tag{3.62}$$

Finally, equations (3.59), (3.60), (3.61), and (3.62) can be solved to express X_2, X_3, X_8, and X_9, respectively, in terms of X_1.

Now, from (3.55), (3.57), and (3.59), we derive

$$\begin{aligned} Y_1 &= a_0 - X_2 - X_8 = a_0 - X_2 \left(1 + \frac{K_1^8}{K_1^2}\right) \\ &= a_0 - \frac{a_0 K_1^2 X_1}{1 + X_1 \left(K_1^2 + K_1^8\right)} \left(1 + \frac{K_1^8}{K_1^2}\right) \\ &= a_0 - \frac{a_0 X_1 \left(K_1^2 + K_1^8\right)}{1 + X_1 \left(K_1^2 + K_1^8\right)} \\ &= a_0 \left[\frac{1 + X_1 \left(K_1^2 + K_1^8\right) - X_1 \left(K_1^2 + K_1^8\right)}{1 + X_1 \left(K_1^2 + K_1^8\right)}\right] \\ &= \frac{a_0}{1 + X_1 \left(K_1^2 + K_1^8\right)} \\ \therefore \quad & \boxed{Y_1 = \frac{a_0}{1 + X_1 \left(K_1^2 + K_1^8\right)}} \end{aligned} \tag{3.63}$$

Similarly, we derive

$$\begin{aligned}
Y_2 &= b_0 - X_3 - X_9 = b_0 - X_3\left(1 + \frac{K_1^9}{K_1^3}\right) \\
&= b_0 - \frac{b_0 K_1^3 X_1}{1 + X_1\left(K_1^3 + K_1^9\right)}\left(1 + \frac{K_1^9}{K_1^3}\right) \\
&= b_0 - \frac{b_0 X_1\left(K_1^3 + K_1^9\right)}{1 + X_1\left(K_1^3 + K_1^9\right)} \\
&= b_0\left[\frac{1 + X_1\left(K_1^3 + K_1^9\right) - X_1\left(K_1^3 + K_1^9\right)}{1 + X_1\left(K_1^3 + K_1^9\right)}\right] \\
&= \frac{b_0}{1 + X_1\left(K_1^3 + K_1^9\right)} \\
\therefore \quad & \boxed{Y_2 = \frac{b_0}{1 + X_1\left(K_1^3 + K_1^9\right)}} \qquad (3.64)
\end{aligned}$$

A final simplification yields a univariate rational function only in X_1 equating to a constant e_0:

$$\begin{aligned}
e_0 &= X_1 + X_2 + X_3 + X_8 + X_9 \qquad \text{(by (3.53))} \\
&= X_1 + X_1\frac{a_0 K_1^2}{1 + X_1\left(K_1^2 + K_1^8\right)} \\
&\quad + X_1\frac{b_0 K_1^3}{1 + X_1\left(K_1^3 + K_1^9\right)} \quad \text{(by (3.59),(3.60))} \\
&\quad + X_1\frac{a_0 K_1^8}{1 + X_1\left(K_1^2 + K_1^8\right)} \\
&\quad + X_1\frac{b_0 K_1^9}{1 + X_1\left(K_1^3 + K_1^9\right)} \quad \text{(by (3.61),(3.62))}
\end{aligned}$$

or

$$\begin{aligned}
e_0 &= X_1\Bigg[1 + a_0\frac{K_1^2 + K_1^8}{1 + X_1\left(K_1^2 + K_1^8\right)} \\
&\qquad + b_0\frac{K_1^3 + K_1^9}{1 + X_1\left(K_1^3 + K_1^9\right)}\Bigg] \qquad (3.65)
\end{aligned}$$

Since the terms $(K_1^2 + K_1^8)$ and $(K_1^3 + K_1^9)$ appear frequently, in order to express the preceding equations in a simpler form, we introduce short-hand notations shown

below. Let

$$s_{28} \equiv K_1^2 + K_1^8, \quad s_{39} \equiv K_1^3 + K_1^9, \quad \text{and } x \equiv X_1.$$

In the simplified from, the equation (3.65) becomes

$$\begin{aligned}
&x\left(1 + a_0\frac{s_{28}}{1+s_{28}x} + b_0\frac{s_{39}}{1+s_{39}x}\right) = e_0 \\
&x\left(\frac{(1+s_{28}x)(1+s_{39}x) + a_0 s_{28}(1+s_{39}x) + b_0 s_{39}(1+s_{28}x)}{(1+s_{28}x)(1+s_{39}x)}\right) = e_0 \\
&x\left((1+s_{28}x)(1+s_{39}x) + a_0 s_{28}(1+s_{39}x) + b_0 s_{39}(1+s_{28}x)\right) \\
&\qquad\qquad = e_0(1+s_{28}x)(1+s_{39}x),
\end{aligned}$$

or

$$\begin{aligned}
&(s_{28}s_{39})x^3 + (s_{28} + s_{39} + s_{28}s_{39}[a_0 + b_0 - e_0])\,x^2 \\
&\quad + (1 + s_{28}[a_0 - e_0] + s_{39}[b_0 - e_0])\,x - e_0 = 0
\end{aligned} \tag{3.66}$$

Now the cubic polynomial equation (3.66) must be solved for the unknown $x = X_1$, and then the solution can be substituted into (3.59)–(3.64) in order to solve for the rest of the variables. We may obtain the solutions in their symbolic form using Mathematica ([48]) as the three possible roots may be easily expressed in radicals. More to the point, we only need to solve for

$$\begin{aligned}
\left(\frac{P_{11}}{P_{01}}\right)_{\mathrm{I}} &= \boxed{\left(\frac{[TP_{11}{}^{1}] + [TP_{11}{}^{2}]}{[TP_{01}{}^{1}] + [TP_{01}{}^{2}]}\right)_{\mathrm{I}}} = \frac{X_2 + X_8}{X_3 + X_9} \\
&= \left(\frac{a_0 K_1^2 x}{1+s_{28}x} + \frac{a_0 K_1^8 x}{1+s_{28}x}\right) \Big/ \left(\frac{b_0 K_1^3 x}{1+s_{39}x} + \frac{b_0 K_1^9 x}{1+s_{39}x}\right)
\end{aligned}$$

or

$$\begin{aligned}\left(\frac{P_{11}}{P_{01}}\right)_{\mathrm{I}} &= \left(\frac{a_0 s_{28} x}{1+s_{28}x}\right) \Big/ \left(\frac{b_0 s_{39} x}{1+s_{39}x}\right) \\ &= \boxed{\frac{a_0}{b_0}\,\frac{s_{28}}{s_{39}}\,\frac{1+s_{39}x}{1+s_{28}x}} \\ &= \frac{a_0}{b_0}\,\frac{s_{28}}{s_{39}}\,\frac{s_{39}+1/x}{s_{28}+1/x}\end{aligned} \tag{3.67}$$

where x is a solution of (3.66).

3.5.2 Model II

As before, starting with (3.54), we may obtain the following linear equalities:

$$Y_3 = c_0 - X_4 - X_6 \tag{3.68}$$

$$Y_4 = d_0 - X_5 - X_7 \tag{3.69}$$

Since

$$\frac{X_4}{X_6} = \frac{K_1^4 X_1 Y_3}{K_1^6 X_1 Y_3} = \frac{K_1^4}{K_1^6} \quad \Longrightarrow \quad X_6 = \frac{K_1^6}{K_1^4} X_4 \tag{3.70}$$

$$\frac{X_5}{X_7} = \frac{K_1^5 X_1 Y_4}{K_1^7 X_1 Y_4} = \frac{K_1^5}{K_1^7} \quad \Longrightarrow \quad X_7 = \frac{K_1^7}{K_1^5} X_5 \tag{3.71}$$

we obtain

$$\begin{aligned} X_4 &= K_1^4 X_1 Y_3 = K_1^4 X_1 \left(c_0 - X_4 \left[1 + \frac{K_1^6}{K_1^4}\right]\right) \\ &= K_1^4 X_1 c_0 - X_1 X_4 \left(K_1^4 + K_1^6\right) \\ \therefore \quad & \boxed{X_4 = \frac{c_0 K_1^4 X_1}{1 + X_1 \left(K_1^4 + K_1^6\right)}} \end{aligned} \tag{3.72}$$

and

$$\begin{aligned} X_5 &= K_1^5 X_1 Y_4 = K_1^5 X_1 \left(d_0 - X_5 \left[1 + \frac{K_1^7}{K_1^5} \right] \right) \\ &= K_1^5 X_1 d_0 - X_1 X_5 \left(K_1^5 + K_1^7 \right) \\ \therefore \quad & \boxed{X_5 = \frac{d_0 K_1^5 X_1}{1 + X_1 \left(K_1^5 + K_1^7 \right)}} \end{aligned} \tag{3.73}$$

Furthermore, from (3.70) and (3.72), we obtain

$$\begin{aligned} X_6 &= \frac{K_1^6}{K_1^4} X_4 = \frac{K_1^6}{K_1^4} \, \frac{c_0 K_1^4 X_1}{1 + X_1 \left(K_1^4 + K_1^6 \right)} \\ &= \boxed{\frac{c_0 K_1^6 X_1}{1 + X_1 \left(K_1^4 + K_1^6 \right)} = X_6} \end{aligned} \tag{3.74}$$

and from (3.71) and (3.73),

$$\begin{aligned} X_7 &= \frac{K_1^7}{K_1^5} X_5 = \frac{K_1^7}{K_1^5} \, \frac{d_0 K_1^5 X_1}{1 + X_1 \left(K_1^5 + K_1^7 \right)} \\ &= \boxed{\frac{d_0 K_1^7 X_1}{1 + X_1 \left(K_1^5 + K_1^7 \right)} = X_7} \end{aligned} \tag{3.75}$$

Finally, equations (3.72), (3.73), (3.74), and (3.75) can be solved to express X_4, X_5, X_6, and X_7, respectively, in terms of X_1.

From (3.68), (3.70), and (3.72), we derive

$$\begin{aligned} Y_3 &= c_0 - X_4 - X_6 = c_0 - X_4 \left(1 + \frac{K_1^6}{K_1^4} \right) \\ &= c_0 - \frac{c_0 K_1^4 X_1}{1 + X_1 \left(K_1^4 + K_1^6 \right)} \left(1 + \frac{K_1^6}{K_1^4} \right) \\ &= c_0 - \frac{c_0 X_1 \left(K_1^4 + K_1^6 \right)}{1 + X_1 \left(K_1^4 + K_1^6 \right)} \\ &= c_0 \left[\frac{1 + X_1 \left(K_1^4 + K_1^6 \right) - X_1 \left(K_1^4 + K_1^6 \right)}{1 + X_1 \left(K_1^4 + K_1^6 \right)} \right] \\ &= \frac{c_0}{1 + X_1 \left(K_1^4 + K_1^6 \right)} \\ \therefore \quad & \boxed{Y_3 = \frac{c_0}{1 + X_1 \left(K_1^4 + K_1^6 \right)}} \end{aligned} \tag{3.76}$$

Similarly, we derive

$$\begin{aligned}
Y_4 &= d_0 - X_5 - X_7 = d_0 - X_5\left(1 + \frac{K_1^7}{K_1^5}\right) \\
&= d_0 - \frac{d_0 K_1^5 X_1}{1 + X_1\left(K_1^5 + K_1^7\right)}\left(1 + \frac{K_1^7}{K_1^5}\right) \\
&= d_0 - \frac{d_0 X_1\left(K_1^5 + K_1^7\right)}{1 + X_1\left(K_1^5 + K_1^7\right)} \\
&= d_0\left[\frac{1 + X_1\left(K_1^5 + K_1^7\right) - X_1\left(K_1^5 + K_1^7\right)}{1 + X_1\left(K_1^5 + K_1^7\right)}\right] \\
&= \frac{d_0}{1 + X_1\left(K_1^5 + K_1^7\right)} \\
\therefore \quad & \boxed{Y_4 = \frac{d_0}{1 + X_1\left(K_1^5 + K_1^7\right)}} \qquad (3.77)
\end{aligned}$$

Putting it all together, we derive the univariate rational equation for X_1.

$$\begin{aligned}
e_0 &= X_1 + X_4 + X_5 + X_6 + X_7 \qquad \text{(by (3.54))} \\
&= X_1 + X_1\frac{c_0 K_1^4}{1 + X_1\left(K_1^4 + K_1^6\right)} \\
&\quad + X_1\frac{d_0 K_1^5}{1 + X_1\left(K_1^5 + K_1^7\right)} \quad \text{(by (3.72),(3.73))} \\
&\quad + X_1\frac{c_0 K_1^6}{1 + X_1\left(K_1^4 + K_1^6\right)} \\
&\quad + X_1\frac{d_0 K_1^7}{1 + X_1\left(K_1^5 + K_1^7\right)} \quad \text{(by (3.74),(3.75))}
\end{aligned}$$

or

$$\begin{aligned}
e_0 &= X_1\Bigg[1 + c_0\frac{K_1^4 + K_1^6}{1 + X_1\left(K_1^4 + K_1^6\right)} \\
&\qquad + d_0\frac{K_1^5 + K_1^7}{1 + X_1\left(K_1^5 + K_1^7\right)}\Bigg] \qquad (3.78)
\end{aligned}$$

As before, we abbreviate the terms $(K_1^4+K_1^6)$ and $(K_1^5+K_1^7)$ by short-hand notation, shown below. Let

$$s_{46} \equiv K_1^4 + K_1^6, \quad s_{57} \equiv K_1^5 + K_1^7, \quad \text{and} \quad y \equiv X_1.$$

Note that, in order to avoid confusion, we have introduced a different abbreviation for X_1 (i.e., y) intentionally since the equation to be solved in this case differs from (3.66). Then (3.78) can be expressed as

$$\begin{aligned}
&y\left(1+c_0\frac{s_{46}}{1+s_{46}y}+d_0\frac{s_{57}}{1+s_{57}y}\right)=e_0\\
&y\left(\frac{(1+s_{46}y)(1+s_{57}y)+c_0s_{46}(1+s_{57}y)+d_0s_{57}(1+s_{46}y)}{(1+s_{46}y)(1+s_{57}y)}\right)=e_0\\
&y\left((1+s_{46}y)(1+s_{57}y)+c_0s_{46}(1+s_{57}y)+d_0s_{57}(1+s_{46}y)\right)\\
&\qquad\qquad =e_0(1+s_{46}y)(1+s_{57}y),
\end{aligned}$$

or

$$\begin{aligned}
&(s_{46}s_{57})y^3+\left(s_{46}+s_{57}+s_{46}s_{57}[c_0+d_0-e_0]\right)y^2\\
&\quad+\left(1+s_{46}[c_0-e_0]+s_{57}[d_0-e_0]\right)y-e_0=0
\end{aligned}\tag{3.79}$$

Again, the cubic polynomial equation (3.79) must be solved for $y=X_1$, and then the solution can be substituted into (3.72)–(3.77) for the rest of the variables. Actually, we only need

$$\begin{aligned}
\left(\frac{P_{12}}{P_{02}}\right)_{\text{II}}&=\boxed{\left(\frac{[TP_{12}^{\ 2}]+[TP_{12}^{\ 1}]}{[TP_{02}^{\ 2}]+[TP_{02}^{\ 1}]}\right)_{\text{II}}}=\frac{X_4+X_6}{X_5+X_7}\\
&=\left(\frac{c_0K_1^4y}{1+s_{46}y}+\frac{c_0K_1^6y}{1+s_{46}y}\right)\Big/\left(\frac{d_0K_1^5y}{1+s_{57}y}+\frac{d_0K_1^7y}{1+s_{57}y}\right)
\end{aligned}$$

or

$$\begin{aligned}
\left(\frac{P_{12}}{P_{02}}\right)_{\text{II}}&=\left(\frac{c_0s_{46}y}{1+s_{46}y}\right)\Big/\left(\frac{d_0s_{57}y}{1+s_{57}y}\right)\\
&=\boxed{\frac{c_0}{d_0}\ \frac{s_{46}}{s_{57}}\ \frac{1+s_{57}y}{1+s_{46}y}}\\
&=\frac{c_0}{d_0}\ \frac{s_{46}}{s_{57}}\ \frac{s_{57}+1/y}{s_{46}+1/y}
\end{aligned}\tag{3.80}$$

where y solves (3.79).

3.5.3 Full Model

As noted in section 3.4, the system (3.52) of equations for the Full Model is simply the union of the systems (3.53) and (3.54) for models I and II, respectively, with the exception of the conservation rule for $[T]$, i.e., the equation for X_1. Therefore, while the equation for X_1 itself must be handled separately, the derivations from sections 3.5.1 and 3.5.2 can be duplicated to obtain equations for all the variables in terms of X_1. For convenience, we gather the resulting equations in one place, as shown below.

$$\boxed{X_2 = \frac{a_0 K_1^2 X_1}{1 + X_1 (K_1^2 + K_1^8)}} \quad \text{(see (3.59))} \tag{3.81}$$

$$\boxed{X_3 = \frac{b_0 K_1^3 X_1}{1 + X_1 (K_1^3 + K_1^9)}} \quad \text{(see (3.60))} \tag{3.82}$$

$$\boxed{X_4 = \frac{c_0 K_1^4 X_1}{1 + X_1 (K_1^4 + K_1^6)}} \quad \text{(see (3.72))} \tag{3.83}$$

$$\boxed{X_5 = \frac{d_0 K_1^5 X_1}{1 + X_1 (K_1^5 + K_1^7)}} \quad \text{(see (3.73))} \tag{3.84}$$

$$\boxed{X_6 = \frac{c_0 K_1^6 X_1}{1 + X_1 (K_1^4 + K_1^6)}} \quad \text{(see (3.74))} \tag{3.85}$$

$$\boxed{X_7 = \frac{d_0 K_1^7 X_1}{1 + X_1 (K_1^5 + K_1^7)}} \quad \text{(see (3.75))} \tag{3.86}$$

$$\boxed{X_8 = \frac{a_0 K_1^8 X_1}{1 + X_1 (K_1^2 + K_1^8)}} \quad \text{(see (3.61))} \tag{3.87}$$

$$\boxed{X_9 = \frac{b_0 K_1^9 X_1}{1 + X_1 (K_1^3 + K_1^9)}} \quad \text{(see (3.62))} \tag{3.88}$$

$$\boxed{Y_1 = \frac{a_0}{1 + X_1\,(K_1^2 + K_1^8)}} \qquad \text{(see (3.63))} \tag{3.89}$$

$$\boxed{Y_2 = \frac{b_0}{1 + X_1\,(K_1^3 + K_1^9)}} \qquad \text{(see (3.64))} \tag{3.90}$$

$$\boxed{Y_3 = \frac{c_0}{1 + X_1\,(K_1^4 + K_1^6)}} \qquad \text{(see (3.76))} \tag{3.91}$$

$$\boxed{Y_4 = \frac{d_0}{1 + X_1\,(K_1^5 + K_1^7)}} \qquad \text{(see (3.77))} \tag{3.92}$$

It remains to derive the univariate equation in X_1. Since the terms $(K_1^2 + K_1^8)$, $(K_1^3 + K_1^9)$, $(K_1^4 + K_1^6)$, and $(K_1^5 + K_1^7)$ appears frequently in the following derivation, as in the previous sections, we abbreviate these terms with the short-hand notation given below. As we did in sections 3.5.1, and 3.5.2, let

$$s_{28} \equiv K_1^2 + K_1^8, \qquad s_{39} \equiv K_1^3 + K_1^9,$$
$$s_{46} \equiv K_1^4 + K_1^6, \qquad s_{57} \equiv K_1^5 + K_1^7,$$

and let

$$z \equiv X_1.$$

Note again that a different symbol for X_1 has to be employed to avoid confusion with the variables used in equations (3.66) and (3.79).

$$
\begin{aligned}
e_0 &= X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 \quad \text{(by (3.52))} \\
&= z + z\frac{a_0 K_1^2}{1 + s_{28}z} + z\frac{b_0 K_1^3}{1 + s_{39}z} \\
&\quad + z\frac{c_0 K_1^4}{1 + s_{46}z} + z\frac{d_0 K_1^5}{1 + s_{57}z} \quad \text{(by (3.81)–(3.84))} \\
&\quad + z\frac{c_0 K_1^6}{1 + s_{46}z} + z\frac{d_0 K_1^7}{1 + s_{57}z} \\
&\quad + z\frac{a_0 K_1^8}{1 + z s_{28}} + z\frac{b_0 K_1^9}{1 + z s_{39}} \quad \text{(by (3.85)–(3.88))} \\
&= z\Bigg[1 + a_0\frac{K_1^2 + K_1^8}{1 + s_{28}z} + b_0\frac{K_1^3 + K_1^9}{1 + s_{39}z} \\
&\quad + c_0\frac{K_1^4 + K_1^6}{1 + s_{46}z} + d_0\frac{K_1^5 + K_1^7}{1 + s_{57}z}\Bigg] \qquad (3.93)
\end{aligned}
$$

or

$$
\begin{aligned}
e_0 &= z\Bigg[1 + \frac{a_0 s_{28}}{1 + s_{28}z} + \frac{b_0 s_{39}}{1 + s_{39}z} \\
&\quad + \frac{c_0 s_{46}}{1 + s_{46}z} + \frac{d_0 s_{57}}{1 + s_{57}z}\Bigg] \qquad (3.94)
\end{aligned}
$$

or

$$
\begin{aligned}
&(1 + s_{28}z)(1 + s_{39}z)(1 + s_{46}z)(1 + s_{57}z)e_0 = \\
&\quad = z[(1 + s_{28}z)(1 + s_{39}z)(1 + s_{46}z)(1 + s_{57}z) + \\
&\quad + a_0 s_{28}(1 + s_{39}z)(1 + s_{46}z)(1 + s_{57}z) + \\
&\quad + b_0 s_{39}(1 + s_{28}z)(1 + s_{46}z)(1 + s_{57}z) + \\
&\quad + c_0 s_{46}(1 + s_{28}z)(1 + s_{39}z)(1 + s_{57}z) + \\
&\quad + d_0 s_{57}(1 + s_{28}z)(1 + s_{39}z)(1 + s_{46}z)] \qquad (3.95)
\end{aligned}
$$

Since we now have a 5^{th} order polynomial equation in z to solve, and since its roots cannot be expressed symbolically in a closed form, we must resort to a purely

numerical approach. Nonetheless, the match-to-mismatch ratio signals can be obtained in terms of z.

$$\left(\frac{P_{11}}{P_{01}}\right)_{\text{full}} = \boxed{\left(\frac{[TP_{11}{}^{1}] + [TP_{01}{}^{2}]}{[TP_{01}{}^{1}] + [TP_{01}{}^{2}]}\right)_{\text{full}}} = \frac{X_2 + X_8}{X_3 + X_9}$$
$$= \boxed{\frac{a_0}{b_0}\ \frac{s_{28}}{s_{39}}\ \frac{1 + s_{39}z}{1 + s_{28}z}} = \frac{a_0}{b_0}\ \frac{s_{28}}{s_{39}}\ \frac{s_{39} + 1/z}{s_{28} + 1/z} \qquad \text{(see (3.67))} \qquad (3.96)$$

and

$$\left(\frac{P_{12}}{P_{02}}\right)_{\text{full}} = \boxed{\left(\frac{[TP_{12}{}^{2}] + [TP_{12}{}^{1}]}{[TP_{02}{}^{2}] + [TP_{02}{}^{1}]}\right)_{\text{full}}} = \frac{X_4 + X_6}{X_5 + X_7}$$
$$= \boxed{\frac{c_0}{d_0}\ \frac{s_{46}}{s_{57}}\ \frac{1 + s_{57}z}{1 + s_{46}z}} = \frac{c_0}{d_0}\ \frac{s_{46}}{s_{57}}\ \frac{s_{57} + 1/z}{s_{46} + 1/z} \qquad \text{(see (3.80))}, \qquad (3.97)$$

where z solves (3.95).

3.6 Additional Models

Next, for the purpose of comparison, we will consider two additional models: *Simple Model*, where the target has exactly one region for the probe to hybridize with, and *Extended Full Model*, where the target has three regions for hybridization and the multiplexed assay involves three pairs of "match" and "mismatch" probes. In particular, while the simple model allows us to understand how just the mismatch probe should be designed optimally, the extended full model gives us insight into the extent to which a system of three or more multiplexed probe pairs can be designed by considering only two probe pairs at a time.

3.6.1 Simple Model

We consider a situation where the target has exactly one region for the probe to hybridize with. Thus, we have three possible states to model: unbound targets, targets bound to "match" probes in the region of interest, and lastly, targets bound to "mismatch" probes in the region of interest—all other possible hybridization states are ignored.

Possible States

We consider the following three states:

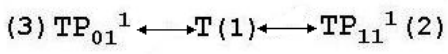
(1) T

(Target is unbound.)

(2) TP_{11}^{1}, (3) TP_{01}^{1}

(Target is bound by "specific" hybridization.)

State Transition Diagram



The set of reversible reactions operating between unbound and bound states can be written as shown below.

$$T + P_{11} \underset{k_{2,1}}{\overset{k_{1,2}}{\rightleftharpoons}} TP_{11}^{1}$$

$$T + P_{01} \underset{k_{3,1}}{\overset{k_{1,3}}{\rightleftharpoons}} TP_{01}^{1}$$

Dynamics

The following are the ODE's (ordinary differential equations) describing the dynamics of the system.

$$\begin{aligned}
\frac{d[T]}{dt} &= k_{2,1}[TP_{11}^{\ 1}] - k_{1,2}[T][P_{11}] \\
&\quad + k_{3,1}[TP_{01}^{\ 1}] - k_{1,3}[T][P_{01}] \qquad (3.98) \\
\frac{d[TP_{11}^{\ 1}]}{dt} &= k_{1,2}[T][P_{11}] - k_{2,1}[TP_{11}^{\ 1}] = 0 \quad \text{(at equilibrium)} \qquad (3.99) \\
\frac{d[TP_{01}^{\ 1}]}{dt} &= k_{1,3}[T][P_{01}] - k_{3,1}[TP_{01}^{\ 1}] = 0 \quad \text{(at equilibrium)} \qquad (3.100)
\end{aligned}$$

Thus at equilibrium, the ODE's yield the following algebraic equations.

$$\begin{aligned}
k_{1,2}[T][P_{11}] &= k_{2,1}[TP_{11}^{\ 1}] \\
\Longrightarrow X_2 = [TP_{11}^{\ 1}] &= \frac{k_{1,2}}{k_{2,1}}[T][P_{11}] = K_1^2[T][P_{11}] \\
k_{1,3}[T][P_{01}] &= k_{3,1}[TP_{01}^{\ 1}] \\
\Longrightarrow X_3 = [TP_{01}^{\ 1}] &= \frac{k_{1,3}}{k_{3,1}}[T][P_{01}] = K_1^3[T][P_{01}]
\end{aligned}$$

$$\text{ratio1} = \frac{K_1^2[T][P_{11}]}{K_1^3[T][P_{01}]} = \frac{X_2}{X_3} \qquad (3.101)$$

We augment the above equations with the linear constraints corresponding to the conservation rules.

$$\begin{aligned}
T: &\quad [T] + [TP_{11}^{\ 1}] + [TP_{01}^{\ 1}] = [T]_0 = e_0 \\
P_{11}: &\quad [P_{11}] + [TP_{11}^{\ 1}] = [P_{11}]_0 = a_0 \\
P_{01}: &\quad [P_{01}] + [TP_{01}^{\ 1}] = [P_{01}]_0 = b_0
\end{aligned}$$

Finally, we gather the system of equations to be solved, with the appropriate change of variables.

$$\begin{array}{ll} X_1 = [T] & X_1 + X_2 + X_3 = e_0 \\ X_2 = [TP_{11}{}^1] & X_2 = K_1^2 X_1 Y_1 \\ X_3 = [TP_{01}{}^1] & X_3 = K_1^3 X_1 Y_2 \\ Y_1 = [P_{11}] & X_2 + Y_1 = a_0 \\ Y_2 = [P_{01}] & X_3 + Y_2 = b_0 \end{array}$$

After simplification, we have

$$\begin{aligned} Y_1 &= a_0 - X_2 \\ Y_2 &= b_0 - X_3 \end{aligned}$$

$$\begin{aligned} X_2 = K_1^2 X_1 (a_0 - X_2) &= K_1^2 X_1 a_0 - K_1^2 X_1 X_2 \\ \Longrightarrow \boxed{X_2} = \frac{a_0 K_1^2 X_1}{1 + K_1^2 X_1} &= \boxed{\frac{a_0 K_1^2}{K_1^2 + \frac{1}{X_1}}} \\ X_3 = K_1^3 X_1 (b_0 - X_3) &= K_1^3 X_1 b_0 - K_1^3 X_1 X_3 \\ \Longrightarrow \boxed{X_3} = \frac{b_0 K_1^3 X_1}{1 + K_1^3 X_1} &= \boxed{\frac{b_0 K_1^3}{K_1^3 + \frac{1}{X_1}}} \end{aligned}$$

Finally, we get the following equation involving rational functions in one variable X_1.

$$X_1 = e_0 - X_2 - X_3 = e_0 - a_0 \frac{K_1^2 X_1}{1 + K_1^2 X_1} - b_0 \frac{K_1^3 X_1}{1 + K_1^3 X_1} \tag{3.102}$$

Simplifying equation (3.102) for X_1, we have the following equation with $w \equiv X_1$.

$$\begin{aligned} e_0 &= w + a_0 \frac{K_1^2}{1 + K_1^2 w} w + b_0 \frac{K_1^3}{1 + K_1^3 w} w \\ &= \boxed{w \left(1 + a_0 \frac{K_1^2}{1 + K_1^2 w} + b_0 \frac{K_1^3}{1 + K_1^3 w} \right) = e_0} \end{aligned} \tag{3.103}$$

We may solve (3.103) for w numerically or symbolically (e.g., in Mathematica). Writing ratio1 in terms of the roots of the above equation, we get

$$\begin{aligned}
\text{ratio1} &= \frac{X_2}{X_3} = \frac{a_0 K_1^2 X_1}{1 + K_1^2 X_1} \frac{1 + K_1^3 X_1}{b_0 K_1^3 X_1} \\
&= \frac{a_0 K_1^2}{K_1^2 + \frac{1}{X_1}} \frac{K_1^3 + \frac{1}{X_1}}{b_0 K_1^3} \\
&= \frac{a_0}{b_0} \cdot \frac{K_1^2}{K_1^3} \cdot \frac{K_1^3 + \frac{1}{X_1}}{K_1^2 + \frac{1}{X_1}} \qquad (3.104) \\
&= \frac{a_0}{b_0} \cdot \frac{K_1^2}{K_1^3} \cdot \frac{1 + \frac{1}{X_1 K_1^3}}{\frac{K_1^2}{K_1^3} + \frac{1}{X_1 K_1^3}}. \qquad (3.105)
\end{aligned}$$

According to (3.104), if $X_1 \gg 1$ then we have ratio1 $\equiv (a_0/b_0)$. On the other hand, if $K_1^2/K_1^3 \sim \frac{1}{X_1 K_1^3}$, i.e., $K_1^2 \sim \frac{1}{X_1}$, then the ratio simplifies to the following, indicating that the ratio depends on the initial concentration of the probes and their thermodynamic parameters.

$$\begin{aligned}
\text{ratio1} &\sim \frac{a_0}{b_0} \frac{K_1^2}{K_1^3} \frac{(X_1 K_1^3 + 1)/X_1 K_1^3}{2/X_1 K_1^3} \\
&= \frac{1}{2} \frac{a_0}{b_0} \frac{1}{X_1 K_1^3} \left(X_1 K_1^3 + 1\right) \qquad (3.106) \\
&= \frac{1}{2} \frac{a_0}{b_0} \left(1 + \frac{1}{X_1 K_1^3}\right) \\
&= \frac{1}{2} \frac{a_0}{b_0} \left(1 + \frac{K_1^2}{K_1^3}\right) \qquad (3.107)
\end{aligned}$$

We need to further investigate what the proper initial target concentration $[T]_0 = e_0$ must be to optimize the expected observation of competition:

- If the initial target concentration is too dense then the expected ratio $\equiv \frac{a_0}{b_0}$, where a_0 = the initial concentration of the matched probe and b_0 = the

initial concentration of the mismatched probe. As discussed in section B.1.1, these two parameters are usually set to be equal, i.e., $a_0 = b_0$. Thus, in this situation, we *cannot* distinguish *pm* from *mm*.

- If the initial target concentration is too diluted then the expected ratio *can* distinguish *pm* and *mm*, but signal strength is so low that the detected intensities "drown" in noise.

3.6.2 Extended Full Model

The final mathematical model (Extended Full Model) involves multiplexed hybridization of a single target with three different probes and can be used to verify that the effects suggested by pairwise probe analysis extend to probe triples correctly.

In this scheme, we will consider one target, three possible binding sites and three probe pairs, one for each binding site, as shown in the figure.

Possible States

We consider the following states:

(1) T (Target is unbound.)

(2) $TP_{11}^{\ 1}$, **(3)** $TP_{01}^{\ 1}$, **(4)** $TP_{12}^{\ 2}$, **(5)** $TP_{02}^{\ 2}$, **(6)** $TP_{13}^{\ 3}$, **(7)** $TP_{03}^{\ 3}$

(Target is bound by "specific" hybridization;

P_{ij} hybridizes to site j.)

(8) $TP_{11}^{\ 2}$, **(9)** $TP_{01}^{\ 2}$, **(10)** $TP_{11}^{\ 3}$, **(11)** $TP_{01}^{\ 3}$, **(12)** $TP_{12}^{\ 1}$, **(13)** $TP_{02}^{\ 1}$,

(14) $TP_{12}^{\ 3}$, **(15)** $TP_{02}^{\ 3}$, **(16)** $TP_{13}^{\ 1}$, **(17)** $TP_{03}^{\ 1}$, **(18)** $TP_{13}^{\ 2}$, **(19)** $TP_{03}^{\ 2}$

(Target is bound by "cross-hybridization";

P_{ij} hybridizes to site k, $j \neq k$.)

State Transition Diagram

The state transition diagram for this model is not shown, as it involves 19 states and is cumbersome to display. The state interaction can be easily inferred from (3.108).

The set of reversible reactions operating between unbound and bound states can be written as shown below, where K_i^j denotes the affinity constant for going from

state $i(=1)$ to state $j \in \{2, 3, \ldots, 19\}$.

$$\begin{cases} T + P_{11} & \rightleftarrows & TP_{11}^{\,k} \\ T + P_{01} & \rightleftarrows & TP_{01}^{\,k} \\ T + P_{12} & \rightleftarrows & TP_{12}^{\,k} \\ T + P_{02} & \rightleftarrows & TP_{02}^{\,k} \\ T + P_{13} & \rightleftarrows & TP_{13}^{\,k} \\ T + P_{03} & \rightleftarrows & TP_{03}^{\,k} \end{cases}$$

Site k	1	2	3
States l:	**(2)**	**(8)**	**(10)**
	(3)	**(9)**	**(11)**
	(12)	**(4)**	**(14)**
	(13)	**(5)**	**(15)**
	(16)	**(18)**	**(6)**
	(17)	**(19)**	**(7)**

(3.108)

Dynamics

The following ODE's describe the dynamics of the system.

$$\frac{d[TP_{ij}^{\,k}]}{dt} = k_{1,l}[T][P_{ij}] - k_{l,1}[TP_{ij}^{\,k}] \qquad (3.109)$$

(18 eqns)

where

$i \in \{0, 1\}$, probe $j \in \{1, 2, 3\}$,

site $k \in \{1, 2, 3\}$, and $l(i, j, k)$ is given in (3.108).

$$\begin{aligned} \frac{d[T]}{dt} &= -\sum_{i=0}^{1}\sum_{j=1}^{3}\sum_{k=1}^{3} k_{1,l(i,j,k)}[T][P_{ij}] + \sum_{i,j,k} k_{l,1}[TP_{ij}^{\,k}] \\ &= \sum_{i=0}^{1}\sum_{j=1}^{3}\sum_{k=1}^{3} \Big\{ k_{l,1}[TP_{ij}^{\,k}] - k_{1,l}[T][P_{ij}] \Big\} \end{aligned} \qquad (3.110)$$

At equilibrium, $\frac{d[\vec{X}]}{dt} = 0$, where $\vec{X} = (X_1, \ldots, X_{19})^T$ and

$$
\begin{array}{lll}
X_1 = [T] & & \\
X_2 = [TP_{11}^{\ 1}] & X_8 = [TP_{11}^{\ 2}] & X_{14} = [TP_{12}^{\ 3}] \\
X_3 = [TP_{01}^{\ 1}] & X_9 = [TP_{01}^{\ 2}] & X_{15} = [TP_{02}^{\ 3}] \\
X_4 = [TP_{12}^{\ 2}] & X_{10} = [TP_{11}^{\ 3}] & X_{16} = [TP_{13}^{\ 1}] \\
X_5 = [TP_{02}^{\ 2}] & X_{11} = [TP_{01}^{\ 3}] & X_{17} = [TP_{03}^{\ 1}] \\
X_6 = [TP_{13}^{\ 3}] & X_{12} = [TP_{12}^{\ 1}] & X_{18} = [TP_{13}^{\ 2}] \\
X_7 = [TP_{03}^{\ 3}] & X_{13} = [TP_{02}^{\ 1}] & X_{19} = [TP_{03}^{\ 2}]
\end{array}
$$

Applying this equilibrium condition to (3.109) yields

$$k_{1,l}[T][P_{ij}] = k_{l,1}[TP_{ij}^{\ k}]$$

or

$$\boxed{K_1^l \equiv \frac{k_{1,l}}{k_{l,1}} = \frac{[TP_{ij}^{\ k}]}{[T][P_{ij}]}} \tag{3.111}$$

while (3.110) becomes the sum of the previous 18 equations and thus provides no additional information.

Mass conservation rules add the following linear constraints:

$$[P_{ij}]_0 = [P_{ij}] + \sum_{k=1}^{3} [TP_{ij}^{\ k}] \quad \text{for } i \in \{0,1\}, j \in \{1,2,3\} \tag{3.112}$$

$$[T]_0 = [T] + \sum_{i,j,k} [TP_{ij}^{\ k}] \tag{3.113}$$

For simplification, we rename the variables as follows:

$$
\begin{array}{llll}
X_1 = [T] & & & \\
X_2 = [TP_{11}^{\ 1}] & X_8 = [TP_{11}^{\ 2}] & X_{14} = [TP_{12}^{\ 3}] & Y_1 = [P_{11}] \\
X_3 = [TP_{01}^{\ 1}] & X_9 = [TP_{01}^{\ 2}] & X_{15} = [TP_{02}^{\ 3}] & Y_2 = [P_{01}] \\
X_4 = [TP_{12}^{\ 2}] & X_{10} = [TP_{11}^{\ 3}] & X_{16} = [TP_{13}^{\ 1}] & Y_3 = [P_{12}] \\
X_5 = [TP_{02}^{\ 2}] & X_{11} = [TP_{01}^{\ 3}] & X_{17} = [TP_{03}^{\ 1}] & Y_4 = [P_{02}] \\
X_6 = [TP_{13}^{\ 3}] & X_{12} = [TP_{12}^{\ 1}] & X_{18} = [TP_{13}^{\ 2}] & Y_5 = [P_{13}] \\
X_7 = [TP_{03}^{\ 3}] & X_{13} = [TP_{02}^{\ 1}] & X_{19} = [TP_{03}^{\ 2}] & Y_6 = [P_{03}]
\end{array}
$$

We rename the constant parameters as follows:

$$
\begin{array}{ll}
K_1^l, \quad l = 2, \ldots, 19 & \\
a_0 = [P_{11}]_0, & b_0 = [P_{01}]_0, \\
c_0 = [P_{12}]_0, & d_0 = [P_{02}]_0, \\
e_0 = [P_{13}]_0, & f_0 = [P_{03}]_0, \\
g_0 = [T]_0. &
\end{array}
$$

And finally, obtain the following simplified equations:

$$
\begin{aligned}
K_1^l[T][P_{ij}] &= [TP_{ij}^{\ k}] \\
\Longrightarrow \quad & \boxed{X_l = K_1^l X_1 Y_n}, \\
& \text{where } n \text{ depends on } l(i,j,k) \\
Y_n^0 &= Y_n + \textstyle\sum_{l \in f^{-1}(n)} X_l \\
\Longrightarrow \quad & \boxed{Y_n = Y_n^0 - \sum_{l \in f^{-1}(n)} X_l} \text{ (probe conservation)} \\
& \boxed{X_1^0 = \sum_{l=1}^{19} X_l} \qquad \text{(target conservation)}
\end{aligned}
$$

In these equations we have written $f^{-1}(n)$ to denote the set of states involving probe Y_n, so that, according to (3.108), we have

$$\begin{aligned} f^{-1}(1) &= \{2, 8, 10\} & f^{-1}(4) &= \{5, 13, 15\} \\ f^{-1}(2) &= \{3, 9, 11\} & f^{-1}(5) &= \{6, 16, 18\} \\ f^{-1}(3) &= \{4, 12, 14\} & f^{-1}(6) &= \{7, 17, 19\} \end{aligned}$$

Reducing the equations further, we get:

$$n = 1 \begin{cases} X_2 &= K_1^2 X_1 Y_1 \\ X_8 &= K_1^8 X_1 Y_1 = \dfrac{K_1^8}{K_1^2} X_2 \\ X_{10} &= \dfrac{K_1^{10}}{K_1^2} X_2 \end{cases} \qquad n = 4 \begin{cases} X_5 &= K_1^5 X_1 Y_4 \\ X_{13} &= \dfrac{K_1^{13}}{K_1^2} X_5 \\ X_{15} &= \dfrac{K_1^{15}}{K_1^2} X_5 \end{cases}$$

$$n = 2 \begin{cases} X_3 &= K_1^3 X_1 Y_2 \\ X_9 &= \dfrac{K_1^9}{K_1^3} X_3 \\ X_{11} &= \dfrac{K_1^{11}}{K_1^3} X_3 \end{cases} \qquad n = 5 \begin{cases} X_6 &= K_1^6 X_1 Y_5 \\ X_{16} &= \dfrac{K_1^{16}}{K_1^6} X_6 \\ X_{18} &= \dfrac{K_1^{18}}{K_1^6} X_6 \end{cases}$$

$$n = 3 \begin{cases} X_4 &= K_1^4 X_1 Y_3 \\ X_{12} &= \dfrac{K_1^{12}}{K_1^4} X_4 \\ X_{14} &= \dfrac{K_1^{14}}{K_1^4} X_4 \end{cases} \qquad n = 6 \begin{cases} X_7 &= K_1^7 X_1 Y_6 \\ X_{17} &= \dfrac{K_1^{17}}{K_1^7} X_7 \\ X_{19} &= \dfrac{K_1^{19}}{K_1^7} X_7 \end{cases}$$

Now, we consider the equations where $l \in \{2, 8, 10\}$ and $n = 1$:

$$
\begin{aligned}
Y_1 &= Y_1^0 - (X_2 + X_8 + X_{10}) \\
X_2 &= K_1^2 X_1 Y_1 = K_1^2 X_1 \big\{ Y_1^0 - (\underbrace{X_2 + X_8 + X_{10}}) \big\} \\
&\quad \left[\begin{array}{l} X_2 + X_8 + X_{10} \\ = X_2 + \frac{K_1^8}{K_1^2} X_2 + \frac{K_1^{10}}{K_1^2} X_2 \\ = \frac{X_2}{K_1^2} [K_1^2 + K_1^8 + K_1^{10}] \\ = X_2 s_{2,8,10} / K_1^2, \\ \quad \text{where } \boxed{s_{i,j,k} = K_1^i + K_1^j + K_1^k} \end{array} \right] \\
&= K_1^2 X_1 Y_1^0 - K_1^2 X_1 \frac{X_2}{K_1^2} s_{2,8,10} \\
&= K_1^2 X_1 Y_1^0 - X_1 X_2 s_{2,8,10} \\
X_2 &+ X_1 X_2 s_{2,8,10} = K_1^2 X_1 Y_1^0 \\
&\Longrightarrow X_2 \big(1 + X_1 s_{2,8,10}\big) = K_1^2 X_1 Y_1^0 \\
&\therefore X_2 = K_1^2 Y_1^0 \frac{X_1}{1 + s_{2,8,10} X_1}
\end{aligned}
$$

Let

$$\boxed{t(2, 8, 10) = \frac{Y_1^0}{1 + s_{2,8,10} X_1}}$$

Then, we have

$$\boxed{X_2 = K_1^2 X_1 t(2, 8, 10)} \tag{3.114}$$

$$\boxed{X_8 = K_1^8 X_1 t(2, 8, 10)} \tag{3.115}$$

$$\boxed{X_{10} = K_1^{10} X_1 t(2, 8, 10)} \tag{3.116}$$

Similarly, we obtain

$$\boxed{\begin{pmatrix} X_3 \\ X_9 \\ X_{11} \end{pmatrix} = \begin{pmatrix} K_1^3 \\ K_1^9 \\ K_1^{11} \end{pmatrix} X_1 t(3,9,11)} \tag{3.117}$$

where $\boxed{t(3,9,11) = \dfrac{Y_2^0}{1 + s_{3,9,11} X_1}}$

$$\boxed{\begin{pmatrix} X_4 \\ X_{12} \\ X_{14} \end{pmatrix} = \begin{pmatrix} K_1^4 \\ K_1^{12} \\ K_1^{14} \end{pmatrix} X_1 t(4,12,14)} \tag{3.118}$$

where $\boxed{t(4,12,14) = \dfrac{Y_3^0}{1 + s_{4,12,14} X_1}}$

$$\boxed{\begin{pmatrix} X_5 \\ X_{13} \\ X_{15} \end{pmatrix} = \begin{pmatrix} K_1^5 \\ K_1^{13} \\ K_1^{15} \end{pmatrix} X_1 t(5,13,15)} \tag{3.119}$$

where $\boxed{t(5,13,15) = \dfrac{Y_4^0}{1 + s_{5,13,15} X_1}}$

$$\boxed{\begin{pmatrix} X_6 \\ X_{16} \\ X_{18} \end{pmatrix} = \begin{pmatrix} K_1^6 \\ K_1^{16} \\ K_1^{18} \end{pmatrix} X_1 t(6,16,18)} \tag{3.120}$$

where $\boxed{t(6,16,18) = \dfrac{Y_5^0}{1 + s_{6,16,18} X_1}}$

and

$$\boxed{\begin{pmatrix} X_7 \\ X_{17} \\ X_{19} \end{pmatrix} = \begin{pmatrix} K_1^7 \\ K_1^{17} \\ K_1^{19} \end{pmatrix} X_1 t(7,17,19)} \tag{3.121}$$

$$\text{where } \boxed{t(7,17,19) = \frac{Y_6^0}{1 + s_{7,17,19} X_1}}$$

After this manipulation we have equations for all X_j's in terms of X_1, $j \neq 1$. Next, we obtain equilibrium probe concentrations:

$$\begin{aligned} Y_1 &= Y_1^0 - (X_2 + X_8 + X_{10}) \\ &= Y_1^0 - s_{2,8,10} Y_1^0 X_1 \frac{1}{1 + s_{2,8,10} X_1} \\ &= Y_1^0 \left\{ 1 - \frac{X_1 s_{2,8,10}}{1 + s_{2,8,10} X_1} \right\} = Y_1^0 \frac{1}{1 + s_{2,8,10} X_1} \\ \therefore \quad & \boxed{Y_1 = \frac{Y_1^0}{1 + s_{2,8,10} X_1} = t(2,8,10)} \end{aligned} \tag{3.122}$$

Similarly, we have

$$\boxed{Y_2 = t(3,9,11)} \tag{3.123}$$

$$\boxed{Y_3 = t(4,12,14)} \tag{3.124}$$

$$\boxed{Y_4 = t(5,13,15)} \tag{3.125}$$

$$\boxed{Y_5 = t(6,16,18)} \tag{3.126}$$

$$\boxed{Y_6 = t(7,17,19)} \tag{3.127}$$

It remains to get the univariate polynomial equation for X_1—"the main equation."

$$\begin{aligned}
X_1^0 &= \sum_{l=1}^{19} X_l \\
&= X_1 + (X_2 + X_8 + X_{10}) + (X_3 + X_9 + X_{11}) + (X_4 + X_{12} + X_{14}) \\
&\qquad (X_5 + X_{13} + X_{15}) + (X_6 + X_{16} + X_{18}) + (X_7 + X_{17} + X_{19}) \\
&= X_1 + (K_1^2 + K_1^8 + K_1^{10}) X_1 t(2,8,10) \\
&\qquad + (K_1^3 + K_1^9 + K_1^{11}) X_1 t(3,9,11) \\
&\qquad + (K_1^4 + K_1^{12} + K_1^{14}) X_1 t(4,12,14) \\
&\qquad + (K_1^5 + K_1^{13} + K_1^{15}) X_1 t(5,13,15) \\
&\qquad + (K_1^6 + K_1^{16} + K_1^{18}) X_1 t(6,16,18) \\
&\qquad + (K_1^7 + K_1^{17} + K_1^{19}) X_1 t(7,17,19) \\
&= X_1 \big(1 + s_{2,8,10} t(2,8,10) + s_{3,9,11} t(3,9,11) + s_{4,12,14} t(4,12,14) \\
&\qquad + s_{5,13,15} t(5,13,15) + s_{6,16,18} t(6,16,18) + s_{7,17,19} t(7,17,19)\big)
\end{aligned}$$

Therefore,

$$\begin{aligned}
X_1^0 = X_1 \Bigg\{ & 1 + Y_1^0 \frac{s_{2,8,10}}{1 + s_{2,8,10} X_1} + Y_2^0 \frac{s_{3,9,11}}{1 + s_{3,9,11} X_1} \\
& + Y_3^0 \frac{s_{4,12,14}}{1 + s_{4,12,14} X_1} + Y_4^0 \frac{s_{5,13,15}}{1 + s_{5,13,15} X_1} \\
& + Y_5^0 \frac{s_{6,16,18}}{1 + s_{6,16,18} X_1} + Y_6^0 \frac{s_{7,17,19}}{1 + s_{7,17,19} X_1} \Bigg\}
\end{aligned} \tag{3.128}$$

which is a 7^{th} order polynomial in X_1. As in other models, (3.128) can be solved for X_1 numerically (e.g., in Mathematica).

3.7 Obtaining Thermodynamic Parameters

3.7.1 Nearest-Neighbor Model

The model of hybridization discussed so far treats the dynamics in terms of kinetic mass-action reactions and ignores both the mixing properties of the molecules and the exact physics of hybridization except for simply acknowledging that the thermodynamics parameters depend on base-pair composition. Recall that the process of hybridization involves the formation of base pairs between Watson-Crick–complementary bases. Namely, base pairing of two single stranded DNA molecules is determined by the fact that A (adenine) is complementary to T (thymine), and C (cytosine) is complementary to G (guanine). Such base pairing is due to the formation of hydrogen bonds between the complementary bases; thus, this interaction is characterized primarily by the composition of the interacting strands. Another physical interaction, *base stacking*, characterizes the hybridization process, and it has been shown to depend on the sequence rather than the composition of the strands. As base stacking depends on the short-range interactions, it is thought to be adequately described by the Nearest-Neighbor (NN) model.

In the NN model, it is assumed that the stability of a given base pair is determined by the identity and orientation of the neighboring base pairs. Thus, each thermodynamic parameter of the hybridization process, such as the change in enthalpy (ΔH), entropy (ΔS), and free energy (ΔG), is calculated as a sum of the contributions from each nearest-neighbor pair along a strand, corrected by some symmetry and initiation parameters. As the enthalpy and entropy terms may be assumed to be independent of temperature, they can be computed as follows ([9],

[40]):

$$\Delta H = \sum_x \Delta H_x + \Delta H(\text{init}) + \Delta H(\text{sym}) \tag{3.129}$$

$$\Delta S = \sum_x \Delta S_x + \Delta S(\text{init}) + \Delta S(\text{sym}) \tag{3.130}$$

where the terms ΔH_x and ΔS_x are tabulated for all ten possible NN dimer duplexes, as are the initiation and symmetry terms. The free energy computation is analogous:

$$\Delta G = \sum_x \Delta G_x + \Delta G(\text{init}) + \Delta G(\text{sym}) \tag{3.131}$$

with the initiation and symmetry terms tabulated. The values ΔG_x for the dimer duplexes have been tabulated at 25°C ([9]) and at 37°C ([40]). Since ΔG depends on the temperature, the values ΔG_x for the dimer duplexes can be easily calculated from the corresponding ΔH_x and ΔS_x parameters by

$$\Delta G_x(T) = \Delta H_x - T\Delta S_x \tag{3.132}$$

The ten distinct dimer duplexes arise as follows. Following the notation of Breslauer *et al.* ([9]), we denote each dimer duplex with a "slash-sign" separating antiparallel strands, e.g., AG/TC denotes 5'-AG-3' Watson-Crick base-paired with 3'-TC-5'. Alternately, $\frac{AG}{TC}$ is equivalent to AG/TC. The table below lists all

sixteen ($= |\{A, T, C, G\}|^2 = 4^2$) possible dimers, identifying the equivalent ones.

$$\boxed{\frac{AA}{TT}} \quad \frac{AC}{TG} \equiv \frac{GT}{CA} \quad \frac{AG}{TC} \equiv \frac{CT}{GA} \quad \boxed{\frac{AT}{TA}}$$

$$\boxed{\frac{CA}{GT}} \quad \frac{CC}{GG} \equiv \frac{GG}{CC} \quad \boxed{\frac{CG}{GC}} \qquad \boxed{\frac{CT}{GA}}$$

$$\boxed{\frac{GA}{CT}} \quad \boxed{\frac{GC}{CG}} \qquad \boxed{\frac{GG}{CC}} \qquad \boxed{\frac{GT}{CA}}$$

$$\boxed{\frac{TA}{AT}} \quad \frac{TC}{AG} \equiv \frac{GA}{CT} \quad \frac{TG}{AC} \equiv \frac{CA}{GT} \quad \frac{TT}{AA} \equiv \frac{AA}{TT}$$

Since our simulations involve oligonucleotide probes, we used the parameters for the initiation of duplex formation drawn from the results in the 1998 paper of SantaLucia ([40]). There, two different initiation parameters were introduced to account for the differences between duplexes with terminal $A \cdot T$ and duplexes with terminal $G \cdot C$. The additional "symmetry" parameter accounts for the maintenance of the $C2$ symmetry of self-complementary duplexes ([10]).

The table of parameters used in our simulations, drawn from [40], is duplicated in Table 3.1 for convenience. The following example illustrates how the free energy can be computed according to (3.131) using the values from Table 3.1.

Example

```
5'  C-G-A-A-G-T  3'
    * * * * * *
3'  G-C-T-T-C-A  5'
```

Table 3.1: Unified oligonucleotide ΔH, ΔS, and ΔG NN parameters in 1M NaCl. [Reproduced with permission from [40] (Copyright 1998, National Academy of Sciences, U.S.A.)]. ΔG is computed at 37°C.

	ΔH	ΔS	ΔG
Interaction	kcal/mol	cal/K·mol	kcal/mol
AA/TT	-7.9	-22.2	-1.00
AT/TA	-7.2	-20.4	-0.88
TA/AT	-7.2	-21.3	-0.58
CA/GT	-8.5	-22.7	-1.45
GT/CA	-8.4	-22.4	-1.44
CT/GA	-7.8	-21.0	-1.28
GA/CT	-8.2	-22.2	-1.30
CG/GC	-10.6	-27.2	-2.17
GC/CG	-9.8	-24.4	-2.24
GG/CC	-8.0	-19.9	-1.84
Init. w/term. $G \cdot C$	0.1	-2.8	0.98
Init. w/term. $A \cdot T$	2.3	4.1	1.03
Symmetry correction	0	-1.4	0.43

$$\begin{aligned}
\Delta G &= \Delta G(CG/GC) + \Delta G(GA/CT) + \Delta G(AA/TT) + \Delta G(AG/TC) \\
&\quad + \Delta G(GT/CA) + \Delta G(\text{init. w}/G \cdot C) + \Delta G(\text{init. w}/A \cdot T) + 0 \\
&= -2.17 - 1.30 - 1.00 - 1.28 - 1.44 + 0.98 + 1.03 \\
&= -5.18 \text{ kcal/mol}
\end{aligned}$$

Since the duplex is not self-complementary, $\Delta G(\text{sym}) = 0$.

3.7.2 Affinity Constants

We further recall that at equilibrium, the affinity constants K_1^j are given by (3.21), replicated here for convenience, as described in section 3.3.1:

$$K_1^j = \exp\left[-\Delta G/RT\right],$$

and ΔG due to stacking interactions is calculated as above. Also, we note that with the affinity constant values computed, we are ready to compute the "ratios of perfect match to mismatch values" for a particular initial target and probe concentrations.

3.8 Observed Competition among Probes

As discussed in sections 3.3.1, 3.3.2, and 3.3.3, we can compute the equilibrium TP concentrations from the initial target and probe concentrations. We have computationally simulated the hybridization process for a large number of target/probe sequences used in practice, and observed a difference in pm/mm ratio for probe 1 under Partial Model ($P_1 + T$) vs. Full Model ($P_1 + P_2 + T$). A similar effect was observed for probe 2. These experiments indicated that the *direction* of the shift depends on the affinity constants and can be empirically characterized to be a function of the products of the affinity constants of the perfect match and mismatch probes.

For instance, we examined the behaviors of exon 11 probes A and B (treated as probes 1 and 2, respectively) under the full hybridization model, discussed in section 3.3.1, as well as under partial hybridization models (sections 3.3.2 and 3.3.3), to observe the following:

1. Ratio $[TP_{A,pm}]/[TP_{A,mm}]$ for A (i.e., probe pair $\{P_{A,pm}, P_{A,mm}\}$) shifts *up* in the presence of probe B (i.e., probe pair $\{P_{B,pm}, P_{B,mm}\}$).

2. Symmetrically, ratio $[TP_{B,pm}]/[TP_{B,mm}]$ for B shifts *down* in the presence of A.

3. We address the following questions: How can the shift direction be predicted? How does it depend on the sequences of the probe pairs in question?

3.8.1 Heuristic Development

Our empirical study was conducted as follows. Let us consider two probes, each having associated with it the pair $\{P_{\cdot,pm}, P_{\cdot,mm}\}$. For each probe, the pm/mm ratio shifts *up* or *down* in the presence of the other probe. The direction of the shift was determined to be a function of the relative sizes of the affinity constants K, where cross-bound states can be neglected. For a given probe, let $K_T\text{pm}$, $K_T\text{mm}$ denote the affinity constants for *This* probe's binding site with pm and mm, respectively; let $K_O\text{pm}$, $K_O\text{mm}$ be the *Other* probe's affinity constants with pm and mm.

Let us view the competition effect as a binary function on the space of affinity constants (+1 for *up*, −1 for *down* shift) and consider the projection of the affinity constant space

$$\mathbb{R}^4 = \{K_T\text{pm}, K_T\text{mm}, K_O\text{pm}, K_O\text{mm}\}$$

onto the plane $\mathcal{L}$ with axes $\log(K_T\text{pm}/K_O\text{pm})$ and $\log(K_T\text{mm}/K_O\text{mm})$. On this plane, the competition effect function values can be clearly separated by the line $x+y=0$. This condition holds for physical exon 11 probes, as shown in Figures 3.2 and 3.3.

The empirically determined condition can be described by the following logically

equivalent statements:

$$\begin{aligned} &\text{pm/mm ratio shifts } \textit{up} \\ &\iff \quad y < -x \\ &\iff \quad \log(K_T\text{mm}/K_O\text{mm}) < -\log(K_T\text{pm}/K_O\text{pm}) \\ &\iff \quad K_T\text{pm}\,K_T\text{mm} < K_O\text{pm}\,K_O\text{mm} \end{aligned} \tag{3.133}$$

Thus, the signal for *This* probe improves whenever (3.133) holds.

In order to test the heuristic computationally, we generated more points for the competition effect function by perturbing existing probe sequences in one base and pairing one actual exon 11 probe with one perturbed probe. The results of this empirical investigation of the competition effect on these probe pairs are graphically presented in Figure 3.4.

3.9 Experimental Validation

In order to further verify the performance of our heuristic, we proposed the following experiments. The pm/mm ratios should be measured for the probes as listed below under Partial model (i.e., the probe and its alternate are present alone with the target) and Full model (i.e., the specified probes, each with an alternate, are present with the target) for:

- Actual probe pairs:
  - AB, AC, AD, BC, BD, CD
  - In each case, both probes should be used alternately as *This* and as *Other* probe.

- Actual/Perturbed probe pairs that show a change of shift direction in simulation:

This	*Other*			
A	D_5A		D_5T	
D	A_2G	A_6G	A_8G	A_14C

- Actual/Perturbed probe pairs (to be used as controls) that do not show change:

This	*Other*	
A	D_5C	D_6T
D	A_2C	A_2T

The remarkable consistency with which our heuristic conforms with the results of the simulation suggests that the heuristic can be used reliably in place of the simulation to predict the competition effect, i.e., the direction of the shift. This predictive power can be used in experiment design (e.g., for HLA typing).

Example

Let $A = C381$, $B = A327$, and $C = D359$ from exon 11, with the alternates used in the experiments. Pairwise computational analysis indicated that: $A327$ improves the signal for $C381$ and $D359$ improves the signal for $A327$. Our heuristic implies that $D359$ automatically improves the signal for $C381$. This conclusion was tested using the extended model, as described in detail in section 3.6.2. Recall that the setup for this model includes three probes (each with an alternate) and three possible binding sites on the target for each probe; the "perfect match" for each

probe is designed to match the corresponding binding site on the target. In this example, we compared the ratio curves for the first probe from the Full and Partial models with the curve from the Extended model, as shown in Figure 3.5.

Note that, in Figure 3.5, the pm/mm ratio curve for $C381$ in the presence of both $A327$ and $D359$ (the blue curve) lies above both the red curve (the ratio for $C381$ in the presence of $A327$ alone) and the green curve (the ratio for $C381$ alone with the target). This indicates that for a given initial target concentration, i.e., a given point on the $x-$axis, the pm/mm ratio for $C381$ goes up in the presence of $A327$, which is consistent with pairwise analysis; the ratio increases further when $D359$ is added to the mix, confirming the heuristic prediction.

3.10 Conclusion

In this chapter we present mathematical models of the competitive probe-target hybridization process. Simulations based on the implementations of these models and the heuristic developed and presented in section 3.8.1 generate results that are in agreement with experimental results observed in the laboratory. Prediction of competition effects based on *in silico* experiments can be used for the design of better biological experiments. Possible applications include experiment design for genotyping and mutation analysis.

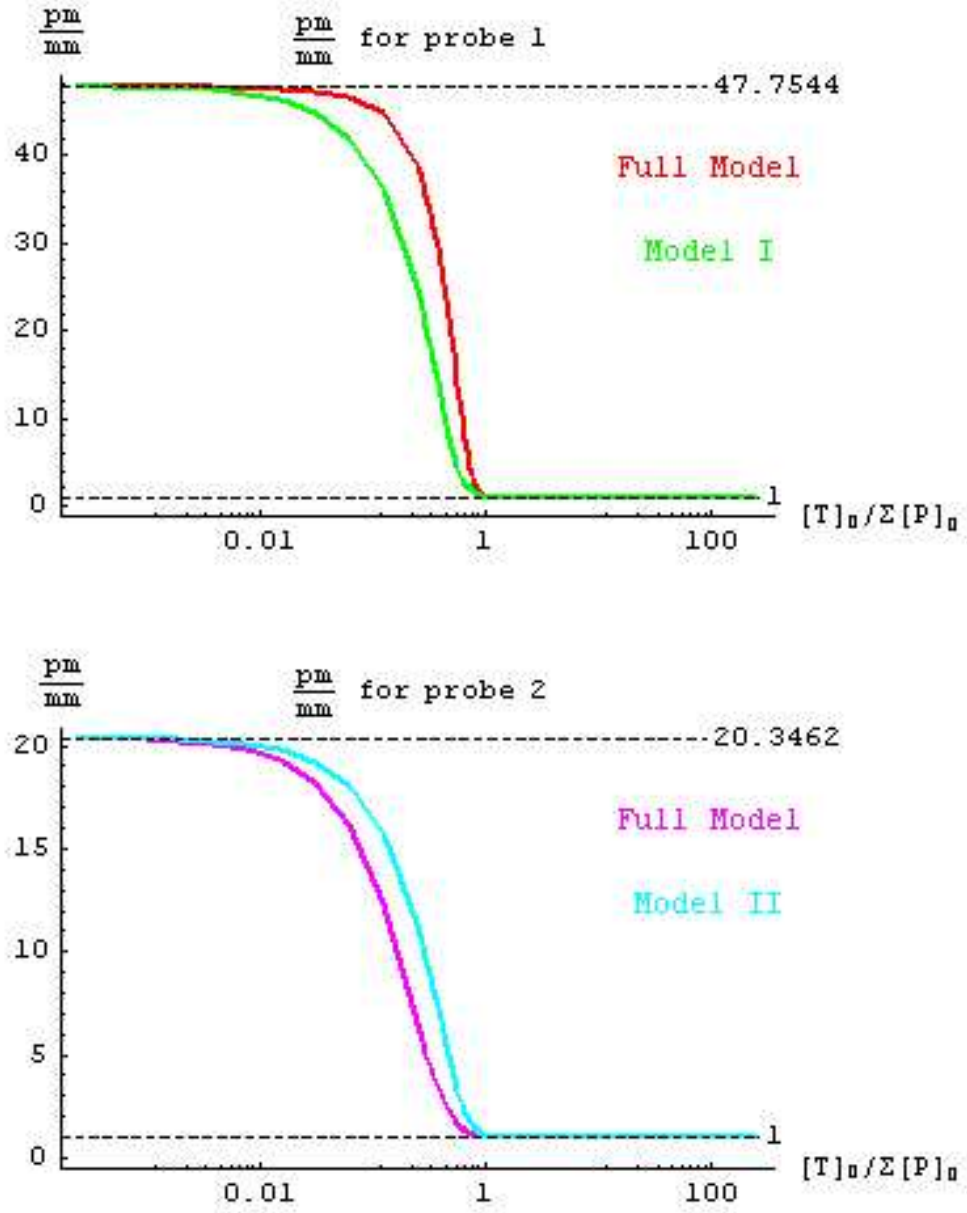


Figure 3.1: pm/mm ratios for probe A (top graph) and probe B (bottom graph), plotted against scaled initial target concentration.

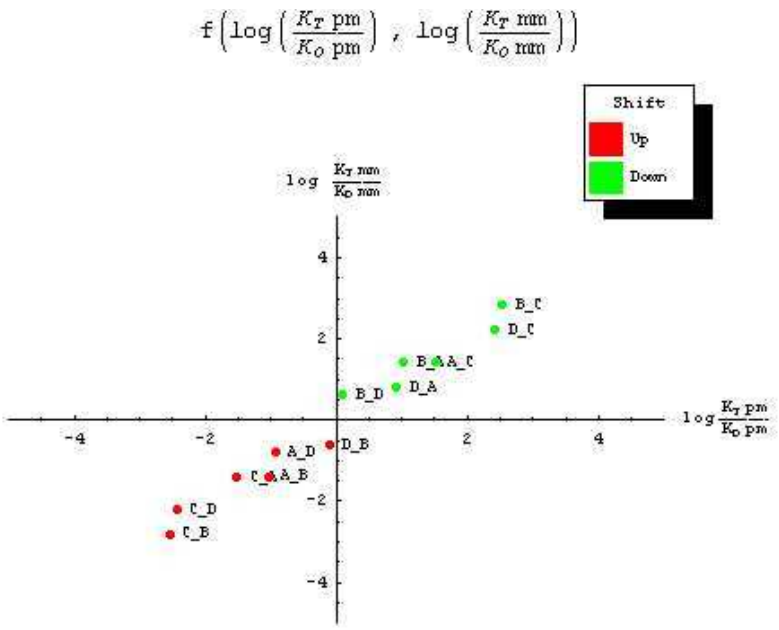


Figure 3.2: Competition effect binary function on exon 11 probes.

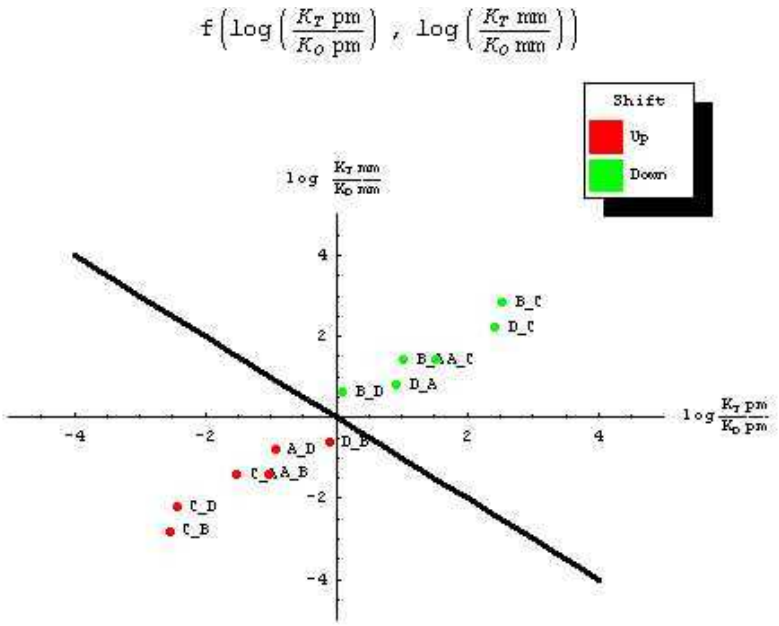


Figure 3.3: Competition effect binary function on exon 11 probes, shown with the separatrix $y = -x$.

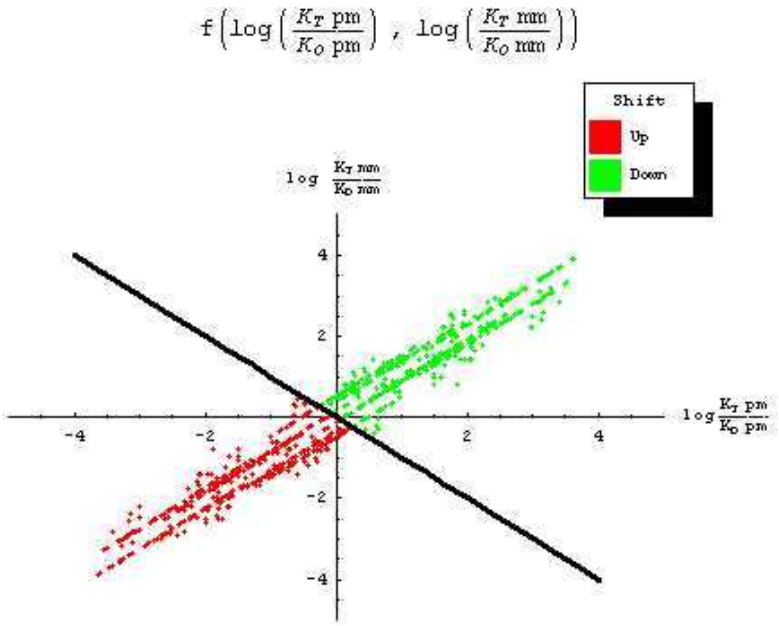


Figure 3.4: Testing the heuristic computationally: each probe pair contains one actual exon 11 probe and one perturbed probe.

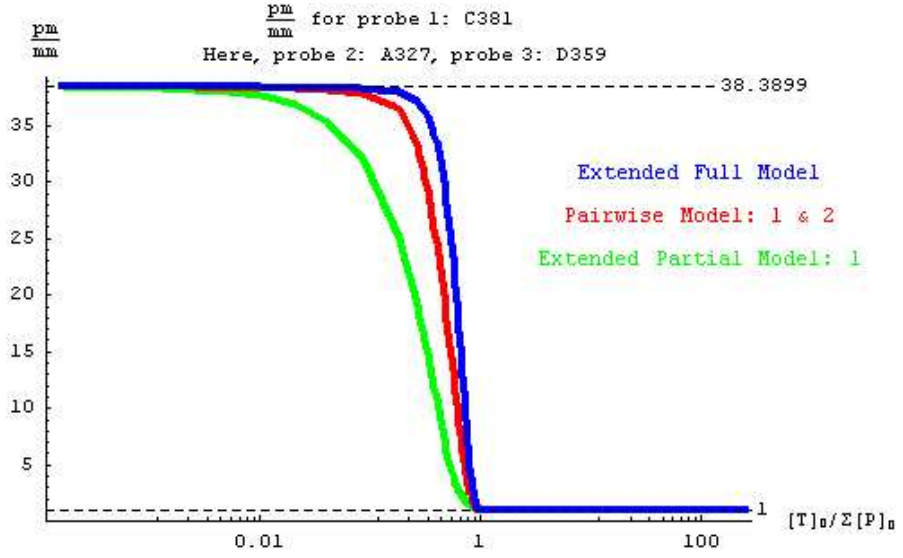


Figure 3.5: Example: pm/mm ratios for three probes

Chapter 4

HLA Typing

ABSTRACT

The problem of HLA typing has many biological implications; particularly, knowing the correct allele is essential to ensure the compatibility of the donor organ with the recipient. Most of the currently used techniques are time-consuming and lack optimality.

We present a graph model on the set of potential probes, formulate the HLA typing problem mathematically as an optimization problem on our graph model, and present an algorithm for solving the optimization problem. The processes of translating the typing problem to the graph model and the optimizing probe set back to the experiment design for HLA typing are described in detail.

Some experimental results on a simple example problem are presented. Extensions of our graph model to more detailed physical models are discussed.

4.1 Problem Definition

Human leukocyte antigen (HLA) region on chromosome 6 (e.g., [15]) is highly polymorphic—the sequence of this region varies from person to person. Many different possible sequences, or "alleles," are known. (Currently, close to 1,100 alleles are known.) Given a DNA sequence, the biological problem lies in determining which allele, or "HLA type," it contains.

The biological implications are many: the allele may predict the presence or absence of diseases, dictate the course of treatment for a patient, or, most notably, determine the compatibility of a potential transplant recipient with the donor organ or bone marrow. One of the approaches to finding the right allele is to design a microarray experiment that gives the allele as an answer. In fact, HLA typing by sequence hybridization with sequence-specific oligonucleotide probes (SSOP) is currently practiced by the National Marrow Donor Program (NMDP) for donor-recipient matching ([34]), alongside with the more traditional serology method ([35], [11]). In many current methods utilizing a popular test format, the DNA samples to be classified are amplified with locus specific primers and spotted onto the microarray chips, resulting in multiple copies of identical chips; each chip is then hybridized to a different probe ([6], [15]). This methodology necessitates a new design process every time a new set of patient samples must be classified. Some of the recent work done in this area is described in [36], [17], [22], and [15].

In the approach proposed here, the sequence-specific probes will be placed on

the microarray chip, and each patient sample will be applied to the chip to allow hybridization with some of the chip-bound probes. With properly selected probes, the same chip can be used for all classifications. Thus, we must design a set of probes to be used in a series of hybridization experiments on the target sequence. How many probes to use, and what their sequences should be, are design questions.

Of course, a general solution should be one that allows us to "recognize" all existing alleles, or decide that the given DNA sequence contains an allele that is not in the "known" list. Such an allele may be a new, previously unknown allele, or one of the very rare alleles that occur so infrequently that they are not considered HLA types.

4.1.1 Mathematical Formulation

Definitions

Let us denote the different HLA types, or alleles, by T_j, $j = 1, \ldots, N$. (Here, we set $N = 1100$.) Let a given microarray be denoted by μA_k, $k \in \mathbb{N}$, where a microarray is defined by a set of hybridization probes and their two-dimensional arrangement on the chip surface. The process of querying the given DNA sequence (hereafter referred to as a "target" sequence) by a hybridization experiment can be denoted by the expression

$$(T_j, \mu A_k) \longrightarrow \mathcal{D} \longrightarrow \widehat{T}_j, \tag{4.1}$$

where T_j is the true allele contained in the target sequence, μA_k is the microarray used in the query, $\mathcal{D}$ is the data output of the hybridization experiment, and $\widehat{T}_j$ is the allele inferred from the data. Both processes in (4.1) are described below.

The problem of HLA typing can then be formulated as that of designing *the best microarray*, namely, the set and arrangement of probes, which "works" for *all* known HLA types (i.e., $\forall j$). In our notation, this means finding μA_k which solves the optimization problem

$$\begin{aligned} & \min \sum_{\text{type } j} w_j \ \mathbf{E}\left[\mathbb{I}_{T_j \neq \widehat{T_j}}\right] \\ \iff \quad & \min \sum_{\text{type } j} w_j \ \Pr\left(T_j \neq \widehat{T}_j\right). \end{aligned} \tag{4.2}$$

Here, $\mathbb{I}_X$ is the indicator function

$$\mathbb{I}_X = \begin{cases} 1, & \text{if } X \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

and w_j is the weight assigned to type j. Initially, we set $w_j = 1 \ \forall j$. Later, it may be desirable to weigh different HLA types differently, based on the frequency of their occurrence in human population or some other criteria.

There are several processes that need to be considered and described in detail: obtaining data $\mathcal{D}$ from an experiment based on allele T_j and microarray μA_k, inferring allele $\widehat{T}_j$ from the data, generating potential microarrays for the typing experiments, and selecting the optimal microarray.

$(T_j, \mu A_k) \longrightarrow \mathcal{D}$

Consider the set of probes $\{P_{1,k}, \ldots, P_{n(k),k}\}$ constituting microarray μA_k, neglecting their arrangement for the moment. Ideally, the outcome of the hybridization of the target sequence with each probe P would be binary: 1, if the target contains a subsequence complementary to P, and 0, otherwise. Using $n = n(k)$ probes then

yields a binary string of length n, or, alternately, a vector of length n, as a code for the target sequence.

In practice, hybridization results are not binary; rather, the measurements are the analog intensity values corresponding to the amount of formed probe-target complex for each probe. In addition, in an attempt to "factor out" the non-specific signal, each probe is often present in two versions: one (perfect match, or "pm") perfectly complementary to a region on the target, and the other (mismatch, or "mm") slightly mismatched, the latter usually containing a single base mismatch near the center of the probe. This is the case, for example, in Affymetrix GeneChips ([24]). In such a setup, the signal from probe P is the match-to-mismatch ratio: the ratio of the intensities corresponding to the matched and mismatched probe-target complexes. Furthermore, the signal is log-transformed, so that the hybridization outcome for probe P is really the value of

$$\log\left(\frac{TP_{pm}}{TP_{mm}}\right).$$

The situation is further complicated by the fact that probes may hybridize to positions on the target other than those they were designed to detect—this is known as "cross-hybridization." In addition, the fact that many probes are present in the system may cause the signal (i.e., hybridization outcome) from a given probe to differ from the signal of the same probe in the absence of other probes. This topic was addressed in more detail in Chapter 3.

Thus, the actual result of a hybridization experiment is a vector of n measurements, $\mathcal{D} \in \mathbb{R}^n$.

$\mathcal{D} \longrightarrow \widehat{T_j}$

The next question is, how can this n-vector be used to infer the allele?

First, let us return to the ideal process, where the outcome is $\mathcal{D} \in \{0,1\}^n$. If the probes were chosen in such a way as to give a *distinct* binary string for each known allele (so that the Hamming distance d_H between any pair of data vectors $\mathcal{D}_i, \mathcal{D}_k$ is at least 1), then these n probes are sufficient to identify the allele of the target sequence. All one has to do is query the sequence with the n probes and read off which allele the pattern corresponds to. Furthermore, if we require $d_H(\mathcal{D}_i, \mathcal{D}_k) \geq \alpha$ for some $\alpha > 1$, the discrimination power can be increased and error-correction is possible. This issue is discussed in more detail in section 4.2.2.

In the practical setting, $\mathcal{D} \in \mathbb{R}^n$. Thus, as a first step, some thresholding process must be applied to $\mathcal{D}$ to reduce it to a binary string.

Generating Potential Microarrays

Choosing Informative Probes We must provide a set of n probes, each of length L, that are at least d letters apart (pairwise), for optimal discrimination among the allele sequences.

If L is not specified, we can choose it arbitrarily (say, 20), or allow it to vary from probe to probe.

With no restrictions, we could choose a very large n: for example, why not use every possible 20-mer as a probe? This would result in $4^{20} = 2^{40} = (2^{10})^4 > (10^3)^4 = 10^{12}$, or over a trillion, probes. Such a large set is not desirable, since many of these probes would give the same results, and it is too expensive to produce all of them. Allowing both n and L to vary would give us an even larger number of

potential probe sequences.

The essence of the probe design problem lies in choosing which of the probes are most useful in discriminating among the given allele sequences, and how many (or rather, how few) we can get away with using.

Arranging Probes on the Chip Once a set of probes has been selected using techniques described in earlier sections, there is still a question of how to arrange these probes on the microarray chip. Several studies indicate that the patterns observed in the results of chip experiments may be due to the arrangement of probes on the chip (see [27], [49], and [38]). Specifically, it has been observed that probes are arranged on a chip based on the labels of the genes they represent, and a gene label is often related to the function and/or disorder the gene is involved in. As a result, genes of shared function have similar labels and are coexpressed, generating monochromatic bands on microarray chip scans.

These studies suggest that more thought should be given to the arrangement of probes on the chip, based on some "conceptual" measure of probe distance. The "conceptual" probe distance can use available biological knowledge about the portions of the genome containing the probes in question, as well as a measure of competition between these probes, as discussed in section 3.8. The following recursive technique, described in [32], can ensure that the "nearest" probe pairs are separated by at least a specified minimum physical distance on the chip. It is designed to be disruptive to neighborhoods, defined with respect to the "conceptual" probe distance; thus, it places conceptually nearby probes far apart on the surface.

Introduction Consider a bijective function

$$f : \{0, \ldots, N^2 - 1\} \to \{0, \ldots, N - 1\} \times \{0, \ldots, N - 1\}$$

that maps every pair of "nearby" points in the domain space to a pair of "distant" points in the range space. In particular, we devise a function f with the following property: For every x, y, if $|x - y| \leq 4^\alpha$, then $\|f(x) - f(y)\|_1 \geq N/(2^{\alpha+1})$. We conjecture (but do not have a proof) that this function gives an optimal placement. If the elements of the domain space satisfy other distance properties, this method can be suitably generalized to handle similar properties with respect to the new distance metric.

This function plays an important role in determining how to place a set of oligonucleotide probes on a microarray surface in such a manner that if two probes are close to each other in their genome locations then they are reasonably far apart on the array. Thus, a placement determined by the function minimizes competition among the probes for the genomic targets as well as the systematic biases in the error processes.

Function Definition Inductively, we define a uniform family of functions f_k as follows. Let $k < \lg N$.

$$\begin{aligned} f_{k+1} &: \{0, \ldots, N^2 - 1\} \to \{0, \ldots, N - 1\} \times \{0, \ldots, N - 1\} \\ &: x \mapsto \langle i, j \rangle. \end{aligned}$$

f_{k+1} is defined in terms of f_k

$$f_k : \{0, \ldots, N^2/4 - 1\} \to \{0, \ldots, N/2 - 1\} \times \{0, \ldots, N/2 - 1\}$$

as follows:

$$f_{k+1}(x) = \begin{cases} f_k(\lfloor \frac{x}{4} \rfloor), & \text{if } x \equiv 0 \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle 0, \frac{N}{2} \rangle, & \text{if } x \equiv 1 \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle \frac{N}{2}, 0 \rangle, & \text{if } x \equiv 2 \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle \frac{N}{2}, \frac{N}{2} \rangle, & \text{if } x \equiv 3 \bmod 4. \end{cases} \tag{4.3}$$

This function can be generalized, without its general properties being affected, by simply including a random permutation $\pi_{k+1} : \{0, \ldots, 3\} \to \{0, \ldots, 3\}$ as follows:

$$f_{k+1}(x) = \begin{cases} f_k(\lfloor \frac{x}{4} \rfloor), & \text{if } x \equiv \pi_{k+1}(0) \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle 0, \frac{N}{2} \rangle, & \text{if } x \equiv \pi_{k+1}(1) \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle \frac{N}{2}, 0 \rangle, & \text{if } x \equiv \pi_{k+1}(2) \bmod 4; \\ f_k(\lfloor \frac{x}{4} \rfloor) + \langle \frac{N}{2}, \frac{N}{2} \rangle, & \text{if } x \equiv \pi_{k+1}(3) \bmod 4. \end{cases}$$

Hereafter, k takes the value $(\lg N - 1)$ and the base case is given by the function

$$f_2 : \{0, \ldots, 15\} \to \{0, \ldots, 3\} \times \{0, \ldots, 3\}$$

where

$$\begin{array}{llll} 0 \mapsto \langle 0,0 \rangle, & 1 \mapsto \langle 0,2 \rangle, & 2 \mapsto \langle 2,0 \rangle, & 3 \mapsto \langle 2,2 \rangle \\ 4 \mapsto \langle 0,1 \rangle, & 5 \mapsto \langle 0,3 \rangle, & 6 \mapsto \langle 2,1 \rangle, & 7 \mapsto \langle 2,3 \rangle \\ 8 \mapsto \langle 1,0 \rangle, & 9 \mapsto \langle 1,2 \rangle, & 10 \mapsto \langle 3,0 \rangle, & 11 \mapsto \langle 3,2 \rangle \\ 12 \mapsto \langle 1,1 \rangle, & 13 \mapsto \langle 1,3 \rangle, & 14 \mapsto \langle 3,1 \rangle, & 15 \mapsto \langle 3,3 \rangle \end{array} \tag{4.4}$$

This base map[1] can be described in matrix format as follows:

$$\begin{bmatrix} 0 & 4 & 1 & 5 \\ 8 & 12 & 9 & 13 \\ 2 & 6 & 3 & 7 \\ 10 & 14 & 11 & 15 \end{bmatrix} \tag{4.5}$$

[1]Suggested by Iuliana Ionita.

Taking $N = 2^\ell$, we can place $N^2 = 4^\ell$ probes by applying $f_\ell : \{0, \ldots, 4^\ell - 1\}$, which after $(\ell - 2)$ recursive steps, defined in (4.3), reduces to the base case $f_2 : \{0, \ldots, 15\}$ shown in (4.4), (4.5).

Function Properties Function f_ℓ has the following distance properties. Let $D(i, j)$ be the distance between probes p_i and p_j, when arrayed on a line (by relabeling the probes, we can view this as the index separation $|i - j|$). Let $d(i, j)$ be their distance when arrayed on the surface. Then the mapping f_ℓ guarantees that for all p_i, p_j for which $D(i, j) \leq 4^k$, $d(i, j) \geq 2^{\ell-k-1}$, where $k = 0, \ldots, \ell - 1$. Furthermore, if $d(i, j) = 1$, that is, p_i is placed next to p_j on the surface, then $D(i, j) \geq 3 \cdot 4^{\ell-2}$.

Example:

$$\left[\begin{array}{cccc|cccc}
0 & 16 & 4 & 20 & 1 & 17 & 5 & 21 \\
32 & 48 & 36 & 52 & 33 & 49 & 37 & 53 \\
8 & 24 & 12 & 28 & 9 & 25 & 13 & 29 \\
40 & 56 & 44 & 60 & 41 & 57 & 45 & 61 \\
\hline
2 & 18 & 6 & 22 & 3 & 19 & 7 & 23 \\
34 & 50 & 38 & 54 & 35 & 51 & 39 & 55 \\
\hline
10 & 26 & 14 & 30 & 11 & 27 & 15 & 31 \\
42 & 58 & 46 & 62 & 43 & 59 & 47 & 63
\end{array}\right]$$

Here, $\ell = 3$, so we have $N^2 = 4^\ell = 64$ probes to place.

k	D	d
0	1	4
1	4	2
2	16	1

If $D = 4^k$, then $d = N/2^{k+1} = 2^{\ell-k-1}$.

4.1.2 Related Problems

Some other biological problems, such as identifying an unknown pathogen as a member of a list of known pathogens, be they viral ([39]) or bacterial ([8]), have the

same mathematical formulation as the problem of HLA typing discussed here. We believe that those applications can also benefit from the improvements to existing approaches provided by the work presented here.

The problem of automatic generation of probe sets for DNA microarrays was also addressed recently in [28]. However, the work described in [28] aims for a probe set that is, even in ideal circumstances, asymptotically much larger than the one generated by our approach.

4.2 Optimization Problem on a Graph Model

We now address the problem of selecting the optimal microarray. The problem of choosing the constituent probes can be reduced to a "best independent set" problem. The following sections define the graph model we use, as well as the meaning of the term "best independent set," and describe the optimizing algorithm.

4.2.1 Graph Model Definitions

Notation

Let us begin by introducing the notation. Recall that there are N known alleles, and suppose we have n potential probes. Each probe is described by a "response vector" $\vec{v}_j \in \{0,1\}^N,\ j = 1, \ldots, n$. The response vector data can be represented in

tabular form:

$$\begin{array}{cccccc} & \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ \text{HLA}_1 & 1 & 0 & \cdots & 1 \\ \text{HLA}_2 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \text{HLA}_N & 0 & 0 & \cdots & 1 \end{array} \tag{4.6}$$

Here, column j is the response vector for probe $\vec{v}_j$:

$$\vec{v}_j = (v_j[1], v_j[2], \ldots, v_j[N])^T, \tag{4.7}$$

and row i is the code for allele i, which we will call HLA_i:

$$\text{HLA}_i = (v_1[i], v_2[i], \ldots, v_n[i]). \tag{4.8}$$

Original Graph

Let each potential probe form a vertex in the graph. Conceptually, an edge in the graph should connect two probes with shared characteristics.

We start with essentially a complete edge- and vertex-weighted undirected graph $G = (V, E)$ on n vertices, where n is the number of potential probes. In the most general problem formulation, n can be very large: for each probe length L, there are 4^L possible probes. (For instance, as shown in section 4.1.1, there are over a trillion possible probes of length 20.) Thus, the graph in our probe interaction model can be very large.

We assign weights to each vertex v and to each edge e, $0 \leq w(v), w(e) \leq 1$. The weight of a vertex is initially set to the "information content"[2] of the corresponding

[2]The term "information content" is used here differently than defined in information theory, where it means the minimum amount of information needed to send a string (e.g., [16]).

probe response vector with respect to the HLA typing problem:

$$w(v) = \min\{\%0\text{'s}, \%1\text{'s}\}/100. \tag{4.9}$$

Ideally, if possible, we would like all vertices to have weight 0.5; a vertex with weight too close to zero is uninformative, and the corresponding probe should only be used if it serves to differentiate an allele that is not distinguishable by using other, more informative probes.

The weight of an edge is initially set to the scaled Hamming distance of the probe response vectors represented by its endpoints:

$$w(e) = \text{Hamming distance/vector length}, \tag{4.10}$$

with values close to zero corresponding to sequence-similar probes.

Thresholded Graph

Next, we transform our graph G by thresholding the edges. We select a threshold ρ, the choice of which is discussed in more detail in section 4.2.3, and generate a modified graph $G_{mod} = (V, E_{mod})$, where

$$E_{mod} = \{e \in E : w(e) \leq \rho\}$$

is a set of unweighted edges, and the set of weighted vertices V is unchanged. Hereafter, we work with this modified vertex-weighted graph and denote it by G.

It now makes sense to define an independent set on our graph. An *independent set* is defined as a set of vertices such that for any pair of vertices, there is no edge between them ([19]). More formally, a set of vertices $V' \subset V$ is an independent set if ($u \in V'$ and $v \in V'$) implies that $\{u, v\} \notin E$.

Goal

In section 4.1.1, we defined *the best microarray.* Now we formulate the corresponding concept on our graph model. We define *the best independent set* as a maximum weight yet minimum size independent set. Thus, such a set $S \subset V$ must be an independent set, have maximum weight $w(S) = \sum_{v \in S} w(v)$, and minimum cardinality $|S|$. The condition of independence is meant to preclude any unintended interaction among the chosen probes. Maximum weight will provide S with maximum discrimination power. Minimum size will ensure that we use the smallest collection of probes that does the job.

Since all vertex weights are nonnegative, the requirements of maximum weight and minimum cardinality are clearly contradictory. We can relax the definition somewhat by specifying *a priori* the desired size M of the set, and look instead for the maximum weight independent set of size $\leq M$.

4.2.2 Optimization Algorithm

To achieve this goal, we used a modification of a Maximal Independent Set algorithm, described in a classic paper by Luby ([31]).

Algorithm Pseudocode

Given graph G and set size M:

1. Initialization:

   (a) Initialize a "current-best" list of independent sets, with associated information weights. It will store a list of the best, say, 20, independent sets

seen so far, sorted by information weight.

2. Restart Loop: Execute at least *minRestartNum* times; if the "current-best" list is not full (i.e., does not have 20 independent sets) by then, keep repeating until the list is filled.

   (a) Initialize boosting weights; we accomplish this by setting the boosting weights to the information weights of the vertices:

   $$\forall v \in V,\ w_b(v) \leftarrow w(v).$$

   (b) Boosting Loop: Repeat until no improvements have been made to the "current-best" list for a fixed number of iterations (say, 5 iterations).

      i. Choose a set S of M vertices randomly from V, with

      $$P(v \in S) = \frac{w_b(v)}{\sum_{u \in V} w_b(u)}.$$

      ii. For each edge $\{u, v\}$ in $G|_S$ (the induced subgraph on S), eliminate one of the endpoint vertices. This leaves a set, S_1, of $K \leq M$ independent vertices.

      iii. Adjust the boosting weights of vertices in S; namely, increase the boosting weights of the vertices in S_1:

      $$w_b(v) \leftarrow a\, w_b(v) \quad \forall v \in S_1$$

      and decrease the boosting weights of the vertices in $S - S_1$:

      $$w_b(v) \leftarrow \frac{1}{a}\, w_b(v) \quad \forall v \in S - S_1,$$

      where $a \geq 1$ is some previously selected constant.

iv. If S_1 is not already in the "current-best" list and provides an improvement over some current member of the list, reset the *noImprovements* counter, find the appropriate location for S_1 in the list, and update the list. Otherwise, make a note that no changes to "current-best" list were made on this iteration (i.e., increment the *noImprovements* counter).

(c) If the condition for continuing the restart loop holds (namely, *minRestartNum* restarts have not yet been executed or the "current-best" list is not yet full), reset the *noImprovements* counter and repeat step 2.

Algorithm Description

Our algorithm uses vertex boosting weights (initially set to probe information weights) to define a probability distribution on the vertex set. On each iteration of the boosting loop (step 2b), a random subset of a specified size is chosen according to the current probability distribution (step 2(b)i). All edges in the induced subgraph on this random subset are broken, with one of the terminal vertices thrown out (step 2(b)ii). The boosting weights of the elements of the subset are then modified (step 2(b)iii), so that the vertices that stayed in the subset are more likely, and the vertices that were thrown out are less likely to be chosen on the next iteration. The boosting loop terminates after a certain number of iterations with no improvement to the list of top independent sets (to allow some flexibility, the algorithm keeps track of several of the top independent sets instead of only storing the best one seen so far).

The algorithm also restarts the boosting loop several times with original probe

information weights. This feature has been incorporated to prevent convergence to a local optimum, which is possible for high values of the boosting factor.

What Does It Mean?

We can think of the boosting algorithm (as a whole) as operating on the probability space of all subsets of our graph. Note that step 2(b)ii guarantees that the selected subset is independent, so that the probability distribution is only supported on independent sets (i.e., the distribution is zero on all non-independent sets). In this view, we can expect the algorithm to converge to a probability distribution where the *best independent set* has the highest probability. Each iteration of the boosting loop adjusts the probabilities associated with each vertex in the graph. The subset of interest is always drawn randomly according to the current probability distribution.

If the solution S^* were known *a priori*, its selection by the algorithm could be guaranteed by initializing the boosting weights is step 2a to be

$$\begin{aligned} \forall v \in S^*, \qquad & w_b(v) \leftarrow 1, \\ \forall v \in V - S^*, \qquad & w_b(v) \leftarrow 0. \end{aligned}$$

In other words, the associated probability distribution would have a probability of 1 for each vertex $v \in S^*$, and a probability of 0 for each remaining vertex $v \in V - S^*$.

Given unlimited time for obtaining a solution (and an appropriate set of parameters), we would expect the boosting algorithm to converge to this ideal distribution. However, when time is limited, a "good" (i.e., informative) independent set of size $\leq M$ is only "more likely" than other independent sets of similar size. The algo-

rithm presented here was designed to give a "pretty good" solution in limited time, yet be able to improve on it iteratively when more time is permitted (with minimal modifications to the loop terminating conditions).

Another way to think about it is that the best independent set is, ideally, a fixed point for our algorithm, in the sense that if the algorithm starts at a perturbed location in the subset probability space, it should converge to the optimal set. That is, if the initial probability distribution is heavily favored towards a set that doesn't differ from the best set in many vertices, the algorithm should converge to the best set.

Breaking Edges

In step 2(b)ii of the boosting algorithm, the mechanism for breaking the edge $\{u, v\}$ was not specified. The implemented approach was the following: choose to keep the vertex that has the higher *boosting* weight; if vertices have equal boosting weights, choose one at random (with probability $\frac{1}{2}$).

Choosing Scaling Factor a

In step 2(b)iii of the boosting algorithm, the weights of those vertices that were selected and kept are boosted (scaled up) by a factor of $a \geq 1$, while the weights of discarded vertices are scaled down by the same factor. This has the effect of noting which vertices were chosen for membership in the independent set and increasing the likelihood that these vertices will be chosen in the future, with the reverse effect on the discarded vertices. Here, we discuss how the value of the scaling factor a affects the "memory" of the probability space evolution. Let us examine a single

"restart" of the algorithm (namely, step 2b).

Extreme cases:

- $a = 1$: No memory of previous selections. Ignore the current selection, and choose anew on the next iteration. The boosting algorithm performs an exhaustive search.

- $a = \infty$: Perfect memory. Once a set S of vertices is selected and pruned, and its elements' boosting weights are modified, each of the vertices remaining in the independent set S_1 will have a boosting weight of ∞ and each of those thrown out of the independent set due to conflicts will have a boosting weight of 0. Thereafter, the boosting algorithm will always choose the independent set selected on the first run.

Real values:

The boosting algorithm was executed on the same graph model with several values of $a \in \{2, 1.5, 1.2, 1.1\}$. The executions with higher values of a were observed to terminate a single "restart" after a smaller number of iterations than those with lower values of a.

Choosing M: the Maximum Size of the Independent Set

This section contains a probabilistic analysis of the answer to the following question: What are the bounds on the number of probes, k, that is sufficient to distinguish N known alleles? In order to answer this question, certain assumptions are made on the random distribution from which the known alleles are assumed to be drawn.

Independent Probes Recall the notation introduced in section 4.2.1. Assume that each probe, at each index $i = 1, \ldots, N$, assumes values 0 and 1 independently and with equal probability. Consider k such probes and two alleles (HLA_l and HLA_m). Thus, if HLA_l is fixed:

$$\text{HLA}_l = (\text{HLA}_l[1], \ldots, \text{HLA}_l[k]),$$

then for each j we have

$$\begin{aligned} \Pr\left(\text{HLA}_m[j] = \text{HLA}_l[j]\right) &= \frac{1}{2} \\ \Pr\left(\text{HLA}_m[j] \neq \text{HLA}_l[j]\right) &= \frac{1}{2} \end{aligned}$$

Then for these two HLA vectors,

$$\Pr\left(\text{The Hamming dist bet'n 2 HLA vectors} = x\right) = \binom{k}{x} 2^{-k}, \tag{4.11}$$

which can easily be seen as follows:

$$\begin{aligned} &\Pr\left(\text{The Hamming dist bet'n 2 HLA vectors} = x\right) \\ &= \Pr\left(\text{HLA vectors differ in exactly } x \text{ positions}\right) \\ &= \Pr\left(\begin{array}{l} x \text{ successes in } k \text{ Bernoulli trials,} \\ \text{where} \\ \text{success} = \{\text{HLA}_m[j] \neq \text{HLA}_l[j]\} \\ \text{and } p = \Pr\left(\text{success}\right) = \frac{1}{2} \end{array}\right) \\ &= \binom{k}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{k-x} = \binom{k}{x} 2^{-k} \end{aligned}$$

Thus, for a fixed pair of alleles,

$$\begin{aligned} \Pr\left(x \geq 1\right) &= 1 - \Pr\left(x = 0\right) = 1 - \binom{k}{0} 2^{-k} \qquad \text{(by (4.11))} \\ &= 1 - 2^{-k}, \end{aligned} \tag{4.12}$$

and

$$\begin{aligned}
\Pr\left(\forall_{\text{pairs}}\ x \geq 1\right) &= \prod_{\text{pairs}} \Pr\left(x \geq 1\right) \qquad \text{(by independence)} \\
&= \prod_{\text{pairs}} (1 - 2^{-k}) = (1 - 2^{-k})^{\#\text{ pairs}} \qquad \text{(by (4.12))} \\
&= (1 - 2^{-k})^{\binom{N}{2}} \qquad (4.13)
\end{aligned}$$

since there are N distinct allele vectors and pairs are unordered.

We wish this probability to be bigger than $(1-\epsilon)$ for some fixed small $0 < \epsilon \ll 1$, i.e.,

$$(1 - 2^{-k})^{\binom{N}{2}} \overset{\text{want}}{>} 1 - \epsilon. \qquad (4.14)$$

First, let us bound the left-hand side term:

$$\begin{aligned}
(1 - 2^{-k})^{\binom{N}{2}} &= \left[\left(1 - 2^{-k}\right)^{2^k}\right]^{\binom{N}{2}2^{-k}} > \left(e^{-1-2^{-k}} \right)^{\binom{N}{2}2^{-k}} \\
&= e^{-\binom{N}{2}(1+2^{-k})\,2^{-k}}
\end{aligned}$$

where the inequality comes from the bound (appendix C.1)

$$\left(1 - \frac{1}{n}\right)^n > e^{-1-\frac{1}{n}} \qquad \text{for large } n. \qquad (4.15)$$

For the inequality in (4.14) to work, the bound (4.15) has to be in the correct direction. Suppose we want $a > b$. If we show instead $a > c$ and then choose the parameter so that $c > b$, we can conclude by $a > c > b$ that $a > b$. Therefore, the above inequality chain will work if (4.15) holds.

Hereafter, we will use the symbol $\Longleftarrow$ to indicate steps in the inequality reduction that will satisfy the previous statements whenever the parameter in question is chosen to satisfy the current statement.

Thus, we reduced inequality (4.14) to the following:

$$\begin{aligned} & e^{-\binom{N}{2}(1+2^{-k})\,2^{-k}} > 1-\epsilon \\ \iff \quad & -\binom{N}{2}(1+2^{-k})\,2^{-k} > \ln(1-\epsilon) \end{aligned} \tag{4.16}$$

Next, consider the right-hand term: $\ln(1-x) < -x$ for $0 < x < 1$. Again, we want $a > b$. If we show $b < d$ and then choose the parameter so that $a > d$, we can conclude by $a > d > b$ that $a > b$. This allows us to reduce inequality (4.16) to the following:

$$-\binom{N}{2}(1+2^{-k})\,2^{-k} > -\epsilon \tag{4.17}$$

$$\iff \quad \epsilon > \binom{N}{2}(1+2^{-k})\,2^{-k}$$

$$\iff \quad \frac{4^k}{2^k+1} > (1/\epsilon)\binom{N}{2}, \tag{4.18}$$

since

$$(1+2^{-k})\,2^{-k} = (2^k+1)\,2^{-2k} = (2^k+1)\,4^{-k}. \tag{4.19}$$

Furthermore,

$$\frac{\beta^2}{\beta+1} = \frac{\beta^2+\beta-\beta-1+1}{\beta+1} = \beta - 1 + \frac{1}{\beta} > \beta - 1 \qquad \forall \beta > 0. \tag{4.20}$$

Hence, taking $\beta = 2^k$ yields

$$\frac{4^k}{2^k+1} > 2^k - 1 \tag{4.21}$$

Thus, (4.18) follows if k is chosen to satisfy

$$\begin{aligned} & 2^k - 1 > (1/\epsilon)\binom{N}{2} \\ \iff \quad & \boxed{2^k > (1/\epsilon)\binom{N}{2} + 1} \end{aligned} \tag{4.22}$$

We can now verify the remaining chain of inequalities, from "desired" to "obtained": inequality (4.17) can be extended

$$-\binom{N}{2}(1+2^{-k})\,2^{-k} > -\epsilon > \ln(1-\epsilon)$$
$$\implies \quad e^{-\binom{N}{2}(1+2^{-k})\,2^{-k}} > 1-\epsilon,$$

which in turn can be extended

$$(1-2^{-k})^{\binom{N}{2}} > \; e^{-\binom{N}{2}(1+2^{-k})\,2^{-k}} > 1-\epsilon$$
$$\implies \quad (1-2^{-k})^{\binom{N}{2}} > 1-\epsilon, \qquad \text{as desired.}$$

Therefore, we must choose k (given ϵ, N) to satisfy (4.22):

$$2^k > (1/\epsilon)\binom{N}{2} + 1$$

A simpler bound on k can be obtained by imposing a stronger condition

$$2^k \overset{\text{want}}{>} (2/\epsilon)\binom{N}{2}, \tag{4.23}$$

which implies (4.22) since $(1/\epsilon)\binom{N}{2} > 1$. The right-hand side of (4.23) simplifies to

$$(2/\epsilon)\binom{N}{2} = (2/\epsilon)\frac{N(N-1)}{2} = (1/\epsilon)N(N-1),$$

so that

$$k > \lg N + \lg(N-1) + \lg(1/\epsilon) \tag{4.24}$$

is equivalent to (4.23). Furthermore, since $2\lg N > \lg N + \lg(N-1)$, choosing

$$k > 2\lg N + \lg(1/\epsilon)$$

certainly gives us a value of k that satisfies (4.23). Therefore, requiring

$$\boxed{k > 2\lg N + \lg(1/\epsilon)} \tag{4.25}$$

imposes the *strongest* condition of those listed above. Hence, a value of k that satisfies (4.25) also satisfies (4.22), and therefore the original desired inequality (4.14).

Dependent Probes Since the independence assumptions may be violated, we model this by an error term δ. Suppose a probe fails to contribute to a Hamming distance with probability $(1+\delta)/2$. As before, we consider each position of the HLA code vector as a Bernoulli trial, where success is defined as the event that j^{th} entry of a code vector contributes to the Hamming distance, i.e., $\{\text{HLA}_m[j] \neq \text{HLA}_l[j]\}$, so that

$$\begin{aligned} q &= \Pr(\text{failure}) = (1+\delta)/2 \\ p &= \Pr(\text{success}) = (1-\delta)/2 \end{aligned}$$

Therefore,

$$\begin{aligned} &\Pr(\text{The Hamm dist } = x) \\ &= \binom{k}{x}\left(\frac{1-\delta}{2}\right)^x\left(\frac{1+\delta}{2}\right)^{k-x} \\ &= \binom{k}{x}(1-\delta)^x(1+\delta)^{k-x}2^{-k} \end{aligned} \tag{4.26}$$

Continuing as before, we derive the condition.

$$\begin{aligned} &\Pr(x \geq 1) = 1 - \Pr(x = 0) \\ &= 1 - \binom{k}{0}(1-\delta)^0(1+\delta)^{k-0}2^{-k} \\ &= 1 - (1+\delta)^k 2^{-k}, \end{aligned} \tag{4.27}$$

and

$$\Pr(\forall_{\text{pairs}}\ x \geq 1) = \left(1 - (1+\delta)^k 2^{-k}\right)^{\binom{N}{2}}.$$

We wish this probability to be bigger than $(1-\epsilon)$. In other words,

$$\left(1-(1+\delta)^k 2^{-k}\right)^{\binom{N}{2}} \overset{\text{want}}{>} 1-\epsilon \tag{4.28}$$

- LHS(4.28) $> e^{-\binom{N}{2}\left(1+((1+\delta)/2)^k\right)((1+\delta)/2)^k}$ by (4.15)

$$\Longleftarrow \quad \overset{\text{want}}{>} 1-\epsilon \tag{4.29}$$

$$\begin{aligned}
&\Longleftrightarrow \quad -\binom{N}{2}\left[1+\left(\frac{1+\delta}{2}\right)^k\right]\left(\frac{1+\delta}{2}\right)^k > \ln(1-\epsilon)\\
&\Longleftarrow \quad -\binom{N}{2}\left[1+\left(\frac{1+\delta}{2}\right)^k\right]\left(\frac{1+\delta}{2}\right)^k \overset{\text{want}}{>} -\epsilon\\
&\Longleftrightarrow \quad \epsilon > \binom{N}{2}\left[1+\left(\frac{1+\delta}{2}\right)^k\right]\left(\frac{1+\delta}{2}\right)^k\\
&\Longleftrightarrow \quad \frac{\left(\frac{2}{1+\delta}\right)^{2k}}{\left(\frac{2}{1+\delta}\right)^k+1} > (1/\epsilon)\binom{N}{2}, \qquad (4.30)
\end{aligned}$$

where the last transformation is obtained as in (4.19), replacing 2 by $2/(1+\delta)$. The same substitution in (4.21) (i.e., taking $\beta = (2/(1+\delta))^k$ in (4.20)) yields

$$\frac{\left(\frac{2}{1+\delta}\right)^{2k}}{\left(\frac{2}{1+\delta}\right)^k+1} > \left(\frac{2}{1+\delta}\right)^k - 1 \qquad \forall k \in \mathbb{N} \tag{4.31}$$

Thus, (4.30) follows if k is chosen to satisfy

$$\begin{aligned}
&\left(\frac{2}{1+\delta}\right)^k - 1 > (1/\epsilon)\binom{N}{2}\\
&\Longleftrightarrow \quad \boxed{\left(\frac{2}{1+\delta}\right)^k > (1/\epsilon)\binom{N}{2}+1} \qquad (4.32)
\end{aligned}$$

Again, a simpler bound on k can be obtained by imposing a stronger condition

$$\left(\frac{2}{1+\delta}\right)^k \overset{\text{want}}{>} (2/\epsilon)\binom{N}{2} = (1/\epsilon)N(N-1) \tag{4.33}$$

so that

$$\begin{aligned}
\lg\left[\left(\frac{2}{1+\delta}\right)^k\right] &= k(1-\lg(1+\delta)) \\
&\overset{\text{want}}{>} \quad \lg N + \lg(N-1) + \lg(1/\epsilon) \qquad (4.34) \\
&\Longleftarrow \quad k\,(1-\lg(1+\delta)) > 2\lg N + \lg(1/\epsilon).
\end{aligned}$$

$$\boxed{k > \frac{1}{(1-\lg(1+\delta))}\lg N + \frac{1}{(1-\lg(1+\delta))}\lg(1/\epsilon)} \qquad (4.35)$$

Non-unit Minimum Hamming Distance Finally, we can estimate the necessary size k for (almost) any desired minimum Hamming distance between allele code vectors. We saw above that the Hamming distance between a pair of HLA vectors is a binomial random variable $x \sim S(n,p)$ where # trials $\equiv n = k$, $\Pr(\text{success}) \equiv p = (1-\delta)/2$, and $\Pr(\text{failure}) \equiv q = (1+\delta)/2$:

$$\Pr(x) = \binom{k}{x}(1-\delta)^x(1+\delta)^{k-x}2^{-k}$$

Its mean is $np = k(1-\delta)/2$ and variance is $npq = k(1-\delta^2)/4$. One can then get the following estimate (using Chernoff bounds):

$$\Pr(x \le k(1-\delta)/4) \le \; e^{-k(1-\delta)/16} \qquad (4.36)$$

Chernoff inequality states (see Appendix C.2 for the proof):

$$\Pr(S(n,p) \le (1-\epsilon)np) \le \; e^{-\frac{\epsilon^2}{2}np} \qquad (4.37)$$

We have $n = k$, $p = (1-\delta)/2$, and let $\epsilon = \frac{1}{2}$. Then by (4.37),

$$\begin{aligned}
\Pr(x \le k(1-\delta)/4) &\le \; e^{-\frac{\left(\frac{1}{2}\right)^2}{2}k\frac{1-\delta}{2}} \\
= \quad e^{-\frac{1}{8}k\frac{1-\delta}{2}} &= \; e^{-k(1-\delta)/16}
\end{aligned}$$

Thus, we can estimate for which k

$$\Pr\left(\forall_{\text{pairs}}\ x \geq k(1-\delta)/4\right) \geq 1-\epsilon \tag{4.38}$$

From (4.36), we obtain

$$\Pr\left(x \geq k(1-\delta)/4\right) \geq 1-\ e^{-k(1-\delta)/16}$$

and hence,

$$\begin{aligned}
&\Pr\left(\forall_{\text{pairs}}\ x \geq k(1-\delta)/4\right) \\
&\quad = \quad \Pr\left(x \geq k(1-\delta)/4\right)^{\binom{N}{2}} \\
&\quad \geq \quad \left(1-\ e^{-k(1-\delta)/16}\right)^{\binom{N}{2}} \\
&\quad \overset{\text{want}}{>} \quad 1-\epsilon
\end{aligned} \tag{4.39}$$

$$\begin{aligned}
&\Longleftarrow \quad \text{(expression (4.39))} > \exp\left\{-\binom{N}{2}\left(1+\ e^{-k(1-\delta)/16}\right)\ e^{-k(1-\delta)/16}\right\} \\
&\qquad \overset{\text{want}}{>}\ 1-\epsilon \\
&\Longleftrightarrow \quad -\binom{N}{2}\left(1+\ e^{-k(1-\delta)/16}\right)\ e^{-k(1-\delta)/16} > \ln(1-\epsilon) \\
&\Longleftarrow \quad -\binom{N}{2}\left(1+\ e^{-k(1-\delta)/16}\right)\ e^{-k(1-\delta)/16} > -\epsilon \qquad \text{(see (4.17))} \\
&\Longleftrightarrow \quad \epsilon > \binom{N}{2}\left(1+\ e^{-k(1-\delta)/16}\right)\ e^{-k(1-\delta)/16} \\
&\Longleftrightarrow \quad \frac{e^{k(1-\delta)/8}}{e^{k(1-\delta)/16}+1} > (1/\epsilon)\binom{N}{2} \\
&\Longleftarrow \quad e^{k(1-\delta)/16} - 1 > (1/\epsilon)\binom{N}{2} \qquad \text{(by (4.20))} \\
&\Longleftrightarrow \quad e^{k(1-\delta)/16} > (1/\epsilon)\binom{N}{2} + 1
\end{aligned}$$

$$
\begin{aligned}
&\Longleftarrow \quad e^{k(1-\delta)/16} > (2/\epsilon)\binom{N}{2} = (1/\epsilon)N(N-1) \qquad \text{(see (4.23))} \\
&\Longleftrightarrow \quad k(1-\delta)/16 > \ln(1/\epsilon) + \ln N + \ln(N-1) \\
&\Longleftarrow \quad k(1-\delta)/16 > \ln(1/\epsilon) + 2\ln N \\
&\Longleftrightarrow \quad k > \frac{32}{1-\delta}\ln N + \frac{16}{1-\delta}\ln(1/\epsilon)
\end{aligned}
$$

Therefore, if

$$\boxed{k > \frac{32}{(1-\delta)}\ln N + \frac{16}{(1-\delta)}\ln(1/\epsilon)} \tag{4.40}$$

then

$$\boxed{\Pr\left(\forall_{\text{pairs}}\ x \geq k(1-\delta)/4\right) \geq 1-\epsilon}$$

For this probe set (with $k = k(\epsilon, N, \delta)$), we can obtain arbitrarily high probability (by our choice of ϵ in (4.38)) that all HLA coding vectors have pairwise Hamming distance of at least $k(1-\delta)/4$. Thus, this probe set will be able to correct $k(1-\delta)/8$ errors (by choosing the coding vector closest to that obtained).

It remains to estimate the error term δ. This can be accomplished on a given set S of probes by sampling pairs $\{l, m\}$ of indices on probes from S and examining the resulting 2-vectors on $\{0, 1\}$. We can then estimate the probability of failure to contribute to the Hamming distance (given by $(1+\delta)/2$) by the frequency $f_=$ of observing equal entries in the 2-vector (since each probe with equal entries in the 2-vector fails to contribute to the Hamming distance between alleles l and m). Therefore,

$$\hat{\delta} = 2f_= - 1 \tag{4.41}$$

Let $f_\neq$ denote the frequency of observing unequal entries in the same setting.

Clearly, $f_= + f_{\neq} = 1$, so we can obtain

$$1 - \delta = 1 - (2f_= - 1) = 2(1 - f_=) = 2f_{\neq}. \tag{4.42}$$

The bound in (4.40) becomes

$$\begin{aligned} k &> \frac{32}{2f_{\neq}} \ln N + \frac{16}{2f_{\neq}} \ln(1/\epsilon) \\ &= \frac{16}{f_{\neq}} \ln N + \frac{8}{f_{\neq}} \ln(1/\epsilon) \end{aligned} \tag{4.43}$$

Thus, to generate distinct coding vectors for all alleles (namely, to guarantee a Hamming distance $d_H(c_i, c_j) \geq 1$ w.p. $> 1 - \epsilon$), we should choose $M > k$, where k satisfies (4.35) with δ estimated as in (4.41). We can also select M to allow for error correction of up to $D/2$ errors (guaranteeing w.p. $> (1 - \epsilon)$ a minimum Hamming distance $d_H(c_i, c_j) \geq D$): set

$$D = k(1 - \delta)/4 = kf_{\neq}/2$$

in (4.38), so that $k = 2D/f_{\neq}$ must satisfy (4.40), and, again, choose $M > k$.

4.2.3 Pre-processing

Initial Probe Selection

In section 4.2.1, we stated that starting with all possible probes results in a graph that has too many vertices. This section discusses some pre-processing steps that allow us to eliminate a large portion of this probe set.

Probes that don't hit the HLA region on any allele Many of the possible length-L probes will not provide sequence-specific information about the target.

As such, they may be safely left out of our probe selection process. This will allow us to reduce the starting (perfectly matched) probe set to those probes that are complementary to a subsequence of at least one of the alleles. A simple way to obtain such a set is as follows. Let us assume that the allele sequences are given in the 5' to 3' orientation.

Consider a window of length L along allele T_1. Let us denote the length of the allelic sequence by $len(T_1)$, and index elements of the sequence starting with 1, so that the entire allele sequence can be denoted by

$$T_1[1] \ldots T_1[len(T_1)].$$

We can construct a probe complementary to the allele subsequence seen through the window $[1 \ldots L]$, place the probe in our set, and shift the window by one nucleotide in the direction of the 3'-end. This process can be repeated until the last window $[k \ldots (L + k - 1)]$ reaches the end of the target sequence:

$$\begin{aligned} L + k - 1 &= len(T_1) \\ k &= len(T_1) - L + 1, \end{aligned}$$

generating a set of $(len(T_1) - L + 1)$ probes, each perfectly complementary to T_1.

The process described generates all probes of length L that are perfectly complementary to a length-L subsequence of the target (i.e., allele) sequence. Depending on the form in which the allele sequences are given, it may also be desirable to include probes corresponding to windows that are partially shifted off the allele, i.e., windows showing a portion of the given allele sequence together with the corresponding 5'-tail of the sequence, if the window is shifted off to the left, or the

3'-tail, if the window is shifted off to the right. There are $2(L-1)$ such probes, corresponding to indices

$$[(len(T_1) - L + 2) \ldots (len(T_1) + 1)], \ldots, [len(T_1) \ldots (len(T_1) + L - 1)]$$

for the right-shifted windows and "indices"

$$[0 \ldots (L-1)], \ldots, [(2-L) \ldots 1]$$

for the left-shifted windows.

This process should be repeated for the other alleles $T_2, \ldots, T_N$. To avoid placing duplicate probes in our set, a generated probe should only be added to the set if its sequence is not already present. Alternately, the duplicates can be weeded out subsequently. It should be noted that this may have the added advantage of eliminating probes hitting sequence repeats.

While described in a constructive fashion, the above process has the effect of eliminating probes that hit genomic sequences outside the target region, including probes that hit introns (if allelic sequences are provided in genomic DNA form) from the original collection of all possible probes of length L. We end up with only those probes complementary to subsequences of the HLA region, or sub-words of the pool of all allele sequences.

Non-informative probes In the set created as described in the previous section, some (perhaps many) of the probes will not be able to give any information useful for distinguishing among the alleles. These are the probes drawn from windows that are shared among the alleles: they hybridize to a common subsequence of the alleles. Any such probe will be useless for discriminating alleles—to such a probe,

all alleles will look alike. Therefore, these probes can be safely eliminated from our potential probe set.

To find all such probes, we need only find the common subsequences of length $\geq L$ of all the alleles, identify the probes complementary to these subsequences, and remove these probes from the set. This can be done as the next step in the "refinement" of our starting probe set, or included as a condition in the process for probe addition specified in the previous section.

Potential for cross-hybridization Probes that are likely to hit multiple sites on the target sequence(s), such as those hitting a repeated region, should be eliminated, as is usually done in microarray design, as their use is likely to produce a high level of noise. Each probe is usually expected to have a unique site on the target. (See, for example, [30], [26], or [29].)

Graph Generation

Generating Probe Response Vectors Once a set of initial probes is selected, as described above, we must generate a probe response vector (4.7) for each of these probes. To do that, given probe j, we run string-matching on each of the N alleles for the Watson-Crick complement (defined in section 3.7.1) of probe j, and set

$$v_j[i] = \begin{cases} 1, & \text{if there is a match with allele HLA}_i \\ 0, & \text{otherwise} \end{cases}$$

Choice of Edge Threshold The edge threshold parameter ρ was used in section 4.2.1 to transform the initial complete edge-weighted graph. Its value determines how many edges remain in the graph, as well as how "independent" each

independent set on the graph really is.

- If ρ is too small, there will be very few edges in the graph. Most of the random sets selected by the boosting algorithm will prove to be independent. However, upon examination in post-processing, described in section 4.2.4, we may find that many of these sets do not possess enough discrimination power to discern all N known alleles.

- If, on the other hand, ρ is too large (e.g., $\rho > 0.5$), the graph will be very dense (i.e., have a lot of edges). The algorithm will then have a much harder time finding an independent set of large enough size. The output sets will likely contain much fewer than M vertices, and there may not be enough probes in the candidate sets to discern all N alleles.

- A reasonable value of ρ is obtained by trial and error on a given set of potential probes.

4.2.4 Post-processing: Ensuring Discrimination

The algorithm (section 4.2.2) returns a list of 20 best independent sets, sorted by the total information weight of the constituent vertices. Each set is made up of at most M vertices (probes). While the independence and maximum weight conditions were chosen to steer each selected set towards maximum discrimination power, this desired outcome is not guaranteed. Thus, each of these best independent sets has to be checked for redundancy of the allele coding vectors. Given a set S of probe response vectors, we can extract the N allele coding vectors generated by S (the rows in (4.6)) and compute their pairwise Hamming distances $d_H(c_i, c_j)$,

$1 \leq i < j \leq |S|$. If $\min_{\{i,j\}} d_H(c_i, c_j) == 0$, the set lacks discrimination power: at least two of the codes are the same, so the set will not be able to discern all known alleles. Such a set cannot be used "as is", and must either be discarded or supplemented by additional probes. This may indicate that the set is not truly independent, so our choice of edge threshold ρ, discussed in section 4.2.3, was inappropriate.

It is possible to make the testing more stringent, in order to allow for up to $D/2$ errors in the data, as discussed in section 4.2.2. Those independent sets of the list of best sets that pass the redundancy testing (by satisfying $\min_{\{i,j\}} d_H(c_i, c_j) \geq D$), in fact satisfy a definition stronger than that formulated for the best independent set. Let us denote by *D-best independent set* a best independent set with the additional condition $\min_{\{i,j\}} d_H(c_i, c_j) \geq D$.

Those sets that pass the redundancy test can be reordered by `aveHamDist` $= \text{ave}_{\{i,j\}} d_H(c_i, c_j)$: once the minimum allele code separation is guaranteed, the usefulness of a probe set to the HLA typing problem can be judged by the metric `aveHamDist`.

4.3 Interpreting Results

Given the best independent set generated by the above described algorithm, how do we turn it into a microarray for the HLA typing experiments?

In order to generate the microarray corresponding to an independent set of vectors yielded by the algorithm (followed by post-processing steps from section 4.2.4), we must retrace our steps and recall the DNA sequence for each probe—the se-

quence that was used to generate the probe response vector used in the analysis. A set of vertices in the graph corresponds to a set of probe sequences. The spatial arrangement of these probes on the chip surface can be decided as discussed in section 4.1.1.

4.4 Future Directions

4.4.1 Open problems

Many extensions of the material discussed in this chapter are possible. Two of the most interesting ones are discussed here.

Extending weight functions in the graph model

The graph model discussed here relies entirely on the characteristics of the probe response vectors to define the weights of vertices and edges. While even this simple model generates interesting results, it can be extended to a much more meaningful model by incorporating the physical properties of the probe sequences and their interactions, some of which were analyzed in chapter 3. In particular, annotation of all potential probes with physical properties, such as melting temperature, free energy, entropy, and enthalpy of hybridization, for perfect matches and for closest matches in other alleles can be used to define cost functions that determine the weights. While the vertex weight provides a measure of the performance of the corresponding probe in discriminating among known alleles, the pairwise probe interaction and the resulting competition effects, as described in chapter 3, can be reflected in the edge weights.

Pooling real data from previously tested chips

Another important extension of the method discussed in this chapter involves the use of data from microarray chips used for HLA typing by different companies. Many biotechnology companies are working on the HLA typing problem, in the hope of designing probe sets that give the answer quicker and with more accuracy. The sequences of the probes are generally considered to be proprietary information and thus not shared. As a result, the collection of experimental data from testing the various probes in different combinations and arrangements on the microarray chips generated by different companies is almost never examined as a whole.

We believe that it is possible to employ our probe interaction model to make use of the aggregate experimental data. Suppose the following information can be obtained: a set of microarray chips along with some identifiers, if not the actual sequences, of the probes comprising each chip, and values measuring the performance of each chip in all previously conducted HLA typing experiments. That is, for each chip, there is a list of unique probe identifiers and some measure of how well this chip performed in HLA typing. It is not necessary to know the sequence of each probe, as long as the uniqueness of the identifiers can be verified by the company providing the data. We propose to combine the information from a large number of such previously tested chips to generate a plan for a new microarray chip (i.e., a collection of probe identifiers and their spatial arrangement) with a performance value higher than that of all "input" chips by the following process. The probe content and arrangement for each chip, together with its performance value, can be used to build the graph model. Vertex weights can be inferred from chip membership information. Edge weights can be estimated from conditional

probabilities using pairwise membership information—that is, by considering two chips at a time, quantities such as the conditional probability that probe P_i was used on chip C_j, given that it was used on chip C_k, can be estimated. Once the graph is constructed, the boosting algorithm can be used to generate the best set of probes, as discussed in section 4.2.2.

Chapter 5

Conclusions

In this manuscript, careful attention has been paid to the myriad details of the design of microarray experiments and the analysis of the outcomes; all the work described here has found its way into the experimental work of life scientists. Nevertheless, it is but a single step in the right direction. Many more problems in genomics, as well as other areas of mathematical biology, need careful mathematical treatment and analysis.

The work described in this thesis can be furthered on many fronts; they are:

1. Shrinkage-based similarity metric: The shrinkage metric appears useful in Monte Carlo Markov Chain simulations, for improving the accuracy of the results. This was found in a currently ongoing project with Marc Sobel at Temple University. The same technique, described in chapter 2, will be applied to a new project: to correlate patients' metabolite levels. Here, data has the same relative dimensionality, i.e., $M \gg N$, but instead of measuring the response of many (M) genes under a few (N) experimental conditions, the

levels of a few (N) metabolites in the blood of many (M) patients would need to be correlated. The data comes from spectroscopic measurements; hence, careful analysis will have to be made to understand and compensate for the contribution of potential error sources.

2. Hybridization Models: The material presented in chapter 3 was tested on small examples. The models must be tested more extensively to compare model predictions with real experimental data. This comes in addition to the future work outlined at the end of chapter 3, e.g., generalizing thermodynamic parameter computations to treat mismatches more carefully. Simulations based on these models may also be used in probe design to define weight functions, as described in the chapter on HLA typing.

3. HLA typing/Probe design/Graph optimization problem: A lot more remains to be done on this problem. While the model definitions have been well thought through and described in detail in chapter 4, they have been tested only on a small example set so far. To demonstrate the usefulness of this approach to probe design, it must be tested on real sequence data with very large numbers of probes and long probe response vectors, corresponding to on the order of 10^3 alleles. Heuristics described in chapter 4 work well for the small test problem. However, better theory needs to be developed, to allow analytical results, which, in turn, may also lead to better heuristics in the future. The predictive, or allele-identifying, power of the algorithm presented should be tested on unknown alleles. Furthermore, the algorithm was designed to work on a probability space of alleles, and assumes that all

polymorphisms in the HLA region are combinations of polymorphisms drawn from some (unknown) distribution. Thus, we believe that this method has an advantage over existing methods that tend to overfit to sample data.

Appendix A

Appendices for Chapter 2

A.1 Receiver Operator Characteristic Curves (More Details)

A.1.1 Definitions

As a measure of truth, we take our working hypothesis, namely, the transcriptional activator table (Table 2.1). Thus, if two genes are in the same group, they "belong in the same cluster", and if they are in different groups, they "belong in different clusters". We will generate an ROC curve for each metric used (i.e., one for Eisen, one for Pearson, and one for Shrinkage).

Event: grouping of (cell cycle) genes into clusters;

Threshold: cut-off similarity value at which the hierarchy tree is cut into clusters.

Our cell-cycle gene table consists of 44 genes, which gives us $C(44, 2) = 946$ gene pairs. For each (unordered) gene pair $\{j, k\}$, we define the following events:

TP: $\{j,k\}$ are in the same group and $\{j,k\}$ are placed in the same cluster;

FP: $\{j,k\}$ are in different groups, but $\{j,k\}$ are placed in the same cluster;

TN: $\{j,k\}$ are in different groups and $\{j,k\}$ are placed in different clusters; and

FN: $\{j,k\}$ are in the same group, but $\{j,k\}$ are placed in different clusters.

Thus,

$$\begin{aligned}
\mathrm{TP}(\gamma) &= \sum_{\{j,k\}} \mathrm{TP}(\{j,k\}) \\
\mathrm{FP}(\gamma) &= \sum_{\{j,k\}} \mathrm{FP}(\{j,k\}) \\
\mathrm{TN}(\gamma) &= \sum_{\{j,k\}} \mathrm{TN}(\{j,k\}) \\
\mathrm{FN}(\gamma) &= \sum_{\{j,k\}} \mathrm{FN}(\{j,k\})
\end{aligned}$$

where the sums are taken over all 946 unordered pairs of genes.

Two other quantities involved in ROC curve generation are

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)}. \tag{A.1}$$

Specificity = fraction of negatives detected by a metric

$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)}. \tag{A.2}$$

An ROC curve plots sensitivity, on the y-axis, as a function of (1− specificity), on the x-axis, with each point on the plot corresponding to a different cut-off value. We create a different curve for each of the three metrics.

The following sections describe how the quantities TP(γ), FN(γ), FP(γ), and TN(γ) can be computed using our set notation for clusters. Recall from section 2.4.3:

$$\Big\{x \to \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\Big\}_{x=1}^{\#\text{ of groups}}$$

A.1.2 Computation

TP

$$\mathrm{TP}(\gamma) = \sum_{\{j,k\}} \mathrm{TP}(\{j,k\}) =$$

gene pairs that were placed in the same cluster and belong in the same group.

For each group x given in set notation as

$$x \to \{\{y_1, z_1\}, \ldots, \{y_{n_x}, z_{n_x}\}\},$$

we count pairs from each y_j, i.e.,

$$\mathrm{TP}(x) = \binom{y_1}{2} + \cdots + \binom{y_{n_x}}{2} = \sum_{j=1}^{n_x} \binom{y_j}{2}$$

Totaling over all groups yields

$$\mathrm{TP}(\gamma) = \sum_{x=1}^{\#\text{ groups}} \mathrm{TP}(x) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2}$$

FN

$$\mathrm{FN}(\gamma) = \sum_{\{j,k\}} \mathrm{FN}(\{j,k\}) =$$

gene pairs that belong in the same group but were placed into different clusters.

We must count every pair that got separated.

$$\mathrm{FN}(x) = \begin{cases} \sum_{j=1}^{n_x} \sum_{k=j+1}^{n_x} y_j \cdot y_k & \text{if } n_x \geq 2, \text{ or} \\ 0, & \text{if } n_x = 1. \end{cases}$$

However, when $n_x = 1$, there is no pair $\{j, k\}$ that satisfies the triple inequality $1 \leq j < k \leq n_x$, and hence, we do not have to treat it as a special case.

$$\therefore \mathrm{FN}(\gamma) = \sum_{x=1}^{\#\text{ groups}} \mathrm{FN}(x) = \sum_{x} \sum_{1 \leq j < k \leq n_x} y_j \cdot y_k$$

FP

$$\mathrm{FP}(\gamma) = \sum_{\{j,k\}} \mathrm{FP}(\{j, k\}) =$$

gene pairs that belong in different groups

but got placed in the same cluster.

The expression

$$\sum_{x} \sum_{j=1}^{n_x} y_j \cdot z_j$$

counts every false-positive pair $\{j, k\}$ twice: first, when looking at j's group, and again, when looking at k's group.

$$\therefore \mathrm{FP}(\gamma) \quad = \quad \frac{1}{2} \sum_{x} \sum_{j=1}^{n_x} y_j \cdot z_j$$

TN

$$\mathrm{TN}(\gamma) = \sum_{\{j,k\}} \mathrm{TN}(\{j, k\}) =$$

gene pairs that belong in different groups

and got placed in different clusters.

Instead of counting true-negatives from our notation, we use the fact that we know the other three scores and the total they all add up to.

Complementarity Given a gene pair $\{j,k\}$, exactly one of the events $\{\text{TP}(\{j,k\}), \text{FN}(\{j,k\}), \text{FP}(\{j,k\}), \text{TN}(\{j,k\})\}$ is true, i.e., exactly one of them $= 1$, while the rest $= 0$. This implies

$$\begin{aligned}
&\sum_{\{j,k\}} \text{TP}(\{j,k\}) + \sum_{\{j,k\}} \text{FN}(\{j,k\}) + \\
&\quad + \sum_{\{j,k\}} \text{FP}(\{j,k\}) + \sum_{\{j,k\}} \text{TN}(\{j,k\}) = \\
&\quad = \text{TP}(\gamma) + \text{FN}(\gamma) + \text{FP}(\gamma) + \text{TN}(\gamma) = \\
&\quad = \binom{44}{2} = \frac{44 \cdot 43}{2} = 946 = \text{Total}
\end{aligned}$$

$$\therefore \text{TN}(\gamma) = \text{Total} - (\text{TP}(\gamma) + \text{FN}(\gamma) + \text{FP}(\gamma))$$

A.1.3 Plotting ROC curves

For each cut-off value θ, we can compute $\text{TP}(\gamma)$, $\text{FN}(\gamma)$, $\text{FP}(\gamma)$, and $\text{TN}(\gamma)$ as described in the previous section, with $\gamma \in \{0.0, 0.66, 1.0\}$ corresponding to Eisen, Shrinkage, and Pearson, respectively. Then, the sensitivity and specificity are computed from equations (A.1) and (A.2), and we can plot sensitivity vs (1− specificity), as shown in Figure 2.3.

We can also examine the effect of the cut-off threshold θ on the FN and FP scores individually, as shown in Figure 2.4.

A 3-dimensional plot of (1− specificity) on the $x-$axis, sensitivity on the $y-$axis, and threshold on the $z-$axis offers an interesting view, as shown in Figure A.1.

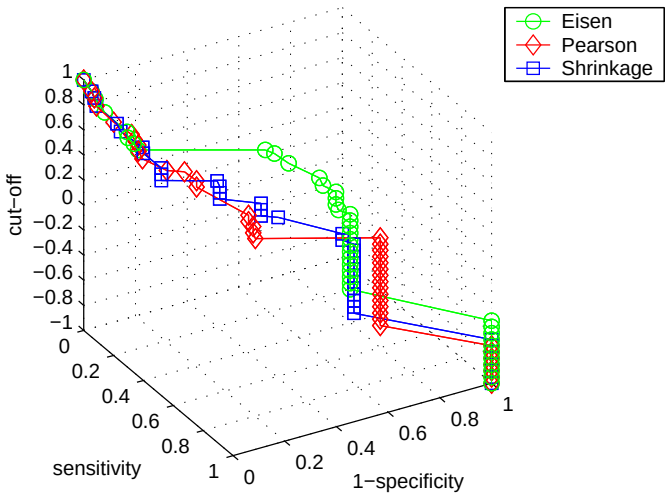


Figure A.1: ROC curves, with threshold plotted on the $z-$axis.

A.2 Computing the Marginal PDF for X_j

$$\begin{aligned} f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(X_j-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau}\, e^{-\frac{\theta^2}{2\tau^2}} d\theta \\ &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{(X_j-\theta)^2}{\sigma^2} + \frac{\theta^2}{\tau^2}\right)} d\theta \end{aligned} \tag{A.3}$$

First, rewrite the exponent as a complete square:

$$\begin{aligned}
\frac{(X_j-\theta)^2}{\sigma^2}+\frac{\theta^2}{\tau^2} &= \frac{1}{\sigma^2\tau^2}\left[\tau^2(X_j-\theta)^2+\sigma^2\theta^2\right] \\
&= \frac{1}{\sigma^2\tau^2}\left[\tau^2 {X_j}^2-2\tau^2X_j\theta+\tau^2\theta^2+\sigma^2\theta^2\right] \\
&= \frac{1}{\sigma^2\tau^2}\left[(\sigma^2+\tau^2)\theta^2-2\tau^2X_j\theta+\tau^2{X_j}^2\right] \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[\theta^2-2\frac{\tau^2}{\sigma^2+\tau^2}X_j\theta+\frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2\right] \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\Bigg[\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 \\
&\quad \underbrace{-\left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2}\Bigg] \qquad \text{(A.4)}
\end{aligned}$$

- $$\begin{aligned}
\frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2-\left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 &= {X_j}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(1-\frac{\tau^2}{\sigma^2+\tau^2}\right) \\
&= {X_j}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right) \\
&= {X_j}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2} \qquad \text{(A.5)}
\end{aligned}$$

Substituting (A.5) into (A.4) yields

$$\begin{aligned}
\frac{(X_j-\theta)^2}{\sigma^2}+\frac{\theta^2}{\tau^2} &= \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}{X_j}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{{X_j}^2}{\sigma^2+\tau^2} \qquad \text{(A.6)}
\end{aligned}$$

Now use the completed square in (A.6) to continue the computation in (A.3).

$$\begin{aligned} & f(X_j) \\ &= \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2} e^{-\frac{1}{2}\frac{{X_j}^2}{\sigma^2+\tau^2}}\,d\theta \\ &= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau}\int_{-\infty}^{\infty}\exp\left[-\left(\frac{\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j}{\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}}\right)^2\right]d\theta \end{aligned}$$

Make the substitution

$$\varphi = \left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)\Big/\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}$$

Then

$$\begin{aligned} d\varphi &= d\theta\Big/\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}} \\ d\theta &= \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\,d\varphi \\ \theta=\pm\infty &\implies \varphi=\pm\infty \end{aligned}$$

and

$$\begin{aligned} f(X_j) &= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\varphi^2}\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\,d\varphi \\ &= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{\pi\sqrt{2(\sigma^2+\tau^2)}}\underbrace{\int_{-\infty}^{\infty} e^{-\varphi^2}d\varphi}_{\sqrt{\pi}} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\,e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}} \end{aligned}$$

Therefore

$$f(X_j) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\ e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}} \tag{A.7}$$

A.3 Calculation of the Posterior Distribution of θ_j

Since the subscript j remains constant throughout the calculation, it will be dropped in this appendix. Herein, θ_j will be replaced by θ, and X_j by X.

$$\begin{aligned}
\pi(\theta|X) &= \frac{f(X|\theta)\,\pi(\theta)}{f(X)} = \frac{f(X,\theta)}{f(X)} \\
&= \frac{(1/2\pi\sigma\tau)\ \exp\left[-\left(\frac{\theta^2}{2\tau^2}+\frac{(X-\theta)^2}{2\sigma^2}\right)\right]}{\left(1\Big/\sqrt{2\pi(\sigma^2+\tau^2)}\right)\ \exp\left[-\frac{X^2}{2(\sigma^2+\tau^2)}\right]} \\
&= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\ \exp\left[-\frac{1}{2}\underbrace{\left(\frac{\theta^2}{\tau^2}+\frac{(X-\theta)^2}{\sigma^2}-\frac{X^2}{\sigma^2+\tau^2}\right)}\right]
\end{aligned}$$

- $$\frac{\theta^2}{\tau^2} + \frac{(X-\theta)^2}{\sigma^2} - \frac{X^2}{\sigma^2+\tau^2} =$$
$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Bigg[\sigma^2(\sigma^2+\tau^2)\theta^2 + \tau^2(\sigma^2+\tau^2)\overbrace{(X-\theta)^2}^{X^2-2X\theta+\theta^2} - \sigma^2\tau^2X^2\Bigg]$$
$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2\left(\sigma^2(\sigma^2+\tau^2)+\tau^2(\sigma^2+\tau^2)\right) - 2\tau^2(\sigma^2+\tau^2)X\theta + X^2\left(\tau^2(\sigma^2+\tau^2)-\sigma^2\tau^2\right)\Big]$$
$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2(\sigma^2+\tau^2)^2 - 2(\sigma^2+\tau^2)\theta\cdot\tau^2X + \tau^4X^2\Big]$$
$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\left((\sigma^2+\tau^2)\theta - \tau^2X\right)^2$$
$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}(\sigma^2+\tau^2)^2\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2$$
$$= \left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2 \Big/ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$$

Therefore,

$$\pi\left(\theta|X\right) = \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\ \exp\left[-\frac{\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right] \tag{A.8}$$

A.4 Proof of the fact that n independent observations from the Normal population $\mathcal{N}(\theta, \sigma^2)$ can be treated as a single observation from $\mathcal{N}(\theta, \sigma^2/n)$

Given the data y, $f(y|\theta)$ can be viewed as a function of θ. We then call it the *likelihood function* of θ for given y, and write

$$l(\theta|y) \propto f(y|\theta).$$

When y is a single data point from $\mathcal{N}(\theta, \sigma^2)$,

$$l(\theta|y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - x}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2\sigma^2}(\theta - x)^2\right], \tag{A.9}$$

where x is some function of y.

Now, suppose that $\vec{y} = (y_1, \ldots, y_n)$ represents a vector of n independent observations from $\mathcal{N}(\theta, \sigma^2)$. We can denote the sample mean by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The likelihood function of θ given such n independent observations from $\mathcal{N}(\theta, \sigma^2)$ is

$$l(\theta|\vec{y}) \propto \prod_i \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] = \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right].$$

Also, since

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2, \tag{A.10}$$

it follows that

$$\begin{aligned} l(\theta|\vec{y}) &\propto \underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i-\bar{y})^2\right]}_{\text{const w.r.t. } \theta} \exp\left[-\frac{1}{2\sigma^2} n(\bar{y}-\theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta-\bar{y})^2\right], \end{aligned} \tag{A.11}$$

which is a Normal function with mean $\bar{y}$ and variance σ^2/n. Comparing with (A.9), we can recognize that this is equivalent to treating the data $\vec{y}$ as a single observation $\bar{y}$ with mean θ and variance σ^2/n, i.e.,

$$\bar{y} \sim \mathcal{N}(\theta, \sigma^2/n). \tag{A.12}$$

Proof of (A.10)**:**

$$\begin{aligned} \sum_{i=1}^{n}(y_i-\theta)^2 &= \sum_i (y_i-\bar{y}+\bar{y}-\theta)^2 \\ &= \sum_i \left[(y_i-\bar{y})^2 + 2(y_i-\bar{y})(\bar{y}-\theta) + (\bar{y}-\theta)^2\right] \\ &= \sum_i (y_i-\bar{y})^2 + 2(\bar{y}-\theta)\sum_i (y_i-\bar{y}) + \sum_i (\bar{y}-\theta)^2 \\ &= \sum_i (y_i-\bar{y})^2 + 2(\bar{y}-\theta)\underbrace{\left(\sum_i y_i - \sum_i \bar{y}\right)}_{n\bar{y}-n\bar{y}=0} + n(\bar{y}-\theta)^2 \\ &= \sum_i (y_i-\bar{y})^2 + n(\bar{y}-\theta)^2 \end{aligned}$$

A.5 Distribution of the Sum of Two Independent Normal Random Variables

Let

$$\begin{aligned} X &\sim \mathcal{N}(0, \alpha^2) \\ Y &\sim \mathcal{N}(0, \beta^2) \end{aligned}$$

be two independent random variables.

Claim: $X + Y \sim \mathcal{N}(0, \alpha^2 + \beta^2)$

(We are only using this result for mean-0 Normal r.v.'s, although a more general result can be proven.)

Proof: (use moment generating functions)

$$\begin{aligned} m_X(t) &= \mathbf{E}\left(e^{tX} \right) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}\alpha}\, e^{-\frac{1}{2\alpha^2}(x-0)^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}\alpha} \int_{-\infty}^{\infty} e^{-\frac{1}{2\alpha^2}\underbrace{[x^2 - 2\alpha^2 tx]}}\, dx \end{aligned} \tag{A.13}$$

Completing the square, we obtain

$$\begin{aligned} x^2 - 2\alpha^2 tx &= x^2 - 2(\alpha^2 t)x + (\alpha^2 t)^2 - (\alpha^2 t)^2 \\ &= (x - \alpha^2 t)^2 - (\alpha^4 t^2) \\ \frac{1}{\alpha^2}(x^2 - 2\alpha^2 tx) &= \big((x - \alpha^2 t)/\alpha\big)^2 - (\alpha^4 t^2)/\alpha^2 \\ &= \left(\frac{x - \alpha^2 t}{\alpha}\right)^2 - \alpha^2 t^2 \end{aligned} \tag{A.14}$$

Using the result of (A.14) in (A.13) yields

$$m_X(t) = \frac{e^{-\frac{1}{2}(-\alpha^2 t^2)}}{\sqrt{2\pi}\alpha} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\alpha^2 t}{\alpha}\right)^2} dx$$

$$\begin{aligned} \text{Let } y &= \frac{x-\alpha^2 t}{\alpha} \\ dy &= \frac{dx}{\alpha} \implies dx = \alpha\, dy \end{aligned}$$

With this substitution, we obtain

$$m_X(t) = \frac{e^{\frac{1}{2}\alpha^2 t^2}}{\sqrt{2\pi}\alpha} \cdot \alpha \underbrace{\int_{y=-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy}_{\sqrt{2\pi}}$$

or

$$m_X(t) = e^{\frac{1}{2}\alpha^2 t^2} \tag{A.15}$$

Similarly

$$m_Y(t) = e^{\frac{1}{2}\beta^2 t^2} \tag{A.16}$$

To obtain the distribution of $X+Y$, it suffices to compute the corresponding moment generating function:

$$\begin{aligned} m_{X+Y}(t) &= \mathbf{E}\left(e^{t(X+Y)}\right) = \mathbf{E}\left(e^{tX}\, e^{tY}\right) \\ &= \mathbf{E}\left(e^{tX}\right)\mathbf{E}\left(e^{tY}\right) \quad \text{by independence of } X \text{ and } Y \\ &= m_X(t)\cdot m_Y(t) \\ &= e^{\frac{1}{2}\alpha^2 t^2} \cdot e^{\frac{1}{2}\beta^2 t^2} \quad \text{by (A.15) and (A.16)} \\ &= e^{\frac{1}{2}(\alpha^2+\beta^2)t^2}, \end{aligned}$$

which is a moment generating function of a Normal random variable with mean 0 and variance $\alpha^2 + \beta^2$. Therefore,

$$X + Y \sim \mathcal{N}(0, \alpha^2 + \beta^2). \tag{A.17}$$

A.6 Properties of the Chi-square Distribution

Let $X_1, X_2, \ldots, X_k$ be i.i.d.r.v.'s from standard Normal distribution, i.e.,

$$X_j \sim \mathcal{N}(0, 1) \quad \forall j.$$

Then

$$\chi_k^2 = X_1^2 + X_2^2 + \cdots + X_k^2 = \sum_{j=1}^{k} {X_j}^2$$

is a random variable from Chi-square distribution with k degrees of freedom, denoted

$$\chi_k^2 \sim \chi_{(k)}^2.$$

It has the probability density function

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma\left(k/2\right)} \, x^{k/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} \, dt. \tag{A.18}$$

The result we are using is

$$\mathbf{E}\left(\frac{1}{\chi_k^2}\right) = \frac{1}{k-2} \quad \text{for } k > 2,$$

which can be obtained as follows:

$$\begin{aligned}
\mathbf{E}\left(\frac{1}{\chi_k^2}\right) &= \int_{\mathcal{R}} \frac{1}{x}\, f(x)\, dx \\
&= \frac{1}{2^{k/2}\Gamma\left(k/2\right)} \int_0^\infty \frac{1}{x}\, x^{k/2-1} e^{-x/2}\, dx \\
&= \frac{1}{2^{k/2}\Gamma\left(k/2\right)} \int_0^\infty x^{k/2-2} e^{-x/2}\, dx \qquad \text{(A.19)}
\end{aligned}$$

Let

$$\begin{aligned}
t &= x/2 \implies x = 2t \\
& \qquad\qquad dx = 2dt \\
x &= 0 \implies t = 0 \\
x &= \infty \implies t = \infty
\end{aligned}$$

$$\begin{aligned}
&\int_0^\infty x^{k/2-2} e^{-x/2}\, dx \\
&= \int_{t=0}^\infty (2t)^{k/2-2}\, e^{-t}\, 2\, dt \\
&= 2^{k/2-2} \cdot 2 \int_0^\infty t^{k/2-2} e^{-t}\, dt. \qquad \text{(A.20)}
\end{aligned}$$

Let

$$\begin{aligned}
u &= e^{-t} \qquad dv = t^{k/2-2}\, dt \\
du &= -e^{-t}\, dt \quad v = \tfrac{t^{k/2-1}}{k/2-1} \qquad \text{for } k > 2
\end{aligned}$$

Integration by parts transforms (A.20) into

$$\begin{aligned}
&= 2^{k/2-1} \left(\frac{1}{k/2-1} \underbrace{e^{-t} t^{k/2-1}\big|_0^\infty}_{\longrightarrow 0} - \int_0^\infty \frac{1}{k/2-1}\, t^{k/2-1} \left(-e^{-t}\right)\, dt \right) \\
&= \frac{2^{k/2-1}}{k/2-1} \underbrace{\int_0^\infty t^{k/2-1} e^{-t}\, dt}_{\Gamma(k/2),\ \text{by (A.18)}} \\
&= \frac{2^{k/2-1}}{k/2-1}\, \Gamma(k/2)
\end{aligned}$$

Substituting this result in (A.19) yields

$$\begin{aligned}
\mathbf{E}\left(\frac{1}{\chi_k^2}\right) &= \frac{1}{2^{k/2}\Gamma(k/2)} \cdot \frac{2^{k/2-1}\Gamma(k/2)}{k/2-1} \\
&= \frac{1}{2\,(k/2-1)} \\
&= \frac{1}{k-2} \quad \text{for } k > 2. \qquad \text{(A.21)}
\end{aligned}$$

A.7 Distribution of Sample Variance s^2

Let $X_j \sim \mathcal{N}(\mu, \sigma^2)$ for $j = 1, \ldots, n$ be independent r.v.'s. We'll derive the joint distribution of

$$\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} \quad \text{and} \quad \frac{(n-1)\,s^2}{\sigma^2}.$$

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2 \\
\frac{(n-1)\,s^2}{\sigma^2} &= \frac{n-1}{\sigma^2} \cdot \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2 \\
&= \sum_{j=1}^{n}\left(\frac{X_j - \bar{X}}{\sigma}\right)^2
\end{aligned}$$

W.L.O.G. can reduce the problem to the case $\mathcal{N}(0,1)$, i.e., $\mu = 0$, $\sigma^2 = 1$: Let $Z_j = (X_j - \mu)/\sigma$. Then

$$\begin{aligned}
\bar{Z} &= \frac{1}{n}\sum Z_j = \frac{1}{n}\sum\left(\frac{X_j - \mu}{\sigma}\right) = \frac{1}{n}\left(\frac{\sum X_j}{\sigma} - \frac{\sum \mu}{\sigma}\right) \\
&= \frac{1}{n}\left(\frac{\sum X_j}{\sigma} - \frac{n\mu}{\sigma}\right) = \frac{1}{\sigma}\left(\frac{\sum X_j}{n} - \mu\right) = \frac{\bar{X} - \mu}{\sigma}
\end{aligned}$$

and hence

$$\frac{\sqrt{n}\,\left(\bar{X} - \mu\right)}{\sigma} = \sqrt{n}\,\bar{Z}. \qquad \text{(A.22)}$$

Also,

$$\begin{aligned}\frac{(n-1)\,s^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum\left(X_j-\bar{X}\right)^2 \\ &= \frac{1}{\sigma^2}\sum\left((X_j-\mu)+\left(\mu-\bar{X}\right)\right)^2 \\ &= \sum\left[\frac{X_j-\mu}{\sigma}-\frac{\bar{X}-\mu}{\sigma}\right]^2=\sum\left(Z_j-\bar{Z}\right)^2 \qquad \text{(A.23)}\end{aligned}$$

By (A.22) and (A.23), it suffices to derive the joint distribution of $\sqrt{n}\,\bar{Z}$ and $\sum_{j=1}^{n}\left(Z_j-\bar{Z}\right)^2$, where $Z_1,\ldots,Z_n$ are i.i.d. from $\mathcal{N}(0,1)$.

Let

$$P=\begin{pmatrix} \text{——}p_1\text{——} \\ \text{——}p_2\text{——} \\ \vdots \\ \text{——}p_n\text{——} \end{pmatrix}$$

be an $n\times n$ orthogonal matrix where

$$p_1=\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$$

and the remaining rows p_j are obtained by, say, applying Gramm-Schmidt to $\{p_1,e_2,e_3,\ldots,e_n\}$, where e_j is a standard unit vector in j^{th} direction in $\mathcal{R}^n$. Let

$$\begin{aligned}\vec{Y}&=P\vec{Z} \\ &=\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \hline & & & \\ & \vdots & & \\ & & & \\ \hline \end{pmatrix}\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}\end{aligned}$$

Then

$$Y_1=\frac{1}{\sqrt{n}}\left(\sum_{j=1}^{n}Z_j\right)=\frac{1}{\sqrt{n}}n\bar{Z}=\sqrt{n}\,\bar{Z}. \qquad \text{(A.24)}$$

Since P is orthogonal, it preserves vector lengths:

$$\begin{aligned} \|\vec{Y}\|^2 &= \|\vec{Z}\|^2 \\ \sum_{j=1}^{n} {Y_j}^2 &= \sum_{j=1}^{n} {Z_j}^2 \\ \implies \left(\sum_{j=1}^{n} {Y_j}^2\right) - Y_1^2 &= \sum_{j=1}^{n} {Z_j}^2 - \left(\sqrt{n}\,\bar{Z}\right)^2 \quad \text{by (A.24)} \end{aligned}$$

Hence

$$\begin{aligned} \sum_{j=2}^{n} {Y_j}^2 &= \sum_{j=1}^{n} {Z_j}^2 - n\bar{Z}^2 = \sum_{j=1}^{n} {Z_j}^2 - 2n\bar{Z}^2 + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} {Z_j}^2 - 2\bar{Z}(n\bar{Z}) + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} {Z_j}^2 - 2\bar{Z}\left(\sum_{j=1}^{n} Z_j\right) + \sum_{j=1}^{n} \bar{Z}^2 \\ &= \sum_{j=1}^{n} \left(Z_j - \bar{Z}\right)^2 \end{aligned} \tag{A.25}$$

Since the Y_j's are mutually independent (by orthogonality of P), we can conclude that

$$\sum_{j=2}^{n} {Y_j}^2 = \sum_{j=1}^{n} \left(Z_j - \bar{Z}\right)^2$$

is independent of

$$Y_1 = \sqrt{n}\,\bar{Z}.$$

Also by orthogonality of P, $Y_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, n$, so

$$\left(\sum_{j=2}^{n} {Y_j}^2\right) \sim \chi^2_{(n-1)} \quad \text{(See Appendix A.6)}$$

and hence, by (A.23) and (A.25),

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{(n-1)} \tag{A.26}$$

Since $\mathbf{E}\left(\chi_k^2\right) = k$, for $\chi_k^2 \sim \chi_{(k)}^2$, we can see that

$$\mathbf{E}\left(\frac{(n-1)\,s^2}{\sigma^2}\right) = n - 1.$$

Also, since

$$\mathbf{E}\left(\frac{(n-1)\,s^2}{\sigma^2}\right) = \frac{n-1}{\sigma^2}\,\mathbf{E}\left(s^2\right),$$

we can conclude that

$$\mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1} \cdot \frac{n-1}{\sigma^2}\,\mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2, \tag{A.27}$$

i.e., s^2 is an unbiased estimator of the variance σ^2.

Appendix B

Appendices for Chapter 3

B.1 Details of Model Implementation

B.1.1 Choice of initial concentration parameters

The pm/mm ratios at equilibrium are computed from systems of equations corresponding to a particular model (equations (3.12)–(3.19) and (3.22)–(3.26) for Full Model; equations (3.32)–(3.35) and (3.36)–(3.38) for Model I; and equations (3.45)–(3.48) and (3.49)–(3.51) for Model II). In all these systems, the second set, made up of conservation equations, involves the initial concentration constants. The solutions will depend on the initial probe and target concentrations. There are several complications that stem from the choice of these parameters.

Initial probe concentrations

To measure competition effects inherent to the probes, the initial probe concentrations must be equal for different probes. If that is not the case, the observed results

will be biased by the unequal starting amounts of probe material.

Initial target concentration

The choice of the initial target concentration must be made carefully as well. If $[T]_0 > \sum[P_{ij}]_0$, no competition effect will be observed since there is plenty of target to go around—each probe will get as much target as it needs.

The idea of "competition" described in this manuscript (see section 3.2 for the initial discussion) is based on the fact that a given target molecule, which can hybridize to either probe 1 or probe 2, or to a variant (alternate) of either of the probes, is more likely at equilibrium to end up hybridized to the probe that "holds it the strongest," i.e., the one with the most negative ΔG of hybridization. On the mass-action scale, this means that a higher proportion of the total target concentration will end up in a complex with the most "attractive" probe, or the best matching probe. However, this argument implicitly assumes that the targets are in short supply and the probes are competing for them.

Initial conditions for actual experiments performed in the lab frequently use a different setup: the initial amount of the target is in huge excess over the probes. This appears to imply that under such conditions no competition effects should be observed—target-probe complexes should be formed for each of the probes, as there is plenty of target to go around! The affinity of the target to a particular probe does not enter the equation. And yet, experiments reveal the presence of competition. For a discussion of how this apparent paradox is resolved, see section B.1.2.

Scaling initial concentrations for comparison

To allow meaningful comparison among match-to-mismatch ratios for a given probe under different models, the initial concentration parameters must be scaled. If that is not done, much more sophisticated post-processing will be required to interpret the differences in the pm/mm ratio values. Scaling the parameters *a priori* also allows for the ratio curves to be plotted on the same set of axes and for the changes to be interpreted as "shifts" of the ratio curves.

B.1.2 Accuracy of entered parameters

The amount of target initially placed in the reaction chamber, together with chamber volume, is usually used to compute the initial target concentration. However, the value of $[T]_0$ computed in this manner may not be accurate. There are steric hindrances in the system. Probes are physically attached to large (relative to probe size) beads, and placed in the reaction chamber. The target molecules, which are much longer than probes, are free to float around the chamber. In order to interact with the probes, the target molecules must diffuse through the chamber. Only a small fraction of the target molecules placed in the reaction chamber end up close enough to the probe molecules to interact (i.e., hybridize) with them. Thus, while the amount of target the experimenter places in the reaction chamber may significantly exceed the total amount of probes, the constraint

$$([T]_0)_{\text{effective}} < \sum [P_{ij}]_0$$

frequently holds for the *effective* initial target concentration, that is, the concentration of target molecules that diffused sufficiently far to reach the probes and

participate in the hybridization reaction. This explains why competition effects, which are observed in the model only when $[T]_0 < \sum[P_{ij}]_0$, are also observed in practice.

Theoretically, one can compute the effective initial target concentration from the initial probe concentration, temperature, and measured pm/mm ratio. If the ratios are obtained for each probe separately, no competition effects will be present; hence the simple model of hybridization, described in section 3.6.1 can be used. Furthermore, if this data is obtained for a sequence of physical $[T]_0$'s, it may also be possible to observe a functional relationship between the physical $[T]_0$ and the effective $[T]_0$.

Suppose that for a given probe, a hybridization experiment is performed involving that probe, its alternate, and the target, and the concentrations of the matched probe-target complex (denoted by X_2) and the mismatched probe-target complex (denoted by X_3) at equilibrium are measured; the pm/mm ratio at equilibrium can be computed from the values of X_2 and X_3. The outcome of the experiment can be predicted *in silico* by the simple model. Recall that during the discussion of the dynamics of the simple model in section 3.6.1, equation (3.104) for the pm/mm ratio was obtained; this equation is repeated here for convenience:

$$\text{ratio1} = \frac{a_0}{b_0} \cdot \frac{K_1^2}{K_1^3} \cdot \frac{K_1^3 + \frac{1}{X_1}}{K_1^2 + \frac{1}{X_1}}$$

Equation (3.104) can be used to solve for the equilibrium concentration of the free target (denoted by X_1) in terms of ratio1, which is, in turn, given in terms of the

measured quantities X_2 and X_3:

$$\text{ratio1} = \frac{X_2}{X_3} \tag{B.1}$$

$$X_1 = \frac{\text{ratio1} - (a_0/b_0) \cdot (K_1^2/K_1^3)}{K_1^2\left((a_0/b_0) - \text{ratio1}\right)} \tag{B.2}$$

Finally, the effective initial target concentration can be obtained from the conservation rule

$$e_0 = ([T]_0)_{\text{effective}} = X_1 + X_2 + X_3, \tag{B.3}$$

where X_1 is given in (B.2). It is worth noting that this computation requires the values of X_2 and X_3, and not just their ratio (i.e., ratio1). This brings up the additional complication of converting the measured quantities (intensities) into the same units as the computed quantities (concentrations), which is discussed in detail in section B.1.3.

B.1.3 Interpreting the results

In the laboratory, to obtain the concentration of a particular complex, one measures instead the total intensity of the fluorophores attached to the molecules of the complex. This intensity is a function of the concentration of the substance in question. The form of this function is generally assumed to be linear in a certain range, growing nonlinear outside the said range. Since the lab measurements are in the units of intensity, and the model predicts concentrations of the substances at equilibrium, direct comparison of the *in silico* and laboratory data does not make sense. However, one can make the argument that since the primary interest is not in the concentrations of individual substances but rather in their ratios (the pm/mm

ratios), the intensity-to-concentration scaling factor cancels out. The investigation described in this manuscript has relied on this assumption.

Nevertheless, one should be careful to verify that the quantities in question do indeed fall into the "linear" range of the intensity function. Should that prove not to be the case, it would no longer be appropriate to treat intensities and concentrations interchangeably; a more careful analysis of this "unit translation" would be prudent.

Furthermore, some of the analysis discussed here, in particular in section B.1.2, requires individual concentration values, making it necessary to formulate the relationship between intensity and concentration explicitly. Information required to obtain the function in question includes the intensity of a single fluor, the number of fluors attached to each target molecule, and the details of how the experimentally measured results are scaled (i.e., the post-processing of the scanned data).

B.2 Future Improvements

B.2.1 Choice of alternate sites

All models discussed in chapter 3, with the exception of Simple Model, allow alternate binding sites for each probe. In the current formulation, those alternate sites are hard-wired to be the matching sites for the other probes involved. This choice of alternate sites fits in with the idea of how competition between probes works, and was convenient to implement, since the portions of the target in question were already stored as the complementary sequences for the other probes. As an added convenience, it also allowed the implementation to avoid string matching, since all the necessary string matching was done as a pre-processing step.

However, it would be more realistic to choose the alternate binding site(s) for each probe based on the sequence of that probe as well as that of the target. One possible approach to selecting potential "alternate sites" for a given probe could be the following. One could generate a landscape of affinity constants ($K12$) and/or melting temperatures (T_m) by convolving the given probe with the long target, i.e., by shifting the probe along the target and computing the quantity of interest at each such alignment, and then threshold it, only keeping the "peaks" as the alternate sites.

B.2.2 Thermodynamics of mismatches

The current implementation of all hybridization models discussed in chapter 3 computes the thermodynamic parameters of hybridization based on the NN model, making use of the parameters for all possible *matching* dimer duplexes, as described in section 3.7.1. Recall that in all these hybridization models, for each probe there is an alternate probe, almost identical to the matching one (in all examples considered, the alternate (mismatching) probe differs from the matching one in only one base). Thus, it is necessary to make regular computations of thermodynamic parameters for target-probe pairs where mismatches occur. Further, more severe mismatches occur when "cross-terms" are considered, where a probe hybridizes to the "wrong" location on the target.

Current implementation

The simplest way to deal with such mismatches, and the one used in the current implementation, is to "lose" the contributions of all mismatched dimers to the

summation term (recall equation (3.131) for ΔG) when the mismatch occurs in the middle of the probe, and to lose the helix initiation parameter contribution if the mismatch occurs on the end of the probe. A single base mismatch in a probe automatically guarantees that the probe is not self-complementary (in the Watson-Crick sense); thus, if the original probe was self-complementary, the contribution from the symmetry term is lost as well.

One should also consider the situation where a matching probe P is *almost* self-complementary, with only one base violating the property. In that case, replacing the offending base appropriately would generate a self-complementary mismatch probe P'. In the computation of ΔG for the hybridization of P' with the target T, the mismatched dimer contributions will be lost, as discussed above, but the contribution of the symmetry term will be gained.

It is also possible for two strategically placed mismatches to turn a matching self-complementary probe into a mismatched self-complementary probe. Thus, the test for self-complementarity should be performed on each probe sequence from scratch, rather than being inferred from the self-complementarity status of the original probe and the editing changes.

More detailed treatment of mismatches

Thermodynamic contributions of different mismatched dimers have been studied as well (see [7], [37], [2], [4], [1], and [3]). These studies showed that different internal mismatches have different effects on the thermodynamic parameters of hybridization—some even stabilize the resulting duplex. One can make use of these available parameters to treat mismatches in much more detail. However, one must

be careful to keep in mind that the parameters for the internal (and some terminal) mismatches were derived using the stabilizing effect of neighboring matching base-pairs. As a result, these parameters may not have the same additive properties as the parameters for matching NN dimers. In any case, potential improvements in the accuracy of the resulting thermodynamic parameters must be weighed against the loss of speed due to more involved computations.

Appendix C

Appendices for Chapter 4

C.1 Exponential Limit Inequality: Proof

Claim: For large n,

$$\left(1-\frac{1}{n}\right)^n > e^{-1-\frac{1}{n}}. \tag{C.1}$$

Proof:

Inequality (C.1) is equivalent to

$$\ln\left[\left(1-\frac{1}{n}\right)^n\right] \overset{\text{want}}{>} -1-\frac{1}{n}. \tag{C.2}$$

Since the series expansion of the logarithm is given by

$$\ln(1-x) = -\sum_{j=1}^{\infty}\frac{x^j}{j} = -x-\frac{x^2}{2}-\frac{x^3}{3}-\cdots \qquad \text{for } |x|<1, \tag{C.3}$$

we can expand the left-hand side of (C.2) as follows:

$$\begin{aligned}\ln\left[\left(1-\frac{1}{n}\right)^n\right] &= n\ln\left(1-\frac{1}{n}\right) = n\left\{-\sum_{j=1}^{\infty}\frac{\left(\frac{1}{n}\right)^j}{j}\right\} \qquad \text{(by (C.3))}\\ &= -n\left\{\sum_{j=1}^{\infty}\frac{1}{jn^j}\right\} = -\sum_{j=1}^{\infty}\frac{1}{jn^{j-1}} = -1-\frac{1}{2n}-\frac{1}{3n^2}-\frac{1}{4n^3}-\cdots \end{aligned} \tag{C.4}$$

Thus, inequality (C.2) reduces to

$$-1 - \frac{1}{2n} - \frac{1}{3n^2} - \frac{1}{4n^3} - \cdots \overset{\text{want}}{>} -1 - \frac{1}{n} \tag{C.5}$$

$$\iff \quad \frac{1}{3n^2} + \frac{1}{4n^3} + \frac{1}{5n^4} + \cdots \overset{\text{want}}{<} -\frac{1}{2n} + \frac{1}{n} = \frac{1}{2n} \tag{C.6}$$

Now,

$$\begin{aligned} &\frac{1}{3n^2} + \frac{1}{4n^3} + \frac{1}{5n^4} + \cdots < \frac{1}{3n^2} + \frac{1}{3n^3} + \frac{1}{3n^4} + \cdots \\ &= \frac{1}{3n^2}\left(1 + \frac{1}{n} + \frac{1}{n^2} + \cdots\right) = \frac{1}{3n^2} \cdot \frac{1}{1 - \frac{1}{n}} \qquad \text{(geometric sum)} \\ &= \frac{1}{3n} \cdot \frac{1}{n-1} \overset{\text{want}}{<} \frac{1}{2n} \qquad \text{(by (C.6))} \end{aligned}$$

Simplifying yields

$$\begin{aligned} &\iff \quad \frac{1}{3(n-1)} \overset{\text{want}}{<} \frac{1}{2} \\ &\iff \quad 2 < 3(n-1) = 3n - 3 \\ &\iff \quad 5 < 3n, \end{aligned}$$

which holds for every $n \geq 2$. Retracing the chain of inequalities, we obtain

$$\boxed{\left(1 - \frac{1}{n}\right)^n > e^{-1-\frac{1}{n}} \quad \forall n \geq 2} \tag{C.7}$$

as desired.

C.2 Chernoff's Inequality: Proof

Claim:

$$\Pr\left(S(n,p) \leq (1-\epsilon)np\right) \leq \ e^{-\frac{\epsilon^2}{2}np}, \quad \epsilon \in (0,1). \tag{C.8}$$

Proof:

S is a Binomial random variable:

$$S(n,p) = X_1 + \cdots + X_n, \tag{C.9}$$

where X_i are i.i.d.r.v.'s with

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } (1-p) \end{cases}, \qquad i = 1, \ldots, n \tag{C.10}$$

Therefore,

$$\mathbf{E}(S) = \sum_{i=1}^{n} \mathbf{E}(X_i) = \sum_{i=1}^{n} p = np. \tag{C.11}$$

Since

$$\begin{aligned} & S \leq (1-\epsilon)np && \text{(C.12)} \\ \iff \quad & S - np \leq -\epsilon np \\ \iff \quad & \lambda(S-np) \leq -\lambda\epsilon np \quad \forall \lambda > 0 \\ \iff \quad & -\lambda(S-np) \geq \lambda\epsilon np \quad \forall \lambda > 0 \\ \iff \quad & e^{-\lambda(S-np)} \geq \ e^{\lambda\epsilon np} \quad \forall \lambda > 0, \end{aligned}$$

it follows that

$$\Pr\left(S \leq (1-\epsilon)np\right) = \Pr\left(e^{-\lambda(S-np)} \geq \ e^{\lambda\epsilon np} \right) \tag{C.13}$$

$$\leq \ \frac{\mathbf{E}\left[e^{-\lambda(S-np)} \right]}{e^{\lambda\epsilon np}} \qquad \text{(by Markov's inequality)} \tag{C.14}$$

For a proof of Markov's inequality, see, e.g., [5].

From (C.9) and (C.10), we know that

$$S - np = \sum_{i=1}^{n} X_i - np = \sum_{i=1}^{n} (X_i - p). \tag{C.15}$$

Therefore,

$$\begin{aligned} \mathbf{E}\left[e^{-\lambda(S-np)} \right] &= \mathbf{E}\left[e^{-\lambda \sum_i (X_i - p)} \right] \qquad (C.16) \\ &= \mathbf{E}\left[\prod_{i=1}^{n} e^{-\lambda(X_i - p)} \right] = \prod_{i=1}^{n} \mathbf{E}\left[e^{-\lambda(X_i-p)} \right] \qquad \text{(by independence)} \\ &= \left\{ \mathbf{E}\left[e^{-\lambda(X_1-p)} \right] \right\}^n \qquad \text{(by (C.10))} \end{aligned}$$

- $$\mathbf{E}\left[e^{-\lambda(X_1-p)} \right] = p\, e^{-\lambda(1-p)} + (1-p)\, e^{-\lambda(-p)} \tag{C.17}$$

$$\begin{aligned} &= e^{\lambda p}\left(p\, e^{-\lambda} + (1-p) \right) = e^{\lambda p}\big(1 + \underbrace{p(\, e^{-\lambda} - 1)}_{u}\big) \\ &\leq e^{\lambda p}\left(e^{p(e^{-\lambda}-1)} \right) \qquad (\text{since } 1 + u \leq e^u \ \forall u) \\ &= e^{p(e^{-\lambda}-1+\lambda)} \leq e^{p\frac{\lambda^2}{2}}, \end{aligned} \tag{C.18}$$

where the last inequality follows from

$$\begin{aligned} & e^{-\lambda} \leq 1 - \lambda + \frac{\lambda^2}{2} \quad \forall \lambda > 0 \\ \Longrightarrow \quad & e^{-\lambda} - 1 + \lambda \leq \frac{\lambda^2}{2} \end{aligned} \tag{C.19}$$

Therefore, by (C.16) and (C.18),

$$\mathbf{E}\left[e^{-\lambda(S-np)} \right] \leq \left(e^{p\frac{\lambda^2}{2}} \right)^n = e^{\frac{\lambda^2}{2} np} \tag{C.20}$$

and

$$\begin{aligned}\Pr\left(S \le (1-\epsilon)np\right) &\le \; e^{-\lambda\epsilon np}\; \mathbf{E}\left[\; e^{-\lambda(S-np)}\right] \\ &\le \quad e^{-\lambda\epsilon np}\; e^{\frac{\lambda^2}{2}np} \;=\; e^{np\left(\frac{\lambda^2}{2}-\lambda\epsilon\right)} \qquad \forall\lambda > 0 \end{aligned} \tag{C.21}$$

Since (C.21) holds for all $\lambda > 0$, it certainly holds for $\lambda = \lambda^*$ which minimizes the expression. Let

$$f(\lambda) = \; e^{np\left(\frac{\lambda^2}{2}-\lambda\epsilon\right)}$$

Optimizing over λ we find:

$$\begin{aligned} f'(\lambda) \;&=\; f(\lambda)\cdot np(\lambda-\epsilon) = 0 \qquad (C.22)\\ \iff \lambda^* \;&=\; \epsilon, \end{aligned}$$

so that

$$f(\lambda^*) = \; e^{np\left(\frac{\epsilon^2}{2}-\epsilon^2\right)} \;=\; e^{-\frac{\epsilon^2}{2}np}$$

and, from (C.21),

$$\boxed{\Pr\left(S \le (1-\epsilon)np\right) \le \; e^{-\frac{\epsilon^2}{2}np} \quad \forall\epsilon \in (0,1)} \tag{C.23}$$

as desired.

It remains to check that the optimizing $\lambda = \lambda^*$ is a minimum of $f(\lambda)$, that is, $f''(\lambda^*) > 0$. By (C.22),

$$\begin{aligned} f''(\lambda^*) \;&=\; f'((\lambda)\cdot np(\lambda-\epsilon) + f(\lambda)\cdot np\;|_{\lambda=\lambda^*} \\ &=\; \underbrace{f'(\lambda^*)}_{0}\cdot np(0) + f(\lambda^*)\cdot np \\ &=\; np\; e^{-\frac{\epsilon^2}{2}np} > 0. \end{aligned}$$

$$\therefore \lambda^* = \epsilon \quad \text{is a minimum.}$$

Bibliography

[1] Allawi, H.T. and SantaLucia, J.Jr. Nearest Neighbor Thermodynamic Parameters for Internal G-A Mismatches in DNA. *Biochemistry* **37**(8), 2170–2179, 1998.

[2] Allawi, H.T. and SantaLucia, J.Jr. Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA: Sequence Dependence and pH Effects. *Biochemistry* **37**(26), 9435–9444, 1998.

[3] Allawi, H.T. and SantaLucia, J.Jr. Thermodynamics and NMR of Internal G-T Mismatches in DNA. *Biochemistry* **36**(34), 10581–10594, 1997.

[4] Allawi, H.T. and SantaLucia, J.Jr. Thermodynamics of Internal C-T Mismatches in DNA. *Nucleic Acids Research* **26**(11), 2694–2701, 1998.

[5] Alon, N. and Spencer, J.H. *The Probabilistic Method,* 2nd edition. John Wiley & Sons, Inc., New York, 2000.

[6] Balazs, I., Beekman, J., Neuweiler, J., Liu, H., Watson, E., and Ray, B. Molecular Typing of HLA-A, -B, and DRB Using a High Throughput Micro Array Format. *Human Immunology* **62**: 850–857, 2001.

[7] Bommarito, S., Peyret, N., and SantaLucia, J.Jr. Thermodynamic Parameters for DNA Sequences with Dangling Ends. *Nucleic Acids Research* **28**(9), 1929–1934, 2000.

[8] Borneman, J., Chrobak, M., Della Vedova, G., Figueroa, A., and Jiang, T. Probe Selection Algorithms with Applications in the Analysis of Microbial Communities. *Bioinformatics* **17**(Suppl. 1): S39–S48, 2001.

[9] Breslauer, K.J., Frank, R., Blocker, H., and Marky, L.A. Predicting DNA Duplex Stability from the Base Sequence. *PNAS USA* **83**, 3746–3750, 1986.

[10] Cantor, C.R. and Schimmel, P.R. *Biophysical Chemistry Part III: The Behavior of Biological Macromolecules.* Freeman, San Francisco, 1980.

[11] Cao, K., Chopek, M., and Fernandez-Vina, M.A. High and intermediate resolution DNA typing systems for class I HLA-A, B, C genes by hybridization with sequence-specific oligonucleotide probes (SSOP). Review. *Reviews in Immunogenetics* **1**(2): 177–208, 1999.

[12] Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B. Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data. NYU CS Technical Report 2003-845 (New York University, New York), 2003. Available for download from `http://www.cs.nyu.edu/csweb/Research/technical_reports.html`

[13] Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B. Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data. *Proceedings of the National Academy of Sciences, USA* **100**(17): 9668–9673, 2003.

[14] Chu, S., DeRisi, J.L., Eisen, M.B., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. The Transcriptional Program of Budding Yeast. *Science* **282**: 699–705, 1998.

[15] Consolandi, C., Busti, E., Pera, C., Delfino, L., Ferrara, G.B., Bordoni, R., Castiglioni, B., Rossi Bernardi, L., Battaglia, C., and De Bellis, G. Detection of HLA Polymorphisms by Ligase Detection Reaction and a Universal Array Format: a Pilot Study for Low Resolution Genotyping. *Human Immunology* **64**: 168–178, 2003.

[16] Cover, T.M. and Thomas, J.A. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, 1991.

[17] Delfino, L., Morabito, A., Longo, A., and Ferrara, G.B. HLA-C High Resolution Typing: Analysis of Exons 2 and 3 by Sequence Based Typing and Detection of Polymorphisms in Exons 1–5 by Sequence Specific Primers. *Tissue Antigens* **52**(3): 251–259, 1998.

[18] DeRisi, J.L., Iyer, V.R., and Brown, P.O. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**: 680–686, 1997.

[19] *Dictionary of Algorithms and Data Structures* at NIST (National Institute of Standards and Technology) website: `http://www.nist.gov/dads/`

[20] Egan, J.P. *Signal Detection Theory and ROC analysis.* Academic Press, New York, 1975.

[21] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. Cluster Analysis

and Display of Genome-wide Expression Patterns. *Proceedings of the National Academy of Sciences, USA* **95**: 14863–14868, 1998.

[22] Ferrara, G.B., Masulli, F., Pera, C., Rovetta, S., and Sensi, R. Fuzzy Modeling for HLA Typing.
`www-dii.ing.unisi.it/aiia2002/paper/BIOINFO/rovetta-aiia02.pdf`
In *AI*IA 2002: Advances in Artificial Intelligence, 8th Congress of the Italian Association for Artificial Intelligence, Siena, Italy, September 10–13, 2002.* Proceedings not yet published.

[23] Hoffman, K. Stein Estimation - A Review. *Statistical Papers*, **41(2)**: 127–158, 2000.

[24] `http://www.affymetrix.com/`

[25] James, W. and Stein, C. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium Mathematical Statistics and Probability*, (ed. Neyman, J.), Vol. **1**: 361–379. University of California Press, 1961.

[26] Kaderali, L. and Schliep, A. An Algorithm to Select Target Specific Probes for DNA Chips. `citeseer.nj.nec.com/kaderali01algorithm.html`, 2001.

[27] Kluger, Y., Yu, H., Qian, J., and Gerstein, M.B. Relationship between gene co-expression and probe localization on microarrays. *Preprint, submitted to Nature Genetics.* Abstract on http://bioinfo.mbb.yale.edu/∼kluger/pipeline/KLUGERetal_NG.pdf

[28] Krause, A., Krautner, M., and Meier, H. Accurate Method for Fast Design of Diagnostic Oligonucleotide Probe Sets for DNA Microarrays. In *Sec-*

ond IEEE International Workshop on High Performance Computational Biology (HiCOMB'2003), Nice, France, April 22, 2003 Online Proceedings `http://hpc.eece.unm.edu/HiCOMB/proceedings.html`

[29] Li, F. and Stormo, G.D. Selection of Optimal DNA Oligos for Gene Expression Arrays. *Bioinformatics* **17**(11): 1067–1076, 2001.

[30] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., and Horton, H. *et al.* Expression Monitoring by Hybridization to High-density Oligonucleotide Arrays. *Nature Biotechnology* **14**: 1675–1680, 1996.

[31] Luby, M. A Simple Parallel Algorithm for the Maximal Independent Set Problem. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing (STOC'85)*: 1–10, 1985.

[32] Mishra, B. Placing Oligonucleotide Probes on a Planar Surface to Reduce Competition among Neighboring Probes. *Unpublished Draft*, 2003.

[33] Naef, F., Lim, D.A., Patil, N., and Magnasco, M. DNA hybridization to mismatched templates: A chip study. *Physical Review E* **65**(4), 040902, 2002.

[34] *National Marrow Donor Program* `http://www.nmdpresearch.org/`

[35] Noreen, H.J., Yu, N., Setterholm, M., Ohashi, M., Baisch, J., Endres, R., Fernandez-Vina, M., Heine, U., Hsu, S., and Kamoun, M. *et al.* Validation of DNA-based HLA-A and HLA-B testing of volunteers for a bone marrow registry through parallel testing with serology. *Tissue Antigens* **57**(3): 221–229, 2001.

[36] Pera, C., Delfino, L., Morabito, A., Longo, A., Johnston-Dow, L., White, C.B., Colonna, M., and Ferrara, G.B. HLA-A Typing: Comparison Between Serology, the Amplification Refractory Mutation System with Polymerase Chain Reaction and Sequencing. *Tissue Antigens* **50**(4): 372–379, 1997.

[37] Peyret, N., Seneviratne, P.A., Allawi, H.T., and SantaLucia, J.Jr. Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches. *Biochemistry* **38**(12), 3468–3477, 1999.

[38] Qian, J., Kluger, Y., Yu, H., and Gerstein, M.B. Spatial artifacts in microarray data: spurious correlations and techniques to remove them. *Preprint, submitted to Biotechniques.* Abstract on http://bioinfo.mbb.yale.edu/∼kluger/pipeline/QYK_artifact.pdf

[39] Rash, S. and Gusfield, D. String Barcoding: Uncovering Optimal Virus Signatures. In *Proceedings of the sixth annual international conference on Computational biology (RECOMB'02)*: 254–261, 2002. ACM Press.

[40] SantaLucia, J. Jr. A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-neighbor Thermodynamics. *PNAS USA* **95**, 1460–1465, 1998.

[41] SantaLucia, J.Jr., Allawi, H.T., and Seneviratne, P.A. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry* **35**(11), 3555–3562, 1996.

[42] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**(5235): 467–470, 1995.

[43] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. Parallel Human Genome Analysis: Microarray-based Expression Monitoring of 1000 Genes. *Proceedings of the National Academy of Sciences, USA* **93**: 10614–10619, 1996.

[44] Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell* **106**: 697–708, 2001.

[45] Sokal, R.R. and Michener, C.D. A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Scientific Bulletin* **38**: 1409–1438, 1958.

[46] Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., and Lepage, M. *et al.* Design and Implementation of Microarray Gene Expression Markup Language (MAGE-ML). *Genome Biology* **3(9)**: research0046.1–0046.9, 2002.

[47] Spellman, P.T., Sherlock, G., Iyer, V.R., Zhang, M., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**: 3273–3297, 1998.

[48] Wolfram, S. *The Mathematica Book.* Cambridge University Press, 4th edition, 1999.

[49] Yu, H., Kluger, Y., Qian, J., and Gerstein, M.B. The effect of chromosome

structure on gene expression. *Preprint, submitted to Nature Biotechnology.* Abstract on http://bioinfo.mbb.yale.edu/∼kluger/pipeline/YQK_NB.pdf