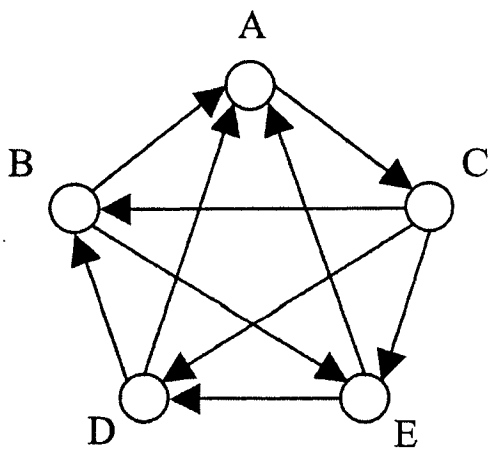
Investigations into the Design and Dissection of Genetic Networks



Eric Libby

Doctor of Philosophy
Department of Physiology

McGill University
Montreal, Canada
May 27, 2007

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-38606-4
Our file Notre référence
ISBN: 978-0-494-38606-4

ABSTRACT

The sequencing of the human genome revealed that the number of genes does not explain why humans are different from other organisms like mice and dogs. Instead, it is how genes interact with each other and the environment that separates us from other organisms. This motivates the study of genetic networks and, consequently, my research. My work delves into the roles that simple genetic networks play in a cell and explores the biotechnological aspects of how to uncover such genes and their interactions in experimental models.

Cells must respond to the extracellular environment to contract, migrate, and live. Cells, however, are subject to stochastic fluctuations in protein concentrations. I investigate how cells make important decisions such as gene transcription based on noisy measurements of the extracellular environment. I propose that genetic networks perform Bayesian inference as a way to consider the probabilistic nature of these measurements and make the best decision. With mathematical models, I show that allosteric repressors and activators can correctly infer the state of the environment despite fluctuating concentrations of molecules. Viewing transcriptional networks as inference modules explains previous experimental data. I also discover that the particular inference problem determines whether repressors or activators are better.

Next, I explore the genetic underpinnings of two canine models of atrial fibrillation: atrial tachypacing and ventricular tachypacing. Using Affymetrix microarrays, I find that the genetic signatures of these two models are significantly different both in magnitude and in class of genes expressed. The ventricular tachypacing model has thousands of transcripts differentially expressed with little

overlap between 24 hours and 2 weeks, suggesting independent mechanisms. The atrial tachypacing model demonstrates an adaptation as the number of genes found changed decreases with increasing time to the point that no genes are changed at 6 weeks. I use higher level analysis to find that extracellular matrix components are among the most changed in ventricular tachypacing and that genes like connective tissue growth factor may be responsible.

Finally, I generalize the main problem of microarray analysis into an evaluation problem of choosing between two competing options based on the scores of many independent judges. In this context, I rediscover the voting paradox and compare two different solutions to this problem: the sum rule and the majority rule. I find that the accuracy of a decision depends on the distribution of the judges' scores. Narrow distributions are better solved with a sum rule, while broad distributions prefer a majority rule. This finding motivates a new algorithm for microarray analysis which outperforms popular existing algorithms on a sample data set and the canine data set examined earlier. A cost analysis reveals that the optimal number of judges depends on the ratio of the cost of a wrong decision to the cost of a judge.

ABRÉGÉ

Le séquençage du génome humain a révélé que le nombre de gènes n'explique pas pourquoi les humains se distinguent des autres organismes comme les souris et les chiens. Plutôt, c'est la manière avec laquelle les gènes communiquent entre eux et l'environnement qui nous sépare des autres organismes. Cela motive l'étude des réseaux génétiques et, par conséquent, ma recherche. Mon travail examine les rôles joués par les réseaux génétiques simples dans une cellule et explore les aspects biotechnologiques qui nous permettrons de dévoiler tels gènes et leurs réseaux dans des modèles expérimentaux.

Les cellules doivent répondre à l'environnement extracellulaire pour se contracter, migrer et vivre. Cependant, les cellules sont soumises aux fluctuations stochastiques des concentrations protéiques. Je dissèque comment les cellules prennent des décisions importantes, par exemple la transcription de gène, basée sur les mesures bruyantes de l'environnement extracellulaire. Je propose que les réseaux génétiques exécutent l'inférence Bayesienne comme méthode d'analyse pour la nature probabiliste de ces mesures et prise de décisions. Avec des modèles mathématiques, je démontre que les répresseurs et activateurs allostériques peuvent correctement déduire l'état environnemental malgré les concentrations fluctuantes des molécules. Définir les réseaux de transcriptions en tant que modules d'inférences explique des données expérimentales précédentes. Je découvre également que le problème d'inférence particulier détermine si les répresseurs ou les activateurs sont meilleurs.

Ensuite, j'explore les fondements génétiques de deux modèles canins pour la fibrillation artérielle : la tachycardie sinusale et la tachycardie ventriculaire.

En utilisant des puces d'Affymetrix, je constate que les signatures génétiques de ces deux modèles sont différentes de façon significative tant dans la couverture que dans la classe des gènes exprimés. Le modèle pour la tachycardie ventriculaire a des milliers de transcrits exprimées différentiellement avec peu de chevauchement entre les échantillons de 24 heures et 2 semaines, sous-entendant des mécanismes indépendants. Le modèle pour la tachycardie sinusale démontre une adaptation par le nombre de gènes trouvés changées diminues durant une période d'observation croissante, au point qu'aucun gène n'est changé à 6 semaines. J'utilise l'analyse de niveau supérieur pour constater que les composantes de la matrice extracellulaire sont parmi les plus variantes dans le modèle pour la tachycardie ventriculaire où un des gènes candidats responsables pour ceci est le facteur de croissance de tissu conjonctif.

Finalement, je généralise le problème principal d'analyse de microarray en un problème d'évaluation pour lequel un choix est fait entre deux options en concurrence basées sur des résultats donnés par plusieurs juges indépendants. Dans ce contexte, je retrouve le paradoxe de Condorcet et deux solutions différentes à ce problème sont comparées : la règle de somme et la règle majoritaire. Je constate que l'exactitude d'une décision dépend sur la distribution des résultats. Les distributions étroites sont mieux résolues avec la règle de somme, alors que les larges distributions préfèrent la règle majoritaire. Cette conclusion incite un nouvel algorithme pour l'analyse de microarray qui remporte les résultats obtenus par les algorithmes communs utilisés jusqu'à présent. Ceci a été démontré en regardant un ensemble de données échantillons et l'ensemble de données canin examiné plus tôt. Une analyse des coûts révèle que le nombre optimal de juges dépend sur le rapport du prix d'une mauvaise décision au coût d'un juge.

CONTRIBUTIONS OF AUTHORS

Chapters 2 and 3 stem from published articles that have been modified for the layout of this thesis. Chapter 5 is a submitted manuscript and Chapters 4 and 6 contain material that may appear in a later manuscript.

Chapter 2 is work done with Ted Perkins and Peter Swain and is published in the *Proceedings of the National Academy of Science.* I generated and analyzed the data presented in figures 1-3 as well as prepared the figures. I also assembled the material associated with these figures found in the appendix to Chapter 2. Ted Perkins did the analysis and preparation of all material related to figure 4. Although Ted Perkins and I helped in writing the manuscript, Peter Swain wrote the majority of the manuscript and supervised the project.

Chapter 3 is work published in *Circulation Research* with authors Sophie Cardin, Eric Libby, Patricia Pelletier, Sabrina Le Bouter, Akiko Shiroshita-Takeshita, Nolwenn Le Meur, Jean Leger, Sophie Demolome, Andre Ponton, Leon Glass, and Stanley Nattel. I performed the microarray analysis and generated all associated data and figures, including the cumulative rank diagrams and the protein interaction network. Sophie Cardin did all of the experimental work on the canine models including the hemodynamic measurements. Patricia Pelletier helped with the classification of the probesets. Sophie Demolombe, Akiko Shiroshita-Takeshita, and Sabrina Le Bouter ran the RT-PCR and Western blot experiments. Andre Ponton conducted the microarray experiment. Jean Leger and Nolwenn Le Meur offered advice on their preferred method of microarray analysis. I wrote sections of the manuscript, but the majority of the writing was done by

Stanley Nattel. Leon Glass supervised my role in the project, and Stanley Nattel supervised the whole collaboration.

I am responsible for all of the work, figures, and writing presented in Chapters 4 and 6. These may support a future manuscript submitted with Leon Glass, who supervised all of this work.

Chapter 5 is a submitted manuscript with the authors Eric Libby and Leon Glass. I produced and analyzed the data for this manuscript as well as created the algorithm for microarray analysis. I also prepared the initial figures, which were sent to Tom Inoue to improve their quality. I wrote the manuscript with the help of Leon Glass and assembled the material in the appendix to Chapter 5. Leon Glass wrote and assembled the section on Latin square score-sheets, but the mathematical proof was a joint effort. Leon Glass supervised the entire project and offered critical input through every stage.

CLAIMS TO ORIGINALITY

I. Chapter 2

We proposed the new paradigm that genetic networks perform Bayesian inference to determine the state of their environment. We showed via deterministic equations and stochastic models that transcriptional regulatory networks can accomplish this task. This model accounted for previous experimental results which the existing logic gate paradigm could not explain.

II. Chapter 3

We contrasted two canine models of atrial fibrillation using canine genetic microarrays. We predicted adaptation in an atrial tachypacing model based on data from 24 hours and 1 week and then confirmed this hypothesis with data from 6 weeks of tachypacing. We showed distinct transcriptional portraits of the two models and demonstrated differential regulation of extracellular matrix components which may produce the fibrosis observed with ventricular tachypacing. We used novel techniques for analyzing and presenting the data including a cumulative rank diagram with a Monte Carlo test for significance. We also assembled a protein network which relates the genetic findings with the phenotypic observations and points to a putative role for connective tissue growth factor.

III. Chapter 5

We generalized the problem of combining probes from microarray data into a more general class of evaluation problems. We rediscovered the voting paradox and proved that Latin square score-sheets produce a cycle that

spans the length of the number of competitors. We showed that when the distribution of the judges' scores are narrow, a sum rule affords higher accuracy than a majority rule. Conversely, when the scores are broadly distributed, a majority rule is better. Based on these results, we created a novel algorithm for microarray analysis that outperforms previous methods on a validation data set. We analyzed the costs of evaluations and revealed that the optimal number of judges depends on the ratio between the cost of a judge and the cost of an incorrect evaluation.

IV. Chapters 4 and 6

We explored and solved problems with microarray analysis that had not been sufficiently addressed including the difficulties annotating probesets and the lack of clarity in the analysis. We applied our algorithm to the original canine data from Chapter 3 and showed regulated pathways and better agreement with the RT-PCR data. This demonstrated that our algorithm could perform on par with existing algorithms when analyzing data with biological variation instead of validation data sets.

ACKNOWLEDGEMENTS

It has been my honor to work with Dr. Leon Glass. He has shaped my approach to science and guided my course of research through my doctoral studies. His enthusiasm, broad academic interest, and passion for life have been an inspiration to me. Because of him, I have gone cross-country skiing, climbed mountains in the Adirondacks, eaten cassoulet, and met distinguished scholars from around the world. Having the chance to talk and work with him has been the best part of my choice to come to McGill University.

I thank the other professors of the Center for Nonlinear Dynamics. It has been enriching to be a part of such a talented and renowned group of scholars. Most of all, I thank Dr. Peter Swain for his friendship and guidance. Beyond squash, Pino's heart-jammer, and white socks, it has been rewarding working with him and discussing lectures and papers.

I also thank my collaborators: Dr. Stanley Nattel, Dr. Ted Perkins, Sophie Cardin, and Patricia Pelletier. The quality of this work would not have been possible without them.

My academic life has been possible because of the love and support of my parents. I thank them for all their emails, phone calls, prayers, care packages, and visits. They made it impossible to forget that I always have my family– even if I am surrounded by frozen tundra.

In this same vein, I thank Nicole Joy for all of her love and support. She has kept me as sane as I can possibly be throughout my doctoral studies. By her baking, pilates, passion for literature, and love of music she has been a constant reminder that there is a lot to do in life.

For whatever social life I have, I thank my friends. Without soccer, softball, karate, Thomson House, sangria, Japanese food, movies, and parties, my doctoral studies may have been shorter but definitely not as enjoyable.

I thank MITACS, NSERC, CIHR, and whatever financial powers have sheltered and fed me.

On a strange note, I thank all the computers I have used: hyper, cnd184, dragoon, Augustus, and Number Five. I probably spent more time with them than people and have developed a special working relationship with each one. Although I rarely kept my Desktop clean or optimized, they performed admirably.

Lastly, I thank every one for their understanding that a short acknowledgement section in a thesis can not properly express my gratitude.

TABLE OF CONTENTS

ABSTRACT i

ABRÉGÉ iii

CONTRIBUTIONS OF AUTHORS v

CLAIMS TO ORIGINALITY vii

ACKNOWLEDGEMENTS ix

LIST OF TABLES xv

LIST OF FIGURES xvi

1 Chapter 1: Introduction 1

1.1 Affymetrix Microarray Analysis 2
1.1.1 Microarray Introduction 2
1.1.2 Affymetrix Microarrays 3
1.1.3 Microarray Analysis Overview 4
1.1.4 Preprocessing 5
1.1.5 The Expression Index 8
1.1.6 Differential Gene Selection 9
1.1.7 Other Algorithms 11
1.1.8 Comparisons 12
1.2 Atrial Fibrillation 13
1.2.1 Atrial Fibrillation Introduction 13
1.2.2 Atrial Fibrillation Mechanisms 14
1.2.3 Ionic Remodelling 15
1.2.4 Atrial Tachypacing Model 17
1.2.5 Ventricular Tachypacing model 18
1.2.6 Structural Remodelling 19
1.2.7 Previous Microarray Analysis 20
1.3 Models of Genetic Networks 23
1.3.1 Gene Expression: The *lac* Operon 23
1.3.2 Modeling Gene Networks 25
1.3.3 Allostery 28
1.3.4 Noise 29

1.3.5 Managing Noise 30
1.4 Research Objectives 31

2 Chapter 2: Gene Networks as Inference Modules 33

2.1 Abstract 34
2.2 Introduction 34
2.3 Cis-regulatory Regions as Inference Modules 39
2.4 Inference in the *lac* Operon 43
2.5 Discussion 45
2.6 Materials and Methods 48
2.6.1 Modeling Genetic Networks 48
2.6.2 Comparison of the Models 49
2.6.3 Stochastic Simulation 50
2.6.4 Fitting a Posterior Probability to an Operon 51
2.7 Acknowledgments 52
2.8 Appendix 52
2.8.1 Allosteric Model of Transcription Factors 52
2.8.2 Promoter Models 55
2.8.3 Generating the posteriors 57
2.8.4 Fitting the Models to the Posteriors 57
2.8.5 Parameter Sensitivity 59
2.8.6 Robustness of the Best-fit Parameters 60
2.8.7 Stochastic Simulation Details 61
2.8.8 The Inverse Gaussian Classification Problem 64
2.8.9 Fitting the Transcription Rate Surface from the *lac* Operon 65

3 Chapter 3: Gene Networks In Atrial Fibrillation 67

3.1 Abstract 68
3.2 Introduction 69
3.3 Materials and Methods 70
3.3.1 Animal Model and Preparation 70
3.3.2 In Vivo Measurements and Cell Isolation 72
3.3.3 RNA Extraction 73
3.3.4 Canine Genome Microarrays 74
3.3.5 RNA Processing on Arrays 75
3.3.6 Statistical Analysis of Microarray Data 76
3.3.7 Real-time RT-PCR 77
3.3.8 Western Blot Analysis 78
3.4 Results 79
3.4.1 Hemodynamics and Electrophysiology 79
3.4.2 Microarray Findings 79

3.4.3 Real-time RT-PCR Results . . . 88
3.4.4 Western-blot Results . . . 89
3.5 Discussion . . . 94
3.5.1 Relationship to Previous Findings . . . 94
3.5.2 Relevance to Mechanisms of AF-related Remodeling . . . 95
3.5.3 Potential Significance . . . 97
3.5.4 Potential Limitations . . . 98

4 Chapter 4: Problems and Properties . . . 101

4.1 Problems . . . 102
4.1.1 Problem 1: Lack of Clarity . . . 102
4.1.2 Problem 2: Probeset Annotations . . . 103
4.1.3 Problem 3: Probesets in Excess and Dearth . . . 104
4.2 Properties . . . 105
4.2.1 Performance . . . 105
4.2.2 Noise . . . 108
4.2.3 Normality . . . 108

5 Chapter 5: Evaluation Theory . . . 113

5.1 Abstract . . . 114
5.2 Figure Skating and the Voting Paradox . . . 115
5.3 Distribution of Microarray Probe Intensities . . . 117
5.4 Analysis of Evaluation: Accuracy and Cost . . . 118
5.5 Majority Rule Algorithm for Microarray Analysis . . . 119
5.6 Cost Analysis for Figure Skating . . . 122
5.7 Discussion . . . 123
5.8 Acknowledgement . . . 124
5.9 Appendix A: Evaluation Accuracy . . . 124
5.9.1 Individual Judge Accuracy . . . 124
5.9.2 Sum Rule . . . 125
5.9.3 Majority Rule Accuracy Calculations . . . 125
5.10 Appendix B: Cost-benefit Analysis . . . 126
5.10.1 Cost-benefit Analysis: Figure Skating . . . 126
5.10.2 Cost-benefit Analysis: Other Examples . . . 127
5.11 Appendix C: Latin Square Tournaments . . . 130
5.11.1 Introduction . . . 130
5.11.2 Basic Properties of Tournaments . . . 131
5.11.3 Basic Properties of Latin Square Tournaments . . . 133
5.11.4 Hamiltonian Cycles in Latin Square Tournaments . . . 135

6 Chapter 6: Algorithm Applications 137

6.1 Solutions to Microarray Analysis Problems 138
6.1.1 Solution 1: Clarity 138
6.1.2 Solution 2: Annotations 139
6.1.3 Solution 3: Multiple Probesets 140
6.2 Application of PBP to the Canine Data 141
6.2.1 Mapping Probes to Genes 141
6.2.2 RT-PCR Comparison 143
6.2.3 Higher Level Analysis 144

7 Chapter 7: Conclusions 147

REFERENCES 151

LIST OF TABLES

Table		page
2–1	Parameter sensitivities for repressor and activator models	60
2–2	Parameter values for simulations	62
2–3	Comparison of scores versus averaging times	63
3–1	Hemodynamic and electrophysiological changes 1	80
3–2	Hemodynamic and electrophysiological changes 2	81
3–3	RT-PCR confirmations VTP	90
3–4	RT-PCR confirmations ATP	91
3–5	Western confirmations	93
5–1	Cost-benefit analysis applications	129
5–2	Latin square score-sheet with $n = 3$	131
5–3	Latin square score-sheet with $n = 7$ with $s_1 = 1$ and $s_7 = 5$	131
6–1	Changed genes from canine data set by PBP	143
6–2	RT-PCR comparison between MBEI-SAM and PBP	144

LIST OF FIGURES

Figure		page
2–1	A two-state classifier problem and its Bayesian solution	37
2–2	Regulatory models and fitting results	41
2–3	Stochastic simulations of two-state inference	44
2–4	Higher dimensional inference	46
2–5	Reaction scheme	53
2–6	Collection of posteriors	58
2–7	Robustness of fits to parameter changes	61
3–1	Scatterplots of ATP (24 hr, 1 wk) and VTP (24 hr, 2 wk)	82
3–2	Scatterplots of 6 wk ATP and ATP sham	83
3–3	Probesets changed in each classification group	85
3–4	Cumulative rank distribution for functional groups in VTP	87
3–5	Cumulative rank distribution for functional groups in ATP	88
3–6	Correlation between microarray and real-time PCR	89
3–7	Western blots	92
3–8	Protein interaction diagram	99
4–1	Intensity distributions for set concentrations	106
4–2	Percent of probes detecting fold changes	107
4–3	Replicate noise versus inter-probe variability	109
4–4	Normalization effects	110
5–1	The voting paradox	116
5–2	Distributions of figure skating scores and microarray probes	118

5–3 Accuracy and cost for different scoring distributions 120
5–4 ROC curves comparing microarray analysis algorithms 122
5–5 Sample score distributions for judges 126
5–6 Digraph for Table 5-3 132

CHAPTER 1
Introduction

This thesis presents the results of three distinct projects which revolve around the theme of the uncovering and understanding of genetic networks. To provide adequate background for each section in the thesis, I have separated the introduction into three components. The introduction begins with a review of genetic microarray technology, focusing particularly on Affymetrix microarrays. The review covers not only the design and use of microarrays, but also the analysis of the data returned from a microarray experiment and methods to discover differentially expressed genes. The next section introduces the cardiac arrhythmia atrial fibrillation. I explore the two different canine models of atrial fibrillation that are further investigated in Chapter 3 and the previous results from studying these models. I conclude this section by briefly presenting the results of previous microarray studies of atrial fibrillation. The third section of the introduction looks at the work delving into the functional properties of genetic networks. This begins with a discussion on the complexity of genetic regulation and uses the *lac* operon as a case study. I then look at different approaches to modeling transcription. I end the section with a discussion of noise and how cells are thought to control it. Finally, I describe the objectives of my thesis work and the specific hypotheses addressed.

1.1 Affymetrix Microarray Analysis

1.1.1 Microarray Introduction

Genetic microarrays measure the expression of tens of thousands of mRNA transcripts simultaneously with probes made of short sequences of DNA affixed to a glass chip [1, 2]. This technology enables researchers to take a snapshot of the expression of an entire genome in a sample of cells. Although it was developed in the late 1990s, at its foundations is the work of Ed Southern who showed decades earlier that nucleic sequences could be quantified with matching nucleic probes [1, 3]. This work inspired the Southern, Northern, and Western blots which measure the concentration of a single DNA sequence, mRNA transcript, or protein, respectively [4, 5]. Genetic microarrays arose later following advances in the high-density synthesis of nucleic probes, fluorescence-based detection, and the use of non-porous solid supports [1]. With the large-scale sequencing of genomes, such as the human genome, researchers had at their disposal databases containing the sequences of tens of thousands of genes [6]. Thus, with known gene sequences and the technology to construct microarrays available, the first microarrays permitted researchers to perform the work of thousands of blots in a single step [6, 7].

The early genetic microarrays were cDNA arrays which used one probe per gene and were fabricated in individual labs [1]. Later, companies like Affymetrix developed high-density oligonucleotide microarrays that used several probes per gene and offered more coverage of the genome [7, 8]. Microarrays appeared in studies classifying biological states, such as mitosis, by their unique transcriptional profiles with numerical techniques like clustering [9]. Other studies looked for individual gene candidates that exhibited differential regulation after the administration of a drug [10, 11]. They have also been used to chart the temporal patterns of gene regulation in the yeast cell cycle [12]. While microarrays can be

used in different types of analysis, we will focus only on their use in discovering differentially expressed genes. As microarrays gained in popularity, many problems surfaced with the reliability and analysis of the results [1, 13]. Currently, results found with microarrays need to be confirmed with other techniques; for example, a selection of changed genes are often validated using reverse-transcriptase polymerase chain reaction (RT-PCR) [14, 15]. Microarrays, therefore, often serve as hypothesis generating mechanisms that highlight genes and pathways for future exploration.

1.1.2 Affymetrix Microarrays

The Affymetrix microarray is one of the most popular microarray platforms [6, 16]. In the area of about one square inch, Affymetrix uses a combination of light directed synthesis and alternating masks to assemble hundreds of thousands of unique 25 base probes on a chip, one base at a time [7]. Affymetrix organizes the microarray into a two dimensional grid where each position contains hundreds of thousands to millions of copies of a given probe [7]. They assign each transcript measured on the microarray a probeset, which is a collection of 10-20 unique probes located at different places on the chip [7, 17]. Each probe of the probeset targets a different section of the transcript's sequence. Since these probes are perfect complements of the target sequences, they are named perfect match probes. Next to each perfect match probe on the grid lies a mismatch probe, identical in sequence except an altered 13th base. The mismatch probes supposedly measure non-specific binding and can serve as a measure of background noise [7]. In practice, they serve as weaker and less consistent versions of the perfect match probes and can even have a higher intensity [18, 19, 20].

A microarray experiment is a multi-step process that is typically performed over a long period of time as a collaboration of multiple labs [21, 22]. Here, I organize the process into three separate stages. Although other groups may divide the process differently [21, 22], this scheme fits the context of this thesis (for full details of the methodology, see [22]). In the first stage, the researchers perform their experiments on a group of cells. After the experiments finish, they select and extract the cells that they want to study with the microarray. These cells are stored until all of the cells for the microarrays have been collected and are ready for hybridization. The second stage begins with the fragmentation of the cells and the corresponding extraction of mRNA. Through a series of RT-PCR experiments, the total amount of mRNA is both amplified and fluorescently tagged. The resulting cRNA, complementary to the original mRNA, is fragmented via sonication or hydrolysis to produce pieces roughly 200 bases in size. In the final stage, the labelled cRNA solution is poured over the microarray and gently mixed. The solution remains on the chip to allow the cRNA time to hybridize to its specific probes. After a set incubation time, the microarrays receive a washing with a salt solution to remove any unbound cRNA, as well as cRNA bound to the wrong probes. A scanner records the fluorescent intensity at each spot on the chip, which corresponds to the amount of cRNA bound at that location. At the end of the microarray experiment, the scanner returns a list of fluorescent intensities for subsequent analysis.

1.1.3 Microarray Analysis Overview

Despite the decade of microarray use, there is still no established method for determining differentially expressed genes [16, 20]. The process of transforming

the raw microarray data into a list of differentially expressed genes has been divided into as many as six steps [23]. At each step there are multiple competing algorithms, often with multiple options, producing combinatorially many different ways to analyze the data [23]. Although many papers have compared algorithms for a specific step [17, 23, 24], the results often depend on the choice of algorithms used in other steps along with the validation data, itself. To complicate matters, there is no perfect validation data set, and the algorithms interact in complex ways with steps before and after them [24, 25]. Thus, there is no clear choice for methodology.

Amidst the choices, there are three suites of algorithms, however, that have gained popularity and appear in many of the papers comparing methodology [20, 24, 26, 27] . The three algorithms are Microarray Analysis Suite 5.0 (MAS 5.0) [28], Model Based Expression Index (MBEI) [29, 30], and Robust Multichip Average (RMA) [18, 24]. Each of these algorithms breaks the process of selecting differentially expressed genes into three distinct steps: 1. preprocessing the data, 2. formulating an expression index, and 3. selecting significantly changed expression indices. Technically, RMA and MBEI refer to algorithms used only in the second step, but both are used with particular algorithms in the first step to the extent that the names encompass both.

1.1.4 Preprocessing

The intensities returned from a microarray experiment depend on a number of factors other than the concentration of the transcripts and the thermodynamics governing the hybridization. The lab doing the microarray analysis, the day of the experiment, the particular scanner, and even the specific arrays can bias the

fluorescent intensities such that replicate arrays can have significantly different average intensities [17, 31]. While issues like lab and day can be controlled in a particular experiment, inter-array variability due to array construction, scanning effects, and pipette errors cannot. Moreover, microarray experiments are expensive and typically consist of just a handful of microarrays for control and experimental groups [13]. Thus, variability causing a particular array to have higher/lower intensities can severely bias the results [13]. Researchers have developed numerous normalization techniques to mitigate this variability [17, 25, 32]. The normalization method determines the ultimate selection of differentially expressed genes more than any other step in the process [16, 25].

While the normalization techniques differ in their approaches, they all share the common goal of reducing inter-array variation. The MAS 5.0 suite usually employs both a background correction and a scaling [28]. For the background correction, the microarray is divided into equally-spaced zones. In each zone, the average background is calculated based on the lowest 2% of probe intensities in that zone [33]. MAS 5.0 then computes the distance of each probe to the center of the zones and uses that to weight the average background experienced by that probe [33]. After subtracting the background, every array in the experiment is divided by a different factor to make the trimmed mean of the chips equal and, thus, make the arrays a similar average intensity [17, 28, 33, 34].

Unlike MAS 5.0, MBEI does not use a background correction. The MBEI algorithm normalizes the data by first choosing an array to serve as a basis for the normalization. One by one, each chip is normalized to the base-line array so that all the microarrays share a similar median. For each chip's normalization, the algorithm compares it with the baseline and finds a group of probes that show the least change, called the invariant set [30, 32]. Certain parameters ensure that

the chosen invariant set spans the range of intensities. Based on the invariant set of probes, a collection of piece-wise linear splines create a nonlinear scaling function. This means that for different ranges of intensities, the normalized chip will be scaled up or down in reference to the baseline. The underlying idea is that since the invariant probes are least likely to correspond to differentially expressed genes, they should serve as the basis for normalization. Different intensities may respond differently to noise and so a set of piece-wise linear splines can represent these nonlinear effects. A drawback to this algorithm is that the normalization depends on the baseline array chosen and different choices of baseline array can have different downstream results. [32, 35]

The algorithm used in conjunction with RMA, called quantile normalization [17], controls variability by forcing every array to use the same distribution for probe intensities. Before applying quantile normalization, RMA reduces every intensity by a background value [18]. This is not a part of quantile normalization *per se* but is part of the RMA algorithm. Quantile normalization begins by ordering the intensities of each chip from greatest to least. The average intensity at each rank is calculated and assigned to each probe at that rank, so the probe with the highest intensity on each chip now has the same value, the mean of all the highest intensities. Likewise, the probes with the second highest intenstiy on each chip are set to their mean, and so forth. After the normalization, the intensities found on each chip are identical. Quantile normalization, therefore, avoids the need to choose a base-line array [17].

1.1.5 The Expression Index

Once the microarrays have been normalized, there is still the problem of determining differential expression based on probe intensities. Algorithms like MAS 5.0, RMA, and MBEI address this problem by combining the probe intensities of a probeset into one value, called the expression index [29]. The expression index serves as a relative measure of the concentration of the probeset's target. It is not an exact concentration but is frequently used to estimate the fold changes of the transcript's concentration for comparison with RT-PCR results and other confirmation data [18].

For each probeset, MAS 5.0 calculates the expression index θ_j by first taking the difference between the log transformed perfect match PM_{ij} and mismatch probes MM_{ij} [33]. If the mismatch probe intensity is higher than the corresponding perfect match probe intensity, then the mismatch probe intensity is set to some small positive threshold, reducing its influence in the final calculation. MAS 5.0 then employs Tukey's biweight algorithm to scale differences by their distance from the median difference and sums them up. Tukey's biweight is a way of removing outliers and producing a robust measure of the difference in perfect match and mismatch probes for a probeset.

$$\log(PM_{ij} - MM_{ij}) = \log(\theta_j) \ + \epsilon_{ij} \ \text{, for probe i, probeset j, and error } \epsilon_{\text{ij}} \quad (1.1)$$

In contrast to MAS 5.0, MBEI estimates the binding affinities of each probe because this variation was found to be five times as great as the variation due to arrays [29]. Adjusting microarrays to similar intensities, therefore, does not remove the largest source of variation. MBEI fits a linear model that relates transcript concentration θ_j, probe binding affinity ϕ_{ij}, the difference in perfect match and

mismatch probe intensities $PM_{ij} - MM_{ij}$, and an error term ϵ_{ij}.

$$PM_{ij} - MM_{ij} = \theta_j \ \phi_{ij} + \epsilon_{ij} \ \text{, for probe i and probeset j} \tag{1.2}$$

MBEI uses a least squares fitting routine and an iterative process to remove probe, probeset, and array outliers. The calculated θ_j values serve as the expression indices in later calculations.

RMA uses the background corrected and quantile normalized data from the first step in a linear model similar to the one found in MBEI [18]. The model, however, uses log transformed perfect match intensities $\log(PM_{ij})$ and ignores the mismatch probes. This $\log(PM_{ij})$ term is decomposed into the sum of three terms: transcript concentration θ_j, each probe's binding affinity ϕ_{ij}, and random error ϵ_{ij}.

$$\log(PM_{ij}) = \theta_j + \phi_{ij} + \epsilon_{ij} \ \text{, for probe i and probeset j} \tag{1.3}$$

The model is fit by a median polish algorithm without the removal of outliers. As in MBEI, the θ_j term is the expression index [18].

1.1.6 Differential Gene Selection

With the expression indices calculated, all that remains is the final problem of finding the differentially expressed probesets. Methods vary from *ad hoc* procedures to applications of standard statistical tests to novel permutation based procedures [14]. The MAS 5.0, RMA, and MBEI algorithm suites do not have a preferred differential gene selection algorithm, and so this last step depends entirely on the analyst.

One popular method is the selection of any probeset with an expression index fold change (ratio of experiment to control) greater than 2 (upregulated) or less

than .5 (downregulated) [34, 36, 37]. This selection criteria has the benefit of simplicity but, unfortunately, has a bias for selecting lowly expressed genes [36]. For instance, a difference of 200 in expression indices has a larger impact on the fold change of a probeset with a control expression index of 10 as opposed to one of 1,000. This method also suffers from the lack of a measure of statistical significance [36]. Instead of setting a constant fold change, a related method called the unusual ratio method looks for the highest and lowest fold changes in an experiment. The major drawback to this method is that it always selects genes as being differentially expressed even if there are none [36].

Researchers have also applied statistical tests, such as the t-test, to expression indices [36, 37, 38]. While this carries a measure of statistical significance, two additional complications arise. First, traditional statistical tests for different means usually assume the data is normally distributed. Evidence suggests that the calculated expression indices, however, are not normal and so invalidates results that depend on this assumption [25, 37]. If normality is not assumed, as in the case with nonparametric tests, there is still an issue with multiple testing [37]. In the case of microarray analysis, tens of thousands of probesets need to be tested for statistically significant changes. If for each statistical test an error is expected 1% of the time, then in 25,000 tests one should expect 250 errors. Researchers have implemented adjustments such as the Bonferroni correction to reduce the number of such errors, but such cases tend to be too restrictive and prevent the discovery of any differentially expressed genes [36, 37, 38].

Unlike the traditional statistical tests which attempt to control the probability of making one mistake in all comparisons, algorithms such as Significance Analysis for Microarrays (SAM) [39] attempt to control the false discovery rate, the percent of selections that are wrong. In SAM a score based on a t-statistic is

calculated between experimental groups. For each score, SAM calculates the number of probesets that would have such a score by chance through randomly shuffling which chips belonged to which groups. In this way, SAM uses the specific expression indices from a microarray experiment to estimate the false discovery rate.

Another method called the Local-pooled-error test [40] also uses a modified test statistic for differential gene selection. It has options to control either the false discovery rate or the family wise error rate, as found in a standard t-test. Instead of using permutations to estimate significance, the variability from probesets with similar expression values are pooled to improve statistical detection.

1.1.7 Other Algorithms

There are many algorithms not covered in the above review [14, 23]. Some algorithms are similar to the above methods and mix different routines to produce an expression index. PQN, for instance, is a method that subtracts a background from the probe intensities, applies a log transformation, takes a trimmed mean that removes the bottom 40% and top 10% of intensities, and uses quantile normalization on the expression indices [41]. These steps make it a hybrid between RMA and MAS 5.0. Another algorithm subtracts a background term from perfect match probe intensities and adds a term designed to stabilize the variance of the probe-level data [42]. The final result is log transformed to obtain an expression index similar to RMA. Geller *et al.* present a different normalization method that uses biological replicates of one sample to calibrate the transformation [43]. They demonstrate that this method produces constant variance and symmetric errors

for all probes regardless of expression level. Other groups used Bayesian models to select differentially expressed genes [44, 45].

1.1.8 Comparisons

There has yet to be a conclusive study on which algorithm is the best. Previous studies often compare multiple methods for one particular step while holding others fixed [14]. These results can reveal which algorithm works best in conjunction with the fixed steps but does not reveal which group of algorithms work the best as a cohesive suite. Validation data sets are also not readily available. Affymetrix provides a Latin square data set on their website [46] which has 42 probesets spiked in at different concentrations over the course of 14 microarray experiments each done with three replicates. The problem with this data set is that only a tiny percentage, $< .005$, probesets are actually present on the microarray, and this does not represent a typical biological situation where half of the transcripts may be present in some concentration. Other validation data sets with spiked in genes suffer similar problems. Comparisons done on microarrays from real biological samples often compare some measure such as the variance within replicates or selected genes based on known biological features [13, 17, 47]. The problem is that not all the concentrations of transcripts on the microarrays are known and so it is difficult to get a complete picture of the number of truly changed genes.

Despite the problems in validation data sets and algorithm combinations, several papers have compared microarray analysis algorithms [17, 25, 26]. The same results appear consistent: RMA and MBEI outperform MAS 5.0 [20, 26]. This has been shown using spike-in data [24], biological data with RT-PCR

confirmations [20], and even analytically [48]. Because RMA and MBEI model the probe affinity's which account for much more variation than inter-array variability, they should be more accurate than MAS 5.0 [48]. Further comparisons between RMA and MBEI are inconclusive. In a study with two spike-in experiments, RMA did better than MBEI [24]. Yet, in a biological experiment confirmed by RT-PCR no method showed a real advantage [27]. Another biological experiment showed that MBEI had higher sensitivity and consistency than RMA [47]. In another comparison, RMA produced the most reproducible results yet genes identified solely by RMA were not confirmed via RT-PCR [20]. In an experiment using the detection of co-expressed operons, MBEI detected these much better than RMA but RMA had the edge in selecting differentially expressed operons [26]. These comparisons often use a test statistic or a fold change measure to detect change instead of a method like SAM which controls the false discovery rate. Since SAM has gained favor as the preferred method for assessing differential expression based on expression indices, it would be interesting to see how these algorithms work in conjunction with it.

1.2 Atrial Fibrillation

1.2.1 Atrial Fibrillation Introduction

Atrial fibrillation is the most common sustained cardiac arrhythmia in the United States: it affects an estimated 2.2 million people [49]. Its occurrence also increases with age, rising from .5% of the United States population in their 50s to 10% of people in their 80s [50]. This makes it an important problem as the population continues to age.

Atrial fibrillation is characterized by rapid and irregular activation of the atria [50]. While the average human heart beats around 60 times per minute at rest and 200 times per minute during exercise, patients with atrial fibrillation experience atrial depolarizations at speeds between 400 and 600 pulses per minute [50]. If this atrial firing rate occurred in the ventricles then the heart would be unable to pump blood effectively, and this would cause death [50]. Fortunately, the atrioventricular node acts as a filter and limits ventricular contractions to ~150 pulses per minute [50]. Although the rapid depolarizations are limited by the atrio-ventricular node, patients with short periods of atrial fibrillation suffer from a number of symptons including chest discomfort, light-headedness, and palpitations [50]. Long term atrial fibrillation can produce severe congestive heart failure in just a few weeks [50] and is responsible for up to 15% of strokes [51]. Furthermore, the loss of proper atrial contraction during atrial fibrillation can lead to a stagnancy of blood in the atria and, consequently, life-threatening blood clots called thromboemboli can form [50].

1.2.2 Atrial Fibrillation Mechanisms

Researchers have investigated the electrical mechanisms that sustain atrial fibrillation for over 100 years [50]. Since the early 20th century, they have attributed the rapid atrial depolarizations to three putative causes: 1. a group of spontaneously firing, ectopic cells, 2. a single re-entry circuit, and 3. a multiple re-entry circuit [50]. The single and multiple re-entry circuit mechanisms both maintain that a self-sustaining loop of depolarization develops and progressively increases the rate of depolarization. Assuming this mechanism, researchers have found that the likelihood of re-entry depends on the wavelength, which is the product of the

refractory period and conduction velocity [52]. The wavelength sets the minimum path length such that re-entry can sustain itself [52]. Shorter refractory periods or conduction velocities decrease the wavelength so that re-entry can happen in a smaller volume.

Of the potential mechanisms, the multiple re-entry circuit has been the dominant theory for the past 50 years [50]. Studies of atrial fibrillation in sheep hearts in the 1990s, however, suggested that a single circuit or ectopic focus [50], possibly acting via the pulmonary vein, may be to blame. Clinically, atrial fibrillation is associated with several risk factors such as advancing age, diabetes, hypertension, congestive heart failure, valve disease, myocardial infarction, and rheumatic and ischemic heart disease [53, 49]. Yet, up to 31% of atrial fibrillation patients have no underlying cardiovascular disease [53]. The different suspected mechanisms for sustaining atrial fibrillation have fostered different treatments including tissue ablation, anti-arrhythmic drugs, and surgically splitting the atria into electrically isolated areas [50]. These treatments each have some level of success but are complicated by other underlying symptons and side effects. Ultimately, it is not clear which mechanism is directly responsible, and it could be the case that the different mechanisms all occur and represent different pathological conditions.

1.2.3 Ionic Remodelling

While the mechanisms discussed above explain how the atria can depolarize so rapidly, they do not address how a re-entry circuit or group of spontaneously active cells develop. Since several cardiac disorders such as congestive heart failure and coronary artery disease predispose people to atrial fibrillation, these diseases

may create a substrate that promotes atrial fibrillation [50, 54]. Much work has been done exploring the machinery that promotes atrial fibrillation [50].

In clinical observations, doctors found that paroxysmal atrial fibrillation tended to advance to chronic atrial fibrillation [53]. Also, defibrillator treatment had a higher success rate when the atrial fibrillation had only existed for a short time. These two observations motivated a study in goats which sought to show that atrial fibrillation is progressive [53]. In the study, atrial fibrillation was induced in goats for short periods of time. With longer pacing times, the researchers found that the atrial fibrillation durations increased to a point that they became sustained [53]. Furthermore, the atrial effective refractory period decreased which would tend to decrease the wavelength and promote re-entry. After 1-week of conversion back to normal sinus rhythm, the effective refractory period returned to normal. Thus, this study demonstrated that atrial fibrillation alters the atrial electrophysiology in such a way that it promotes more atrial fibrillation, introducing the phrase, "atrial fibrillation begets atrial fibrillation [53]." This work implies that for sufficiently fast atrial tachycardia the electrophysiology remodelling would be the same final pathway for any initial mechanism of atrial fibrillation [50].

The electrophysiological changes promoted by atrial fibrillation are collectively described as electrical or ionic remodelling. Besides shortened refractory periods, investigations into this phenomenon revealed conduction abnormalities, altered calcium handling, changes in the sodium and potassium current, and differential expression of the channel proteins, known as connexins, which determine intercellular electrical communication [50, 55, 56]. In particular, calcium handling may be a major determinant of how atrial fibrillation promotes itself. Atrial fibrillation patients and dog models show decreases in L-type calcium channels [56, 57], and blocking the L-type calcium channels can attenuate the electrical remodeling

[55, 56]. Calcium homeostasis may be crucial because with each atrial action potential calcium enters the cell. Rapidly depolarizing atrial cells consequently experience an increased intracellular concentration of calcium [50, 55]. Because high intracellular levels of calcium can be lethal to cells, cells respond by downregulating the genetic expression of the L-type Ca 2+ channel [55, 56]. This decreases the cell's refractory period and in turn promotes atrial fibrillation [50].

1.2.4 Atrial Tachypacing Model

Early animal models for atrial fibrillation were only short term models where atrial fibrillation was induced and maintained through stimulation of the vagus nerve [58]. Based on the success of ventricular tachypacing models inducing ventricular fibrillation, Morillo *et al.* showed that chronic rapid atrial pacing in dogs could serve as a reproducible model of atrial fibrillation [58]. In their model, the hearts of mongrel dogs were paced at 400 pulses per minute for 6 weeks causing the ventricles to beat around 130 times per minute [58]. They found shortened atrial effective refractory periods, enlarged atria, increased size and number of mitochondria, dilation of the rough endoplasmic reticulum and nuclei, and disorganization of atrial fiber orientation without any extra deposition of connective tissue [58].

Dr. Stanley Nattel's lab adopted this atrial tachypacing model and showed that increasing the pacing time for the atria increases the duration of atrial fibrillation [59]. While the effective refractory period decreased to a minimum time at 1 week, the conduction velocity of the atria decreased more slowly reaching a minimum at 6 weeks, which was the length of the study [59]. Regional electrophysiological heterogeneity also increased with longer pacing times. Both conduction

abnormalities and regional heterogeneity were previously noted in human atrial fibrillation patients [59]. This work also established that the sham or control dogs were identical, independent of the time the pacing apparatus was left inside them. Additionally, as in both clinical findings and other studies [58], the left atrium showed more complex electrical activity that indicated different susceptibilities to remodelling within the atria. Other findings included reduced connexin 40 expression and decreases in the L-type Ca2+ channel [59].

1.2.5 Ventricular Tachypacing model

In the clinical setting, the prevalence of atrial fibrillation in congestive heart failure patients can be as high as 40% [53]. Previously, researchers found that tachypacing the right ventricle of dogs produced an animal model of congestive heart failure [60, 61]. This ventricular tachypacing model was later studied as a model for atrial fibrillation [54]. In that study, they paced the right ventricles of mongrel dogs at 240 pulses per minute for three weeks and then 220 pulses per minute for two weeks to limit mortality [54]. In comparison with the atrial tachypacing model, the ventricular tachypacing model produced atrial fibrillation but without changes in the effective refractory period [54, 62]. Although later work showed alteration of the sodium-calcium exchange current, the calcium current, and the slow potassium current [63, 64], the extent of electrical remodelling was reduced in ventricular tachypacing with no reductions in conduction velocity or heterogenity of refractory periods [54, 62]. This supports an alternative mechanism for atrial fibrillation.

Unlike the atrial tachypacing model, the ventricular tachypacing model showed extensive structural remodelling [54] in the atria and a transient immune

response [65]. The hallmarks of the structural remodelling are increased connective tissue deposition and an increased number of fat cells and fibroblasts [54]. This structure also permits a heterogeneity of conduction velocities, presumably due to the disruptions in cell communication caused by the increased extra-cellular matrix components [54]. A later study [65] supported this, showing that the likelihood of atrial fibrillation increases with the extent of fibrosis. Along with fibrosis, the ventricular tachypacing model also experienced an invasion of white blood cells, apoptosis, tissue edema, and cell death in as early as 24 hours of pacing [65, 66]. The immune response, however, was transient. It decreased in magnitude as the pacing time increased [65, 66] and returned to normal after 1 week of pacing [66].

1.2.6 Structural Remodelling

The fibrotic state produced by ventricular tachypacing remains even after pacing stops and the patient recovers from congestive heart failure [62]. Although there are no hemodynamic abnormalities or ionic remodelling present after recovery, persistent atrial fibrillation can still be induced [62, 64, 67]. Thus, structural remodelling may be solely responsible for atrial fibrillation in the ventricular tachypacing model [64]. This hypothesis is supported both by clinical evidence of increased fibrosis in atrial fibrillation patients and the advanced-age risk factor for atrial fibrillation [62].

The protein angiotensin II has been linked to the fibrotic structural remodelling in congestive heart failure [66]. The concentration of atrial angiotensin II increases within six hours of ventricular tachypacing and more than doubles by 24 hours [66]. Angiotensin II enhances phosphorylation of several proteins found

phosphorylated at 6 hours, including MAP kinase, JNK, ERK, and p38 [66]. Moreover, angiotensin II regulates the profibrotic cytokine transforming growth factor beta 1, whose activity is increased in ventricular tachypacing [62]. The increased activity of transforming growth factor beta 1 would explain the finding of increased genetic expression of extracellular matrix components like collagen and matrix metalloproteinases [62]. Interestingly, when an apoptosis blocker enalapril was given to ventricular tachypaced dogs, the angiotensin II concentration and phosphorylated ERK did not increase. Fibrosis reduced from a 9.8% increase in the congestive heart failure dogs to a 5.7% increase, which was still above the control group's fibrosis of .8%. It did not, however, prevent the increased cell death rates, the white blood cell invasion, and the enhanced phosphorylation of p38 and JNK [66]. This implies that structural remodelling involves angiotensin II/apoptosis dependent and independent pathways [66].

1.2.7 Previous Microarray Analysis

Many of the previous microarray studies of atrial fibrillation used tissue samples from human patients undergoing surgery for some other form of cardiac disease. In these cases, there was limited control over the experimental groups because many of the patients took cocktails of medication and even the control groups had a medical abnormality. It was difficult to separate the genetic changes reponsible for atrial fibrillation promotion from those associated with the underlying cardiac disease. Still, these studies identified several genes that may play a role in atrial fibrillation and suggested that atrial fibrillation may have its own transcriptional signature [68]. Due to the recent development of microarray technology

[1], there are few microarray studies of atrial fibrillation. Here, I briefly review a few studies done prior to the work presented in Chapter 2.

Barth *et al.* used Affymetrix microarrays to compare human patients with permanent atrial fibrillation to those with no history of it [69]. All patients in the study suffered from coronary artery disease or cardiac valve damage. When they looked at the right atrial tissue, they found that the atrial fibrillation patients had downregulation of calcium dependent signaling pathways like the calcineurin-NFATc signaling pathway [69]. They found upregulation of growth factors like platelet derived growth factor and extracellular matrix genes like collagen and matrix metalloproteinase 9 [69]. Most of the genes differentially expressed, however, were downregulated: 982 out of 1,434. They compared the atrial fibrillation right atrial samples to the atria and ventricles of the control group and found that they were more similar to the ventricular tissue [69]. Complementing this, they found the downregulation of atria-specific transcripts and less pronounced upregulation of ventricle-specific transcripts [69]. The downregulation of enzymes controlling fatty acid oxidation and upregulation of enzymes controlling glucose utilization also supported the claim that atrial fibrillation is associated with atrial dedifferentiation [69].

Lai *et al.* developed a porcine model of atrial fibrillation by rapidly pacing the right atrium at 600 pulses per minute [70]. After 4 weeks of pacing and 2 weeks of recovery, they obtained atrial tissue and analyzed it using cDNA microarrays. Compared to a control group, they found that 387 genes changed in the left atrium and 81 genes changed in the right atrium [70]. The ventricular isoform of myosin regulatory light chain (MLC-2V) showed the greatest change in upregulation in both atria [70]. Since MLC-2 plays a role in force development and sensitivity

to extracellular calcium, increased expression of MLC-2 may be a response to increased pressure from a rapid ventricular rate [70].

In another study, researchers used cDNA microarrays to compare atrial fibrillation patients undergoing maze therapy with sinus rhythm patients undergoing coronary artery bypass [71]. Between the groups they found 30 genes upregulated and 25 genes downregulated in atrial fibrillation patients [71]. Specifically, five of the most upregulated genes were related to reactive oxygen species: flavin containing monooxygenase 1, monoamine oxidase B, ubiquitin specific protease 8, tyrosinase-related protein 1, and tyrosine 3-monooxygenase. Two of the most downregulated genes were related to antioxidants: glutathione peroxidase 1 and heme oxygenase 2 [71]. The authors suggested that these results indicate that oxidative stress plays an important role in the pathology of atrial fibrillation [71].

Selecting from cardiac surgery patients, Ohki *et al.* used Affymetrix microarrays to compare the gene expression in right atria of patients with atrial fibrillation to those in normal sinus rhythm [72]. Atrial fibrillation patients had upregulation of vascular endothelial growth factor B, Rho C, an antioxidative enzyme called glutathione peroxidase, and inflammatory genes NF-IL6-beta and macrophage migration inhibitory factor. The authors proposed that RhoC, which regulates remodeling of the actin cytoskeleton during cell morphogenesis, may contribute to structural remodeling [72]. Genes for sarcoplasmic reticulum Ca2+ ATPase 2 and connexin 43 were downregulated in atrial fibrillation patients.

Finally, a study of valve replacement patients compared left atrial samples from people with chronic atrial fibrillation to those with no history of atrial fibrillation [73]. Of the 8,167 genes on the cDNA microarray, 31 were found decreased and 35 were found increased [73] in atrial fibrillation. Some of the most upregulated genes include a cell-cycle regulator cyclin dependent kinase inhibitor

1a and a signal transduction gene unc-5 homolog B [73]. The atrial fibrillation patients also had an increased bax/bcl-2 ratio, decreased angiotensin 1 receptor to angiotensin II receptor ratio, upregulation of p21, and down regulation of p27, all of which may be linked to apoptosis [73].

1.3 Models of Genetic Networks

1.3.1 Gene Expression: The *lac* Operon

Gene expression is a regulated, complex process that depends on the interaction of multiple molecules in time and space [74, 75]. At its simplest, RNA polymerase binds a section of DNA at the promotor site for a gene and initiates transcription. An additional layer of regulation adds DNA operator sites which can be bound by regulator molecules [74]. These regulators interact in combinatorial fashion to control the occupancy of the promotor. In this way the expression of one gene can affect the expression of several other genes. In eukaryotes, the complexity increases. Instead of one molecule, RNA polymerase, acting as the transcription machinery, over 50 proteins must be recruited and assembled [74]. The DNA itself is wrapped tightly around histones which can be modified through methylation to change the availability of promotor or operator sites [74]. There are also alternative splicing mechanisms that cut and paste mRNA transcripts into different sequences so that a genetic sequence can give rise to several different proteins. Additional regulation includes degradation, localization, and RNA silencing– not to mention any translational or post-translational control. All of these steps in controlling gene expression occur in parallel for the entire genome.

The *lac* operon in *Escherichia coli* is a paradigmatic example of a regulatory network that controls gene expression [74]. Since Jacob and Monod's initial characterization in the 1960s, it has been one of the most studied gene expression models [76]. In the *lac* operon, a single promotor dictates the expression of a series of three genes that regulate the catabolism of lactose [77, 78]. The first gene, lacZ, codes for β-galactosidase which cleaves lactose into glucose and galactose. The next gene, lacY, produces *lac* permease which resides in the membrane and transports lactose into the cell. The last gene, lacA, produces the enzyme thiogalactoside transacetylase that adds an acetyl group to the galactosides [77].

The *lac* operon is regulated by the LacI repressor molecule which, in the absence of lactose, can bind to one of three operator sites and thus inhibit expression. Each operator site has a different affinity for the repressor and so affords a different level of inhibition. Two repressors can bind to different operator sites and then interact to cooperatively inhibit expression. Along with the repressor binding sites, the *lac* operon has two activator binding sites that when bound by a catabolite activator protein complex, increase expression by $\sim$50 times [77, 79]. The repressor and activator binding sites can interfere with one another such that binding of one may prevent binding of the other [79]. To further complicate the system, the *lac* operon also has multiple promotor sites with different affinities for RNA polymerase [79].

In the presence of both glucose and lactose, the bacteria prefers to break down glucose for energy to the extent that the repressor remains bound and the *lac* operon is transcribed infrequently [77, 79]. In low levels of glucose, cyclic AMP accumulates and binds to catabolite activator protein. This binding enables the catabolite activator protein to bind to the activator binding sites and increase expression [77]. When the bacteria are then exposed to lactose, basal levels of

lac permease, lacY's product, transport lactose into the cell. Once in the cell, β-galactosidase, lacZ's product, converts lactose into 1,6-allolactose [77, 80]. This 1,6-allolactose product binds the *lac* repressor, preventing it from binding the operator sites. This releases the inhibition of the *lac* operon and produces more *lac* permease and β-galactosidase. Consequently, more lactose is imported and converted into 1,6-allolactose, repeating the cycle and establishing a positive feedback loop that maintains the expression of the operon [77]. Hence, the coordinated balance between repressor and activator molecules determines the expression of the *lac* operon.

1.3.2 Modeling Gene Networks

There has been extensive work on the mathematical modeling of genetic networks [81, 82]. Researchers have directed their attentions to specific biological systems like the *lac* operon [79, 80, 83, 84] and phage lambda [85, 86, 87]. By using detailed models and estimating parameters like equilibrium constants, they attempt to fit experimental data as well as make behavioral predictions. Failure of the models to fit the data can point to unknown biological mechanisms. For instance, phage lambda models could not reproduce the stability of the lysogeny state until the later discovery of an additional operator site [85, 74]. Besides investigating particular biological systems, researchers have also explored more theoretical models to deduce general and global properties like oscillations and stability [82, 88]. These models exhibit varying levels of scale and abstraction from boolean switching networks [88, 89] to directed graphs [82, 90] to systems of differential equations [82, 91, 92]. In this section, I will briefly review a couple examples of gene network models to give a sense of the some modeling approaches.

Mathematical models of specific biological systems enjoy the benefits of testing hypotheses on experimental data and the pitfalls of insufficient data to fit the models. Because models of such systems usually need to estimate kinetic rates, there is a tendency to focus on more understood experimental models. Consequently, the *lac* operon has been a popular subject of modeling [79, 80, 83, 84]. To show contrasting modeling approaches, I will briefly present two models of the *lac* operon that only examine the interactions between the repressor and its operator sites, ignoring the activator and its sites.

Yildirim *et al.* [80] present a model of five nonlinear differential equations that govern the dynamics of *lac* permease (P), β-galactosidase (B), lactose (L), operon mRNA (M), and 1,6-allolactose (A). In the model, there are three delays: one for transcription of the operon (τ_M) and two for the translation of *lac* permease (τ_P) and β-galactosidase (τ_B). The model has 24 parameters. Of them, 22 were estimated based on published data and 2 were fit to an experimental data set. The equations are shown below to give a sense of the model. Those variables and parameters which have not been defined represent binding rates, decay rates, production rates, and half saturation constants. For more information consult Yildirim *et al.* [80]. Although the model is too complex to permit a full stability analysis, numerical results show good concordance with three sets of experimental data including two time courses of β-galactosidase activity. They also determine that the system can have up to three steady states and demonstrates bistability. This bistability of the *lac* operon has also been experimentally validated using synthetic biology techniques [93].

$$\frac{dM}{dt} = \alpha_M \frac{1 + K_1(\exp^{-\mu\tau_M} A_{\tau_M})^n}{K + K_1(\exp^{-\mu\tau_M} A_{\tau_M})^n} + \Gamma_0 - \Upsilon_M M$$

$$\begin{aligned}
\frac{dB}{dt} &= \alpha_B \exp^{-\mu\tau_B} M_{\tau_B} - \Upsilon_B B \\
\frac{dA}{dt} &= \alpha_A B \frac{L}{K_L + L} - \beta_A B \frac{A}{K_A + A} - \Upsilon_A A \\
\frac{dL}{dt} &= \alpha_L P \frac{L_e}{K_{L_e} + L_e} - \beta_L P \frac{L}{K_{L_1} + L} - \beta_{L_2} B \frac{L}{K_{L_2} + L} - \Upsilon_L L \\
\frac{dP}{dt} &= \alpha_P \exp^{-\mu(\tau_P + \tau_B)} M_{\tau_P + \tau_B} - \Upsilon_P P
\end{aligned} \tag{1.4}$$

Instead of a differential equations model, Vilar *et al.* [84] use a statistical thermodynamics approach as used by Ackers *et al.* [86]. They divide the occupancy of the promotor into probabilistic states: free and repressed. In the free state, the operator sites are unbound by repressor and the operon is assumed to be transcribed. The repressed state is the opposite case, where at least one operator site is bound by repressor and transcription is prevented. The probability of each state is the product of its free energy ΔG and the number of ways it can exist. For example, if there are N repressors and one operator site, then the number of ways the state of a bound operator can exist is N, one for each repressor bound to that operator. Using this state-based description, Vilar *et al.* equate the repression level R_{O_m} with the probability an operator site is bound [84]. This expression is shown below for the purposes of illustration. Consult Vilar *et al.* [84] to identify parameters or see the derivation.

$$R_{O_m} = 1 + \frac{N \exp^{-(\Delta G_{O_m})} + N \exp^{-(\Delta G_{O_m} + \Delta G_l + \Delta G_{O_a})} + N(N-1) \exp^{-(\Delta G_{O_a} + \Delta G_{O_m})}}{1 + N \exp^{(-\Delta G_{O_a})}} \tag{1.5}$$

The probability the promotor is free is 1-R_{O_m} compared with $\frac{1+K_1 A^n}{K+K_1 A^n}$ as found in Yildirim *et al.* [80]. Using this model, Vilar *et al.* find that the cooperativity

from repressor interactions, when bound to operator sites, is analagous to increasing the number of repressors per cell. Stochastic simulations then show that the cooperativity yields a distribution of *lac* mRNA similar to that obtained with a higher association rate for the repressor, but significantly different from lowering the repressor's dissociation rate.

1.3.3 Allostery

Models of genetic networks often focus on the dynamics induced by gene or protein interactions. Yet, properties of the individual molecules such as allostery can also shape the dynamics [94]. Allostery is the property whereby proteins can assume at least two different conformations with different affinities for ligand binding [94, 95, 96]. It is frequently found in transcription factors, enzymes, and receptors [96]. Even though allosteric proteins can have ligand binding sites in different domains or subunits, binding of one ligand alters the protein's affinity for another [94, 96]. Based on 24 allosteric enzyme systems, Monod *et al.* [95] proposed a model (MWC) in which proteins exist in a tense or a relaxed conformation with different ligand affinities. The protein can switch between conformational states but there is no hybrid form where the protein is both tense and relaxed. Ligand binding reinforces a conformation by reducing the probability of a conformational switch. This system offers a way to control proteins, switching them from "on" states to "off" states via ligand binding.

The MWC allostery model is not uncontested. The sequential model [97] proposes that subunits change conformation one at a time which allows hybrid forms of the protein. Other models mix elements of the sequential and the MWC model, and some claim that allostery results from the average of many

proteins in different conformations [94]. While mathematical models often use the MWC model because of its simplicity and tractability, experimental results are still inconclusive about the correct model of allostery; it may be that the exact mechanism is protein dependent [94, 96].

1.3.4 Noise

While allostery is a property of individual molecules like transcription factors, stochasticity is a property of biological systems as a whole. In particular, the biochemical reactions that govern gene expression require interacting molecules to find one another in time and space and are subject to random collisions. This can lead to fluctuations in protein concentration around a mean, or noise. Advances in synthetic biology and single cell analysis have recently made it possible to measure noise in gene expression quantitatively [98]. In one experiment, Elowitz *et al.*[99] constructed a synthetic plasmid carrying two genes for two different fluorescent proteins. Bacterial cells transformed with the plasmids produced different amounts of the corresponding proteins, resulting in different fluorescence within the same cell. This proved that stochasticity due to processes like RNA polymerase binding has downstream effects in the amount of gene expression. Furthermore, the total flourescence of both proteins differed between cells. This variation was due to a number of factors including differences in the number of ribosomes, RNA polymerases, and degradation proteins. Other experiments have confirmed stochasticity in gene expression in both prokaryotes [100] and eukaryotes [101]. Stochasticity occurs at both transcriptional and translational levels, but not equally. Transcriptional noise contributes more to the total noise in eukaryotes while translational noise has the dominant role in prokaryotes

[98, 101]. The experiments discussed here show that stochastically generated mRNA transcripts can lead to bursts of protein production which in turn generate phenotypic variation. In this way, noise in gene expression can lead to population heterogeneity.

1.3.5 Managing Noise

While stochasticity in gene expression can create population diversity, it is generally thought to be detrimental to routine cellular functions because fluctuations in intracellular protein numbers can interfere with important cell signalling and cause inappropriate switching between steady states [98, 102]. Many researchers have modelled how cells mitigate the effects of noise and in some cases exploit it [103]. One way to reduce noise in gene networks is to use negative feedback [104, 105]. In negative feedback, a gene's product can inhibit its own production, either directly or indirectly. This enables a form of error correction such that genes stochastically expressed are quickly shut down. A type of negative feedback known as integral feedback found in the bacterial chemotaxis system permits robust adaptation [106, 107]. Another method is to use signaling cascades [106, 108, 109, 110] which act as low pass filters, removing high frequency noise by adding delays at each stage. In this manner, short term fluctuations will attenuate so that only persistent changes in protein concentration will dominate. Other mechanisms for controlling noise include redundant pathways, check points [111], and kinetic proofreading [106, 112].

1.4 Research Objectives

In the first part of my thesis, I will address the problem of how cells make decisions about the external environment based on noisy measurements. I will use differential equation models of transcriptional regulation to test whether gene networks can perform Bayesian inference, such that the state of the promotor at equilibrium mirrors the probability of being in a particular external environment. I will then use stochastic simulations to see if this inference can also work in fluctuating environments. I hope to show that viewing gene networks as Bayesian inference modules can explain the results of previous experimental data.

Next, I will analyze microarray data from two canine models of atrial fibrillation, atrial and ventricular tachpacying, observed at early and late time points. I will investigate whether these models have unique transcriptional profiles or share common mechanims. Using different techniques, I will try to link the differentially expressed genes to the observed pathological differences in the two models. I will explore how gene expression changes over time in each model. With higher level analysis looking at pathways and protein interaction, I hope to identify mechanisms or genes responsible for the pathology of atrial fibrillation that merit further study as therapeutic targets.

Finally, I plan to examine microarray analysis algorithms. By returning to the raw data, I hope to find properties of the probes that can be exploited in a new algorithm. I will also generalize the problem of selecting differentially expressed genes in the expectation that it will provide a better understanding of how to approach this problem. I will then develop the techniques and apply them to validation and experimental data to assess performance.

CHAPTER 2
Gene Networks as Inference Modules

This chapter investigates the design principles of simple genetic networks. We examine if a transcriptional regulatory network can act as a Bayesian classifier and infer the state of the environment based on fluctuating concentrations of intracellular molecules. We compare whether repressors or activators are better at processing such noisy information and which parameters are the most sensitive in this task. Finally, we determine if viewing genetic networks as inference modules can resolve issues presented by previous experimental data.

2.1 Abstract

Cells must respond to environmental changes to remain viable, yet the information they receive is often noisy. Through a biochemical implementation of Bayes's rule, we show that genetic networks can act as inference modules, inferring from intracellular conditions the likely state of the extracellular environment and regulating gene expression appropriately. By considering a two state environment, either poor or rich in nutrients, we show that promoter occupancy is proportional to the (posterior) probability of the high nutrient state given current intracellular information. We demonstrate that single gene networks inferring and responding to a high environmental state infer best when negatively controlled, and those inferring and responding to a low environmental state infer best when positively controlled. Our interpretation is supported by experimental data from the *lac* operon, and should provide a basis both for understanding more complex cellular decision making and for the design of synthetic inference circuits.

2.2 Introduction

For cells to interact with their environment, the DNA and regulatory machinery, which are intracellular, require information from the cell surface. This information is conveyed through gene and protein networks and is transferred via biochemical reactions that are potentially significantly stochastic [99, 101, 113, 114]. Stochastic fluctuations will undermine both signal detection and transduction. Cells are therefore confronted with the task of predicting the state of the extracellular environment from noisy and potentially unreliable intracellular signals. For example, a bacterium must decide from intracellular levels of a nutrient whether or

not the nutrient is sufficiently abundant extracellularly to express the appropriate catabolic enzymes. Similarly, a smooth muscle cell must decide from concentrations of second messengers whether or not extracellular hormone levels are high enough to warrant contracting.

Here we consider if, and how, it is possible for biochemical networks to correctly infer properties of the extracellular environment based on noisy, intracellular signals. Suppose that the cell should respond under high concentrations of an extracellular molecule. Suppose further that the concentration of an intracellular signaling molecule is related to the concentration of the extracellular molecule through a signal transduction mechanism. A simple inference network could establish a concentration threshold for the intracellular molecule. Only if the molecule is above threshold is the extracellular concentration judged to be high enough for a cellular response. This network performs poorly, however, in fluctuating extra- and intracellular environments. First, fluctuations lead to input molecules crossing threshold even when the state of the environment is unchanged. Second, a threshold scheme cannot specify the degree of certainty in the inference which may be important for the ultimate response. For example, a bacterium may express a catabolic operon once the degree of certainty in high extracellular levels of a particular nutrient reaches 40%, but it may only shut down other catabolic operons once the degree of certainty is larger, 80% say.

The method of Bayesian inference both accounts for fluctuations and gives a degree of uncertainty in predictions [115]. We postulate that the cellular regulatory machinery may have evolved to perform Bayesian inference on some intracellular inputs. Typically, a cellular decision has two levels: first, predicting the state of the environment; second, choosing the appropriate response. At this second level, the expected costs must be compared with expected benefits [116]. Although

Bayesian theory can handle both problems, we focus here on the first: classification of the local environment.

As an example, consider a bacterium with a nutrient scavenging operon that encodes enzymes to import and catabolize a sugar (Fig. 2–1A and B). Suppose the environment can be in one of two states: a high or a low sugar state — for example, the high and low lactose environments of the small intestine [117]. The intracellular concentration of the sugar depends on the extracellular state, though in a stochastic fashion. To optimize growth, the bacterium must predict the extracellular state from intracellular sugar because expressing the operon involves a significant metabolic cost [116, 118]. Let S be the intracellular sugar level at a particular time. We denote the probability (i.e. the fraction of time) that there are S intracellular sugar molecules given that the environment is in the low sugar state as $P(S|\text{low})$. Similarly, we denote the probability that there is S intracellular sugar molecules given that the environment is in the high sugar state by $P(S|\text{high})$. If fluctuations are negligible, these two distributions will be sharply peaked functions of S, and they will be broader as fluctuations become significant.

The bacterium must determine the probability that its extracellular environment is in a high sugar state based on levels of intracellular sugar. This probability is denoted $P(\text{high}|S)$. A Bayesian approach assumes that some information about the long term probability of environmental states is known. This information could be simply that the environment is expected to be in one of two states, either a low or a high sugar state, and that each state is *a priori* equally likely. In one particular environment (for example, the soil), though, a low sugar state may occur more often on the long term. The *a priori* probability for this state will then be higher. Such *a priori*, or prior, probabilities are denoted $P(\text{high})$ and $P(\text{low})$. Once sugar enters the cell, the *a priori* probabilities are updated based on the levels of sugar

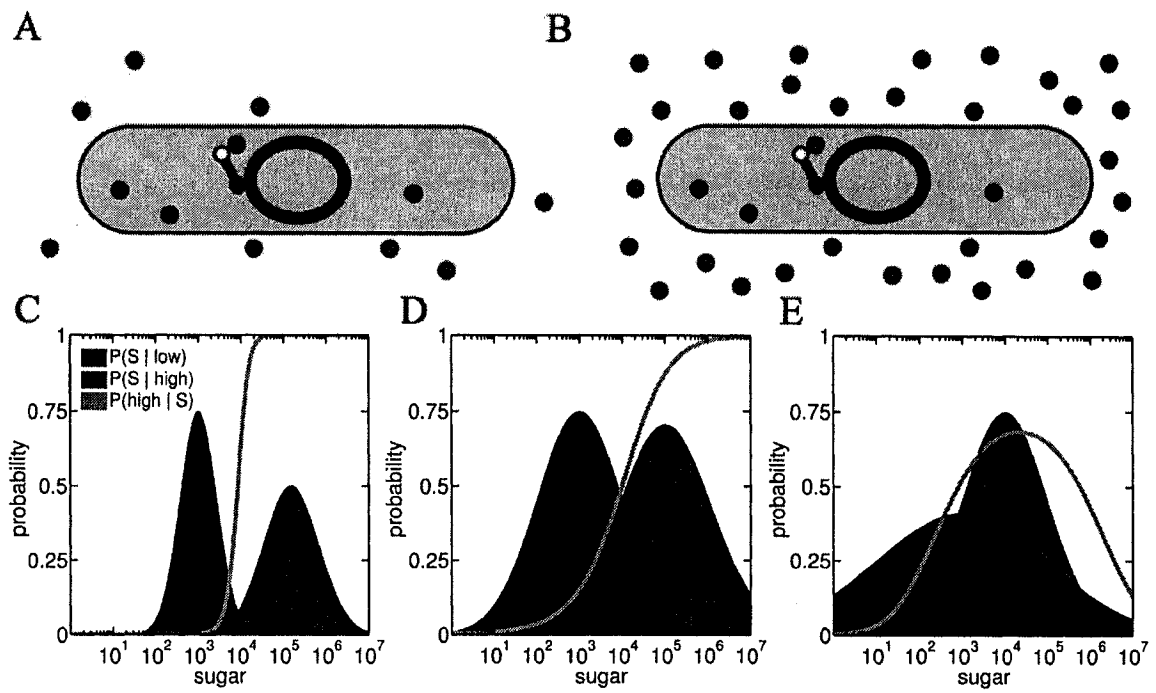


Figure 2–1: A two-state classifier problem and its Bayesian solution — the posterior probability. A cell must infer from intracellular concentrations of a nutrient or signaling molecule (green circles) whether the molecule is in high or low concentrations in the extracellular environment. **A** and **B** Fluctuations in the environment and in molecule detection and transport can lead to similar intracellular concentrations of the molecule for different extracellular conditions. The cellular decision-making machinery, shown as a simple genetic network, must decide from intracellular information the probable state of the extracellular environment. **C** Two distributions for intracellular numbers of a sugar molecule: the low sugar state is in blue, the high sugar state in red. For an intracellular sugar level S, the green curve is the posterior (predicted) probability that the extracellular state is the high sugar state, $P(\text{high}|S)$. **D** For two intracellular distributions that overlap substantially, the posterior probability for the high sugar state transitions gradually from low to high values. **E** The posterior probability, $P(\text{high}|S)$, need not be monotonic. The low sugar state is more probable at both low and high intracellular sugar, and $P(\text{high}|S)$ goes through a maximum.

detected. The more intracellular sugar, the larger the predicted probability of the environment being in the high sugar state (and the smaller the corresponding probability of the low sugar state). This *a posteriori* probability of the high state is $P(\text{high}|S)$. It is referred to as the posterior (predicted) probability of the high state given intracellular sugar S.

Bayes's rule states explicitly how the prior probabilities are correctly updated to their posterior values for the levels of sugar detected [119] (see Materials and Methods):

$$P(\text{high}|S) = \frac{P(S|\text{high})P(\text{high})}{P(S|\text{low})P(\text{low}) + P(S|\text{high})P(\text{high})} \tag{2.1}$$

Intuitively, the more likely a particular intracellular S is in the high extracellular state compared to the low extracellular state (the greater $P(S|\text{high})$ is compared to $P(S|\text{low})$), the higher the posterior probability of a high state environment. For simplicity, we will assume that the environment is *a priori* equally likely to be in either state: $P(\text{high}) = P(\text{low}) = 1/2$. The prior probabilities then play no mathematical role in Eq. 2.1.

Often the posterior distribution, $P(\text{high}|S)$, is a sigmoidal curve. Fig. 2–1C shows two distributions for numbers of sugar molecules: a distribution for a low extracellular sugar state (in blue) and a distribution for a high extracellular sugar state (in red). The corresponding posterior probability curve is shown in green. If the intracellular sugar level, S, is low, there is a high predicted probability that the extracellular state is low, with the converse holding for high intracellular sugar levels. In an intermediate range of S, lying in the overlap between the two state distributions, $P(\text{high}|S)$ switches from low probability to high probability. When fluctuations are more significant and the overlap between the two distributions is greater, the transition is more gradual (Fig. 2–1D). The posterior probability need

not always be sigmoidal: Fig. 2–1E shows a long tailed low state distribution that results in a non-monotonic posterior curve.

We will argue that a single gene can make probabilistic inferences about extracellular states through a biochemical implementation of Bayes's rule. By tuning the kinetic rates of the system, the promoter efficacy — the fraction of time the promoter is capable of initiating transcription — can match the posterior probability of high extracellular sugar. Consider a negatively controlled operon. We view the repressors controlling the gene as detectors that monitor intracellular sugar levels. Repressors thermally flip back and forth between two allosteric forms [95]: one DNA binding and the other non-DNA binding. As each repressor diffuses in the cytosol, it samples intracellular sugar. At low sugar levels, the DNA binding form of the repressor is stable, and the operon is not expressed. At high sugar levels, the non-DNA binding form is stable, leading to expression. Repressor binding sites on the promoter 'read' the allosteric form of cytosolic repressors and control transcription. Promoter efficacy is therefore a readout of the number of non-DNA binding repressors, which in turn are a readout of sugar levels.

2.3 Cis-regulatory Regions as Inference Modules

We tested the ability of different regulatory mechanisms to classify a two state environment. We considered 18 different networks (Fig. 2–2A-C): regulation can be positive or negative, transcription factor can allosterically bind either 1, 2, or 4 sugar molecules, and promoters can be one of three different types. Network input is the number of sugar molecules, which range from zero to approximately 2000 times the number of transcription factors. Network output is promoter efficacy

(i.e. promoter bound by an activator for a positive control and free of repressor for negative control). Rather than specialize to particular sugar distributions for the high and the low states, we generated 50 different pairs of lognormal distributions for S. Each pair corresponded to a different inference problem and had a different, but always sigmoidal, posterior probability. We fit the kinetic rates of each network to minimize the squared error between promoter efficacy and $P(\text{high}|S)$ as a function of S for each of the 50 posteriors (see Materials and Methods). A network that fits this collection of posterior curves well has a network architecture able to solve a variety of (two state) inference problems; it is an inference module.

Networks with higher cooperativity, either through the ability of transcription factor to allosterically bind sugar or through cooperative binding of transcription factors to DNA, perform best (Fig. 2–2D and E). A genetic inference system with low cooperativity is unable to generate a promoter efficacy curve that switches sharply with S [95]. These models thus perform poorly (higher residual in fits) on those inference problems with distinct sugar distributions and therefore strongly sigmoidal posterior probabilities (compare the posterior probabilities for Fig. 2–1C and D).

Less intuitively, negatively controlled inference systems perform significantly better than positively controlled systems (Fig. 2–2F). Positively controlled systems are less able to exploit cooperativity. Activators should bind DNA as sugar levels rise. Consequently, $K_b \gg K_n$ in Fig. 2–2A. For low sugar, the posterior probability is close to zero (Fig. 2–1C and D), and no activators at all should bind DNA. Therefore K_b must be small, and the more activators present, the smaller K_b must be. As $K_b \gg K_n$, both K_b and K_n are small: there is weak sugar binding, and cooperative binding only occurs at high sugar levels. Contrarily, in a negatively controlled system, $K_n \gg K_b$, so that sugar lifts repressor off DNA. For low sugar,

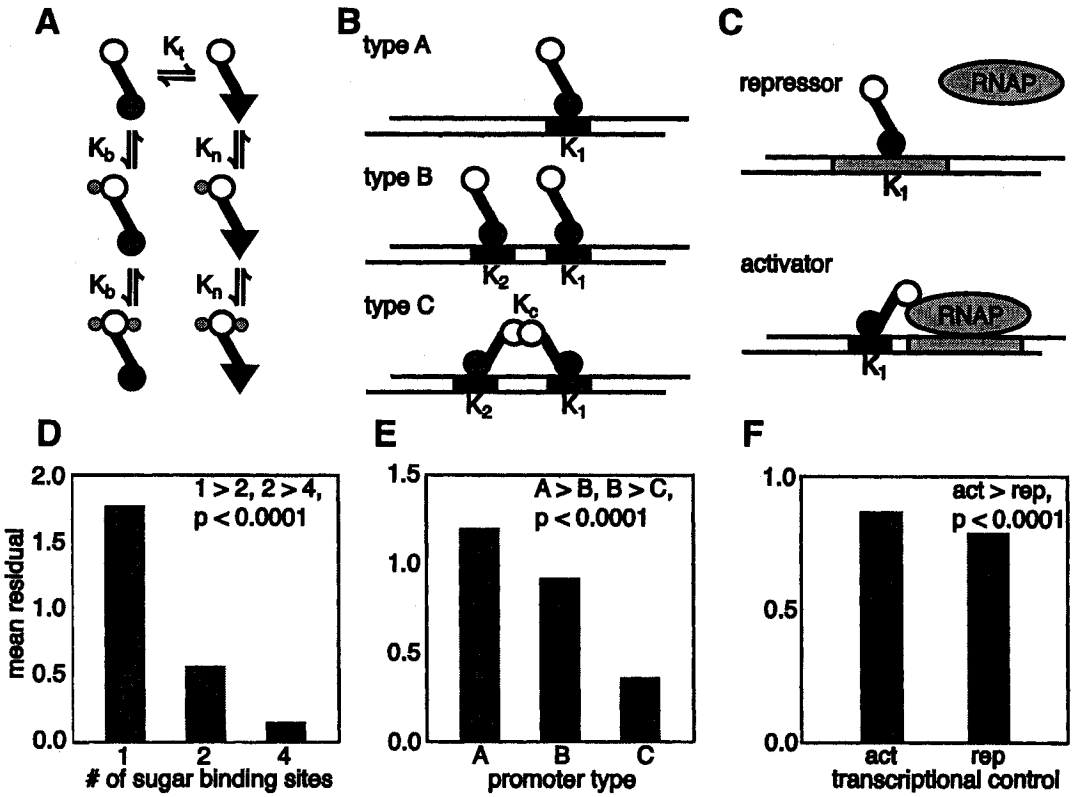


Figure 2–2: A comparison of different regulatory mechanisms for solving the two state discrimination problem; highly cooperative, negatively controlled genetic networks perform the most accurate inference. **A** The Monod-Changeux-Wyman model of an allosteric transcription factor. Association constants are denoted by K's. The protein flips between DNA binding (red circles) and non-DNA binding forms (blue triangles). If $K_b \gg K_n$, sugar stabilizes the DNA binding state. Conversely, if $K_n \gg K_b$, the non-DNA binding state is stabilized. Two sugar binding sites are shown, but we also test models with 1 and 4 binding sites. **B** We consider three different promoters: type A, one active operator site; type B, two active operator sites, but with no cooperative binding between transcription factors; and type C, one active and one inactive operator with cooperative transcription factor binding. **C** Transcription can by regulated either negatively, via repressors that obstruct RNA polymerase (RNAP) binding, or positively, via activators that help stabilize RNAP binding. The RNAP binding site (sigma site) is shown in gray, operators in red. **D** Mean residuals (a high residual implies a poor fit) from fits to 50 different posterior probabilities for the models grouped by the numbers of sugars bound by transcription factor. Models with 4 transcription binding sites perform the best inference (p value for one model type consistently performing better than the other is given in the inset — see Appendix). **E** Mean residuals for models grouped by promoter type. Cooperative promoters perform best (type C). **F** Mean residuals for models grouped by their mode of transcriptional control. Repressors perform better than activators (for over 70% of the fits, corresponding to a p value substantially smaller than 10^{-4}).

just one repressor must bind DNA to maintain a low promoter efficacy. More repressors allow K_b to be smaller giving greater, not less, flexibility in K_n. Altering K_t, the equilibrium between the DNA and non-DNA binding forms in the absence of sugar, can partly offset the inherent frustration in the activator system, but not completely (Fig. 2–2F). Therefore, negatively controlled promoters are best able to tune promoter efficacy to track $P(\text{high}|S)$.

While negatively controlled systems can better match their promoter efficacy to $P(\text{high}|S)$ than positively controlled systems, the opposite holds for matching $P(\text{low}|S)$. This posterior probability satisfies $P(\text{low}|S) = 1 - P(\text{high}|S)$ and so has the opposite behavior to $P(\text{high}|S)$. The argument given above is reversed. Thus, for systems that respond to a low state of the environment, positive control gives the best inference.

Fig. 2–2 demonstrates that model genetic networks can perform inference, with equilibrium promoter efficacy tracking posterior probability; Fig. 2–3 shows that inference can occur in real time in noisy environments. For the two sugar distributions in Fig. 2–1C, we chose the activator and repressor networks that best fit the posterior probability of the high sugar state. We performed a stochastic simulation of each of these networks using the best fit parameters, and let the environment change from a low to a high and back to a low sugar state. In each state, we sampled from the appropriate sugar distribution, mimicking intracellular fluctuations, and producing a time series of intracellular sugar (Fig. 2–3A). For each sugar level, there is a different posterior probability of the high extracellular sugar state (Fig. 2–1C). This instantaneous posterior probability is shown in Fig. 2–3B. Most often, $P(\text{high}|S)$ is very low (near zero) or very high (near one). It should be compared with the response of each network, measured by their promoter efficacies (Fig. 2–3C and D). Both the promoter efficacy of the repressor

network (Fig. 2–3C) and of the activator network (Fig. 2–3D) closely follow the instantaneous posterior probability, although the activator network underestimates the probability of the high sugar state. A quantitative measure of the goodness of fit of each promoter efficacy to $P(\text{high}|S)$ shows that repressor performs more than twice as well as activator (see Appendix).

2.4 Inference in the *lac* Operon

Viewing networks as inference modules gives new interpretations of *in vivo* behavior. For example, Setty *et al.* measured the transcription rate of the *lac* operon in *Escherichia coli* as a function of two inputs: isopropyl β-D-thiogalactoside (IPTG), an analogue of lactose, and cAMP [120]. Traditionally, transcription of the *lac* operon is described as being 'on' in the presence of sufficient cAMP and sufficient lactose, i.e., its *cis*-regulatory region performs a logical AND on the two inputs [121]. Setty *et al.* found more complex behavior: with enough IPTG, there is significant transcription at low cAMP, and transcription increases smoothly, rather than in a switch-like fashion, as cAMP increases (Fig. 2–4A).

The shape of this surface can be explained if the *lac* operon has evolved to solve a two state inference problem. The high state corresponds to a state where the *lac* operon should be expressed — an extracellular environment rich in lactose and poor in glucose, resulting in both high intracellular lactose and cAMP (cAMP concentrations are inversely proportional to glucose levels [122]). The low state, where the *lac* operon should not be repressed, corresponds to an extracellular environment poor in lactose and rich in glucose. We interpret S in Eq. 2.1 as the set of two variables: intracellular IPTG and cAMP concentrations (see Materials

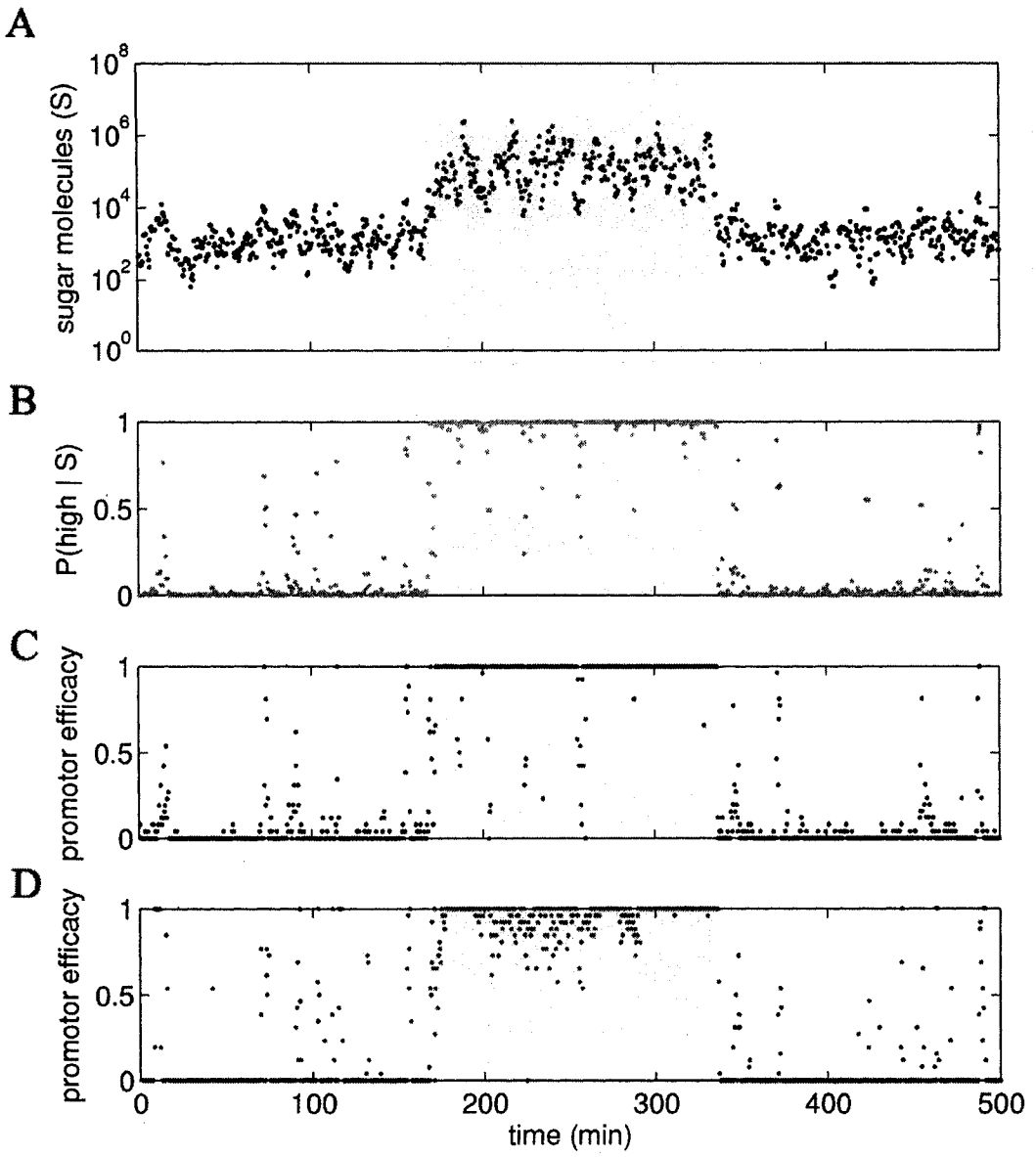


Figure 2–3: Two-state inference by simulated genetic networks. **A** A time series of intracellular sugar molecules as the extracellular environment moves from a low to a high (shaded region) and back to a low sugar state. Histograms of the intracellular sugar distributions are shown in Fig. 2–1C. Sugar was sampled every 25 seconds. In the low state, mean sugar numbers are $\sim 10^3$; in the high state, $\sim 10^5$. **B** The instantaneous posterior probability of the high sugar state, $P(\text{high}|S)$, for the particular sugar level existing at the current time point. Posterior probability points come from the green curve in Fig. 2–1C. **C** The average promoter efficacy for the best repressor network of Fig. 2–2 with 4 sugar binding sites and promoter type C. The actual promoter efficacy is either zero (promoter bound by repressor) or one (unbound promoter). An average over the 25 second period chosen to sample the sugar is shown. **D** The average promoter efficacy for the best activator network of Fig. 2–2 with 4 sugar binding sites and promoter type C.

and Methods). Assuming bivariate lognormal distributions for IPTG and cAMP in each state, we fit the parameters of the distributions so that the posterior probability, $P(\text{high}|S)$, matches the data of Fig. 2–4A (Fig. 2–4B). Two lognormal distributions that generate this posterior are shown in Fig. 2–4C. (Note that the axes represent measured extracellular levels, which are assumed to be proportional to intracellular levels [120].) The *lac* transcription rate is well explained by a two state model in which mean intracellular levels of IPTG are approximately 3 times higher in the high state than in the low and cAMP levels are 10 times higher.

2.5 Discussion

We have argued that a single gene through allosteric control and its *cis*-regulatory region can statistically infer the state of the extracellular environment from intracellular inputs. *Cis*-regulatory regions are often considered to perform logical operations on their input, allowing gene expression only under a particular combination of inputs [123, 124]. Such a view has been especially successful in understanding development [125], where gene expression occurs in an ordered manner. Cell behavior need not, however, follow a pre-determined pattern, and in these cases a cell that infers the state of its environment may have an evolutionary advantage. A genetic network, or more generally a biochemical network, that performs inference allows the cell to optimally interpret fluctuating inputs. Expression of the *lac* operon is a possible example, but inference is also likely to occur in signal transduction networks. Although we have emphasized the sigmoidal character of the posterior probability, networks that perform Bayesian inference need not have a sigmoidal output. Fig. 2–1E shows two sugar distributions that

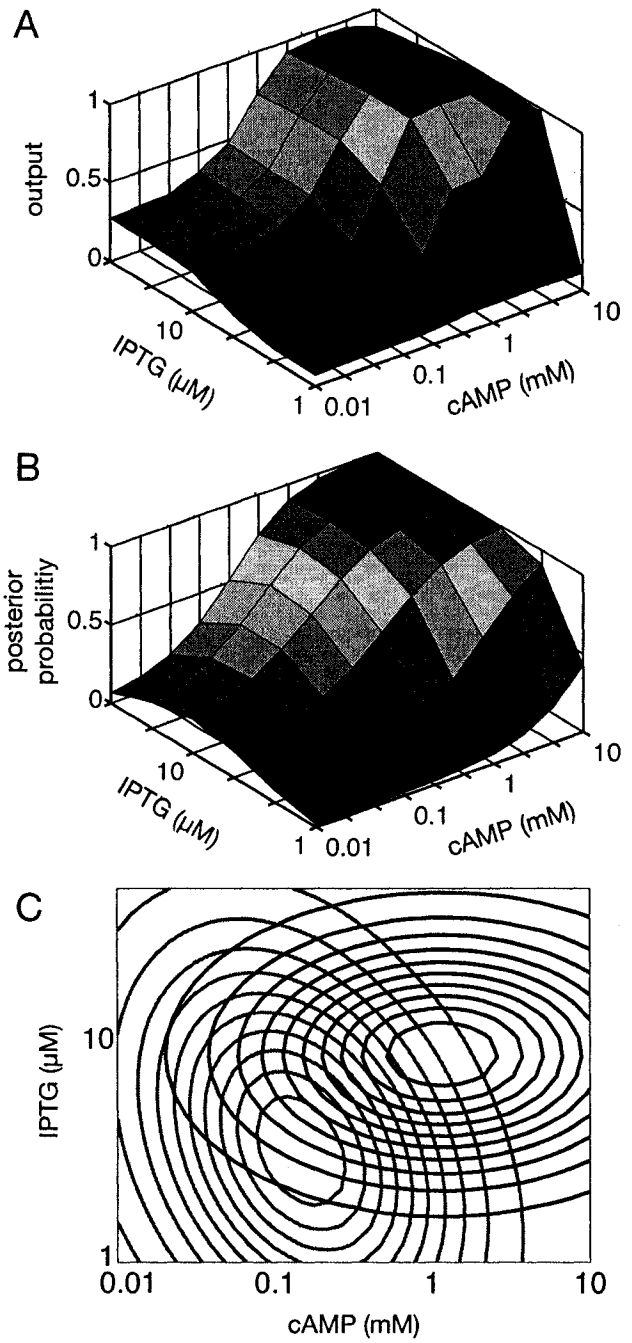


Figure 2–4: Inference by the *lac* operon in *E. coli.* **A** Observed transcriptional output as a function of extracellular concentrations of IPTG and cAMP (both log scaled), normalized to range from zero to one (data from [120]). **B** Posterior probability, fit to the data in **A**, that the environment is in a high state given the concentrations of IPTG and cAMP. **C** A possible two state model for *E. coli*'s view of its extracellular environment. The low state is in red (peak at $\sim 3\mu$M IPTG and 0.2 mM cAMP), the high state is in black (peak at 8μM IPTG and 1.2 mM cAMP). Both states are described by bivariate lognormal distributions.

produce a biphasic posterior probability. Such behavior has been reported, for example in the *E. coli gal* operon [126], and is hard to justify within a logic gate description.

We predict that a positively controlled genetic inference module is more likely to infer the probability of the environment being in a low state and that a negatively controlled system is more likely to infer the probability of the environment being in a high state. For example, the cAMP receptor protein in *E. coli* is an activator and promotes high promoter efficacy of the *lac* operon when glucose levels are low; LacI is a repressor and promotes high promoter efficacy when lactose levels are high [121]. This bias is expected to be stronger for networks with less cooperativity.

Although we have focused on a single estimate of the probability of the extracellular state, cells might be expected to perform long term integration of noisy signals. Such integration could occur by changing the prior probabilities of the high and low states. For example, an *E. coli* previously exposed to lactose has a higher concentration of lactose permease in its cell membrane than one not exposed [127]. This greater permease concentration may reflect an increase in the prior probability of the high extracellular lactose state, i.e., $P(\text{high}) > P(\text{low})$. Eq. 2.1 then predicts a sigmoidal response that favors the high state: the posterior probability curve is shifted towards lower sugar levels. This change mimics the change expected in promoter efficacy of the *lac* operon: higher permease concentrations lead to gene expression (higher promoter efficacy) at lower extracellular lactose levels because lactose more efficiently enters the cell.

In our framework, the output of different networks are distinct functions of their input because each network is solving a different inference problem. For example, if the intracellular distributions of the two extracellular states strongly

overlap, a repressor may have a high allosteric constant (K_t in Fig. 2–2A) to give a more sigmoidal promoter efficacy curve, reflecting the steep posterior probability. The promoter efficacy curve is most sensitive, however, to the inducer binding affinity (K_n for repressors and K_b for activators). Its sensitivity is over three times higher than the next most sensitive parameter (K_t) — see Appendix. If the extracellular environment substantially changes, leading to a new inference problem, the most efficient way to evolve to the new posterior probability is to modify the sugar binding affinity. This modification has the benefit of preserving the connectivities of pre-existing genetic networks.

Cellular inference need not follow the simple two state classifier model proposed here. Multi-state classifiers and real time averaging methods are more appropriate for some problems. Nevertheless, given the prevalence of sigmoidally responding biochemical networks [128], the two state classifier, whose solution is often a sigmoidal posterior probability, may be an essential component of many inference and decision-making networks in cells. Interpreting biochemical networks as inference modules may be an important step for both unraveling cellular behavior and designing selective, synthetic gene circuits.

2.6 Materials and Methods

2.6.1 Modeling Genetic Networks

We use the Monod-Wyman-Changeux model [95] to describe allosteric transcription factors. We assume that both the total amount of sugar and the total amount of transcription factors are conserved. Given these values, we numerically solve for the amount of free sugar and the total amount of transcription factor

in the DNA binding state, irrespective of the number of sugars each individual transcription factor has bound (see Appendix).

To calculate promoter efficacies, we follow a statistical mechanics approach [129] to describe the equilibrium occupancies of the different states of the promoters of Fig. 2–2B (see Appendix).

2.6.2 Comparison of the Models

To test the ability of the models to implement a Bayesian classifier, we fit each model to the posterior probabilities for 50 different two state classification problems. For each problem, we generated two sugar distributions corresponding to a low and a high sugar state. From these distributions, we calculated the posterior probability of being in the high state for each concentration of sugar S:

$$P(\text{high}|S) = \frac{P(S|\text{high})P(\text{high})}{P(S)} \tag{2.2}$$

We can rewrite the expression for the probability of a sugar concentration as:

$$\begin{aligned} P(S) &= \sum_{\text{states}} P(S|\text{state})P(\text{state}) \\ &= P(S|\text{high})P(\text{high}) + P(S|\text{low})P(\text{low}) \end{aligned} \tag{2.3}$$

to derive Eq. 2.1. For simplicity, we assume equal priors; allowing unequal prior probabilities for the two states does not change our results.

We considered two state classification problems generated by Poisson, normal, and lognormal distributions of sugar. The results of Fig. 2–2D-F are for lognormal distributions, but are qualitatively the same independent of the distribution type

chosen. The probability $P(S|\text{state})$ in Eq. 2.1 is therefore

$$P(S|\text{state}_i) = \frac{e^{\frac{-(\ln S-\mu_i)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i S} \tag{2.4}$$

where $i = 1$ for the low state and $i = 2$ for the high state. Each state has a different μ_i and σ_i, which define the mean and standard deviation in log space of the distribution. Fifty posterior probability curves that best gave a range of different inference problems were chosen (see Appendix).

We used a least square fit to score how well a model matches the posterior probability of the high state. To fit we use an interior-reflective Newton method (`lsqnonlin` in Matlab, The Mathworks, Massachusetts). Each posterior probability curve generated has 100 points (evenly spaced in log space), and we fit all 18 models to each curve 500 times with different initial conditions, for a total of 450,000 fits.

The p values for the residual comparisons were computed using a Wilcoxon two-sided signed rank test (`signrank` in Matlab, The Mathworks, Massachusetts). For each fit, we calculated the difference in the residual for a particular pair of models. The null hypothesis was that these differences came from a distribution with median zero.

2.6.3 Stochastic Simulation

We simulated both a repressor and an activator model. We chose a posterior probability from the 50 used in the fitting (the posterior of Fig. 2–1C) and the repressor and activator model that fit it best (parameters are given in the Appendix). The selected repressor and activator models both have four sugar binding sites and promoter type C in Fig. 2–2B. To generate a relatively smooth

time series of sugar levels, we used a Markov chain Monte Carlo method [115] to produce fluctuating, dependent samples of sugar from the appropriate distribution in Fig. 2–1C. For each sugar sample, the cytosolic sugar levels are changed to the new sampled value. A stochastic simulation of the genetic network is then run for a fixed time interval of 25 seconds using the Gillespie algorithm [130] (results for different time intervals are given in the Appendix). A new sugar sample is then taken and the simulation of the genetic network run again. The average value of the promoter efficacy during each simulation run is shown in Fig. 2–3C and D.

2.6.4 Fitting a Posterior Probability to an Operon

We fit the data of Fig. 2–4A to Eq. 2.1 where each state is characterized by two variables: s_1 corresponding to the logarithm of the IPTG concentration and s_2 corresponding to the logarithm of the cAMP concentration. $P(S|\text{high})$ is then a bivariate normal distribution:

$$P(S|\text{high}) \sim \frac{1}{\sqrt{\det(\sigma)}} \exp\left(-\frac{1}{2}\sum_{i,j=1}^{2}(s_i-\mu_i)\sigma_{ij}^{-1}(s_j-\mu_j)\right) \tag{2.5}$$

with μ_1 the mean of s_1, μ_2 the mean of s_2, and σ the covariance matrix of s_1 and s_2, all for the high state. A similar set of parameters is needed to describe the low state. The problem of fitting Eq. 2.1 to a given posterior probability surface is degenerate: different sets of parameters can result in the same posterior surface (see Appendix). However, we can identify a unique posterior probability surface that best fits the *lac* operon data (Fig. 2–4B) along with the family of two state discrimination problems that generate the posterior surface. Fig. 2–4C shows one example of this family.

2.7 Acknowledgments

We thank Uri Alon, Julie Desbarats, Michael Elowitz, Leon Glass, Terry Hebert, Moises Santillan, and particularly Sharad Ramanathan for helpful comments, and Yaki Setty and Uri Alon for supplying the data of Fig. 2–4A. P.S.S. holds a Tier II Canada Research Chair. P.S.S. and T.J.P. are supported by N.S.E.R.C. (Canada) and the MITACS National Centre of Excellence.

2.8 Appendix

2.8.1 Allosteric Model of Transcription Factors

We model transcription factors as allosteric molecules having two states: a DNA-binding state (B) and a non-DNA binding state (N). Following Monod-Wyman-Changeux [95], the presence of sugar causes a shift in the time the transcription factor spends in each state (Fig. 2–5). Sugar can bind to either the B or the N form of the transcription factor, but does so with a different binding affinity (K_b for the DNA-binding state and K_n for the non-DNA-binding state). Only when unbound by sugar can the transcription factor change between its two states. The reaction describing this change has an equilibrium constant of K_t. If $K_b \gg K_n$, sugar preferentially binds to the B state. By binding to the transcription factor, sugar converts B_0 molecules into B_r molecules, more so than N_0 molecules into N_r molecules (where the subscript r denotes that r sugar molecules are bound). The reaction between B_0 and N_0 is no longer at equilibrium and more N molecules convert to B molecules while this equilibrium is restored. The population of transcription factors as a whole is now more in the stronger sugar-binding B state, and so again more B than N molecules are likely to bind

sugar. This positive feedback means that the number of transcription factors in the B state can be a highly nonlinear function of the number of sugar molecules [95]. If $K_n \gg K_b$ the opposite behavior occurs, and sugar drives the transcription factors into the non-DNA binding N state.

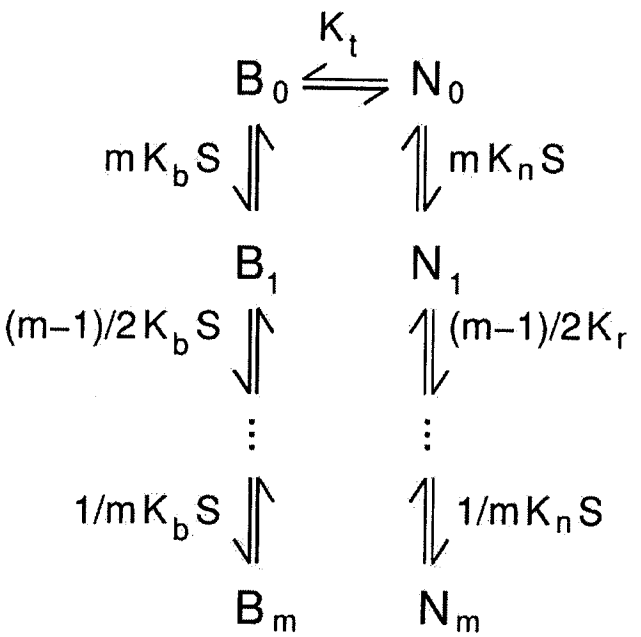


Figure 2–5: An allosteric transcription factor that binds sugar S and exists in a DNA-binding state (B) and a non-DNA binding state (N). Each sugar binding site is assumed identical, and the subscripts denote the number of bound sugar molecules. Consequently, the basic equilibrium association constants for sugar binding, K_b and K_n, are altered by the ratio of the number of sites available for binding sugar (which increase the forward rate of the reaction) to the number of bound sugars (which increase the backward rate).

We assume that both the total amount of sugar and the total amount of transcription factors are conserved:

$$S_{\text{tot}} = S + \sum_{r=0}^{m}(rN_r + rB_r) \tag{2.6}$$

$$T_{\text{tot}} = \sum_{r=0}^{m}(N_r + B_r) \tag{2.7}$$

where m is the number of sugar binding sites. Following [95], we assume that each reaction in Fig. 2–5 is at equilibrium:

$$\begin{aligned} K_t N_0 &= B_0 \\ mK_b S B_0 &= B_1 \\ (m-1)K_b S B_1 &= 2B_2 \\ \vdots &\quad \vdots \\ K_b S B_{m-1} &= mB_m \end{aligned} \tag{2.8}$$

Each equilibrium concentration can be solved in terms of N_0, the amount of transcription factor in the non-DNA binding state unbound by sugar:

$$\begin{aligned} N_r &= \binom{m}{r}(K_n S)^r N_0 \\ B_r &= \binom{m}{r}(K_b S)^r K_t N_0 \end{aligned} \tag{2.9}$$

Using these expressions and carrying out the summations in Eqs. 2.6 and 2.7 with the binomial theorem gives

$$S_{\text{tot}} = S + N_0 m S\left[K_n(1 + K_n S)^{m-1} + K_t K_b(1 + K_b S)^{m-1}\right] \tag{2.10}$$

$$T_{\text{tot}} = N_0\left[(1 + K_n S)^m + K_t(1 + K_b S)^m\right]. \tag{2.11}$$

For a given S_{tot} and T_{tot} and the equilibrium association constants K_t, K_b, and K_n, we numerically solve Eqs. 2.10 and 2.11 for the amount of free sugar, S, and for N_0. We can therefore calculate the total amount of transcription factor in the non-DNA binding state, $N = N_0(1 + K_n S)^m$, and the total amount in the DNA-binding state, $B = N_0 K_t (1 + K_b S)^m$.

2.8.2 Promoter Models

We consider three different models of the promoter (Fig. 2-2B and C). The type A model has just one operator site. The type B model has two operators: a transcription factor at either operator prevents or initiates transcription independently. The final model, type C, has two operators but only one is sufficiently close to the RNAP binding site to directly affect transcription. Nevertheless, a transcription factor bound to the inactive operator can stabilize a transcription factor bound to the active operator. We denote the fraction of time that the promoter is able to initiate transcription at equilibrium as promoter efficacy, P_{eff}.

We follow Shea and Ackers [129] to calculate the occupancy of the promoter at equilibrium. For example, for a negatively controlled type A promoter, which has just one binding site for a repressor, we consider the promoter existing in two states: P_1, bound by repressor, and P_0, not bound by repressor. If K_1 is the association constant for repressor binding and B is the number of repressors that are able to bind DNA, then $P_1 = K_1 B P_0$. The promoter is conserved: $P_0 + P_1 = 1$, if there is only one copy of the promoter. Combining these two equations implies that the promoter efficacy, P_0, obeys $P_0 = 1/(1 + K_1 B)$. We solve for the promoter efficacy for more complicated promoters similarly.

For a negatively controlled system, P_{eff} is the equilibrium fraction of promoter free from repressor. For the different promoter models:

type A:

$$P_{\text{eff}} = \frac{1}{1 + K_1 B} \tag{2.12}$$

type B:

$$P_{\text{eff}} = \frac{1}{1 + (K_1 + K_2)B + K_1 K_2 B^2} \tag{2.13}$$

type C:

$$P_{\text{eff}} = \frac{1 + K_2 B}{1 + (K_1 + K_2)B + \frac{1}{2}(K_1 K_2 + K_1 K_2 K_c)B^2} \tag{2.14}$$

where B is the total amount of transcription factor in the DNA-binding form, K_1 and K_2 are association constants for transcription factor binding to the two operator sites, and K_c determines the degree of cooperativity between two interacting, DNA bound transcription factors.

For positively controlled systems, P_{eff} is the equilibrium fraction of promoter bound by activator. For

type A:

$$P_{\text{eff}} = \frac{K_1 B}{1 + K_1 B} \tag{2.15}$$

type B:

$$P_{\text{eff}} = \frac{(K_1 + K_2)B + K_1 K_2 B^2}{1 + (K_1 + K_2)B + K_1 K_2 B^2} \tag{2.16}$$

type C:

$$P_{\text{eff}} = \frac{K_1 B + \frac{1}{2}(K_1 K_2 + K_1 K_2 K_c)B^2}{1 + (K_1 + K_2)B + \frac{1}{2}(K_1 K_2 + K_1 K_2 K_c)B^2} \tag{2.17}$$

Note when $K_c = 1$, that is, no cooperative interaction between the transcription factors, the type C models do not reduce to the type B models because only one operator is active for type C whereas both are active for type B.

2.8.3 Generating the posteriors

To generate a set of two state classification problems, we assumed that each state can be described by a lognormal distribution:

$$P(S|\text{state}_i) = \frac{e^{\frac{-(\ln S-\mu_i)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i S} \tag{2.18}$$

The low state has a sugar distribution with mean μ_1 and standard deviation σ_1; the high state has a mean μ_2 and standard deviation σ_2. We choose μ_1 to be either 1, 3, or 5; μ_2 to be either 5.1, 6.6, 8.1, or 9.6; σ_1 to be either 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9; and σ_2 to be either 1, 1.25, 1.5, 1.75, or 2. All possible combinations of these parameters were considered, and we chose fifty pairs of distributions that best gave a range of different posterior probabilities (Fig. 2–6).

2.8.4 Fitting the Models to the Posteriors

We used a least square fit to score how well a model matches the posterior probability of the high state. The residuals plotted in Fig. 2-2 are the minimum value of the sum of squares:

$$\sum_i^n \Big[P(S_i|\text{high}) - P_{\text{eff}}(S_i,\lambda)\Big]^2 \tag{2.19}$$

where we have n sugar levels S_i leading to n points on the posterior probability curve, $P(S|\text{high})$, we are trying to fit, and $P_{\text{eff}}(S,\lambda)$ is the model prediction for the promoter efficacy. This prediction is a function of the set of parameters λ: K_t, K_n, K_b, K_1, and K_2 and K_c depending on the promoter type. The minimum value of Eq. 2.19 occurs at the best fit set of parameters λ. To ensure that the fitting algorithm considers only non-negative parameters, we define new variables for each

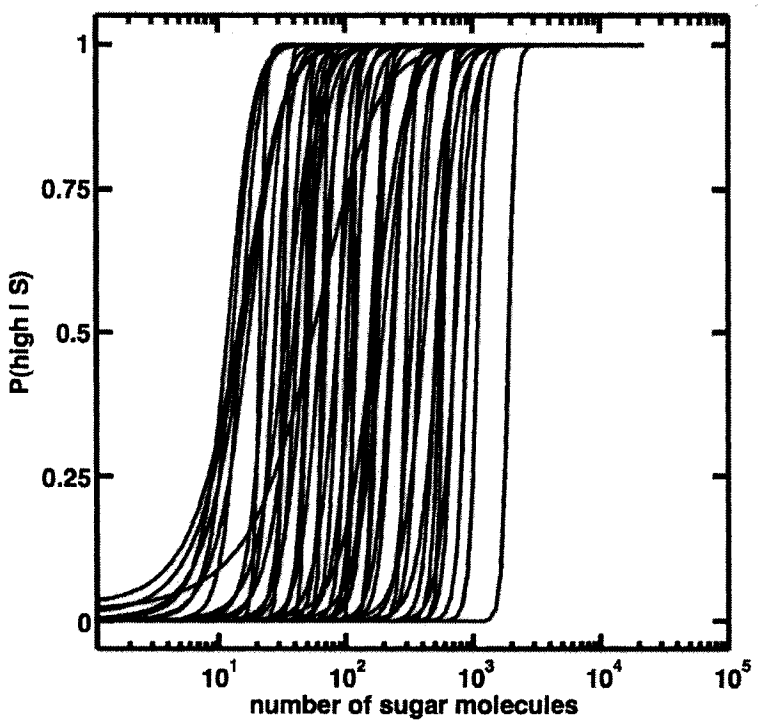


Figure 2–6: The collection of posterior probabilities that were generated as solutions of lognormal two state classification problems and used to compare the different genetic models of Fig. 2-2 as Bayesian classifiers.

parameter in log space. For example, $\kappa_1 = \log(K_1)$, and therefore can range over positive and negative values [131].

To correctly compare the ability of different models to fit a data set, models with more parameters should be penalized because they have more freedom to match the data. Typical methods are the Bayes Information Criterion (BIC) [132] and the Laplace method for model selection [115]. Using both these techniques to compare the different models, the results of Fig. 2-2D–F were qualitatively unchanged. For the Laplace method, we need the maximum likelihood of the data (the 100 posterior probability points in our case) given the model. For each parameter, the maximum likelihood is penalized by a term that is determined by the error in the best fit value of the parameter and by its prior [115]. We use

$$\left(\sum_{i}^{n}\Big[P(S_i|\text{high}) - P_{\text{eff}}(S_i,\lambda)\Big]^2\right)^{-\frac{(n-1)}{2}} \tag{2.20}$$

for the likelihood. This distribution results from assuming that the data have normally distributed errors with zero mean and any non-negative standard deviation [115]. It is maximized when the sum of squares residual, Eq. 2.19, is minimized.

2.8.5 Parameter Sensitivity

The sensitivities of the parameters were calculated as the mean log gain sensitivities [133] of the promoter efficacy. For parameter p_j, the sensitivity, χ_j, is

$$\chi_j = \left\langle \frac{\partial \log P_{\text{eff}}}{\partial \log p_j} \right\rangle \tag{2.21}$$

where the angled brackets denote an average over all sugar concentrations. We analytically calculated the $\partial \log P_{\text{eff}}/\partial \log p_j$ derivative as an implicit function of

$\partial N_0/\partial p_j$ and $\partial S/\partial p_j$ by differentiating the promoter efficacy, such as Eq. 2.15 for example. We calculated these last two derivatives by differentiating Eqs. 2.10 and 2.11 with respect to p_j and numerically solving the resulting equations. Sensitivity values are given in Table 2–1.

Table 2–1: Parameter sensitivities for repressor and activator models. Parameters are defined in Fig. 2-2.

	Repressor model	**Activator model**
K_t	0.07	0.10
K_n	0.21	0.004
K_b	0.004	0.20
K_1	0.07	0.07
K_2	0.02	0.04
K_c	0.06	0.04

2.8.6 Robustness of the Best-fit Parameters

The fits of the promoter efficacy to the posterior probability curves are robust to changes in all but two of the parameters specifying each model. To investigate this robustness, we considered the model that best fit the posterior probability curves of Fig. 2–6. This model is transcriptionally controlled by a repressor that has four sugar binding sites. We varied each parameter individually and calculated the average change in the sum of squares residual, Eq. 2.19, over all the posterior curves. The results shown in Fig. 2–7 reflect Table 2–1: the fit is only significantly sensitive to K_n, the sugar binding affinity for the non-DNA binding form of the repressor, and to a much lesser extent to K_t, the affinity describing transitions between the DNA- and non-DNA binding forms. Nevertheless, the sum of squares residual is so small for this model that the promoter efficacy curves behave like the posterior probability of Fig. 2-1C even if the residual is increased 5000-fold. We

comment on possible implications of the high sensitivity to K_n in the discussion section.

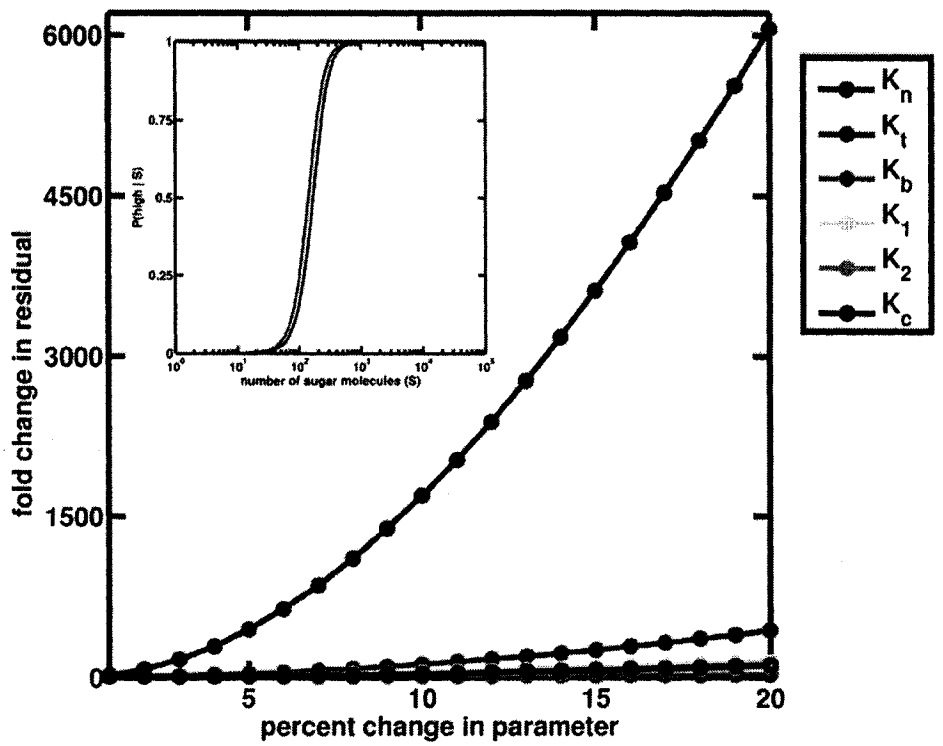


Figure 2–7: Robustness of the sum of squares fit to systematic perturbations in individual model parameters away from their best-fit values. The model that best fits the posterior probabilities of Fig. 2–6 is shown: this model has promoter type C and is negatively regulated by a repressor with four sugar binding sites. Parameters are defined in Fig. 2-2. The inset shows an example of the promoter efficacy curves where K_n is changed by 20%. The curves are very similar despite the residual for the upper red curve being almost 5000-fold larger than the residual of the original blue curve.

2.8.7 Stochastic Simulation Details

To confirm that genetic networks can perform inference in real time with a noisy sugar source, we simulated both a repressor and an activator model with

fluctuating sugar levels. We chose the posterior of Fig. 2-1C and the repressor and activator model that fit it best (parameters are given in Table 2–2).

Table 2–2: Parameter values for the simulation shown in Fig. 2-3. These values are association affinities and are the best fit values of the networks to the posterior probability of Fig. 2-1C. Each association affinity is dimensionless because we simulate with numbers of molecules rather than concentrations. Shown in brackets is the corresponding dissociation rate. These rates, which are not given by a fit to $P(\text{high}|S)$, were chosen so that the network would respond in a reasonable time to changes in sugar levels.

	Repressor model	**Activator model**
K_t	1.27 (10 s)	6.21×10^{-7} (10 s)
K_n	9.45×10^{5} (10 s)	3.04×10^{4} (10 s)
K_b	233 (10 s)	1.33×10^{6} (10 s)
K_1	3.41×10^{6} (0.1 s)	3.62×10^{6} (0.1 s)
K_2	3.34×10^{10} (0.1 s)	1.51×10^{9} (0.1 s)
K_c	88.3 (10 s)	219 (10 s)

To generate a relatively smooth time series of sugar levels, we used a Markov chain Monte Carlo method [115] to sample from the distributions in Fig. 2-1C (the Metropolis algorithm with a Gaussian trial distribution). We sample from the low distribution for 10^4 seconds, then from the high distribution for 10^4 seconds, and the again from the low distribution for another 10^4 seconds. For each sugar sample, the cytosolic sugar levels in the simulation are changed to the new sampled value. A stochastic simulation of the genetic network is then run for a fixed time interval (either 5, 10, 25, 50, or 100 seconds) using the Gibson-Bruck version [134] of the Gillespie algorithm [130]. The probability of a given reaction per unit time is equal to the product of the kinetic rate for the reaction and the number of potential reactants present. The time steps between reactions obey a Markov process. The cytosolic sugar level is then re-sampled using the Markov

chain Monte Carlo method and another Gillespie simulation run for this new level of sugar. The promoter efficacy plotted in Fig. 2-3 is the average promoter efficacy generated during each run of the Gillespie algorithm. Simulations start with one DNA molecule, 25 transcription factors in the DNA binding state and 25 transcription factors in the non DNA binding state.

For each choice of sugar sampling interval, we compared the performance of the two networks (Fig. 2-3C and D) to the instantaneous posterior probability (Fig. 2-3B). The comparison was scored by measuring the mean over time of the absolute difference between promoter efficacy and the instantaneous posterior probability. The results are shown in Table 2–3. Both networks perform better as the sugar sampling interval increases. As the time period grows over which the promoter efficacy is averaged, the average more closely matches the posterior probability of Fig. 2-2C (for a long sampling period, the promoter efficacy will match the posterior probability almost perfectly because we use the best fit parameters for the simulation).

Table 2–3: Comparison scores of the mean absolute difference between the promoter efficacy and the instantaneous posterior probability of the high sugar state. Each score is the average from five simulation runs. A score of zero implies the the promoter efficacy exactly follows the instantaneous posterior probability. The sampling interval is the time between the samples of sugar used to generate the sugar time series.

Sampling interval	**Repressor model**	**Activator model**
5 secs	7.0×10^{-2}	12.5×10^{-2}
10 secs	3.3×10^{-2}	6.3×10^{-2}
25 secs	3.4×10^{-2}	7.4×10^{-2}
50 secs	2.8×10^{-2}	5.0×10^{-2}
100 secs	2.5×10^{-2}	4.4×10^{-2}

Negatively controlled networks consistently performed better because the network is better able to use its cooperativity (see the argument given earlier).

2.8.8 The Inverse Gaussian Classification Problem

A bivariate, or two dimensional, Gaussian distribution is a function of a vector (s_1, s_2) and is specified by a mean vector (μ_1, μ_2) and a 2×2 covariance matrix $\boldsymbol{\sigma}$. For example, μ_1 is the mean of the s_1 variable and σ_{11} its variance. $P(s_1, s_2)$ obeys

$$P(s_1, s_2) \sim \frac{1}{\sqrt{\det(\boldsymbol{\sigma})}} \exp\left(-\frac{1}{2}\sum_{i,j}(s_i - \mu_i)\sigma_{ij}^{-1}(s_j - \mu_j)\right) \tag{2.22}$$

where $\boldsymbol{\sigma}^{-1}$ is the matrix inverse of $\boldsymbol{\sigma}$.

A two state, bivariate Gaussian classification problem is described by the prior probabilities of the two states, $P(\mathrm{II})$ and $P(\mathrm{I}) = 1 - P(\mathrm{II})$; the mean $\boldsymbol{\mu}^{\mathrm{I}}$ and covariance matrix $\boldsymbol{\sigma}^{\mathrm{I}}$ for s_1 and s_2 for state I; and the mean $\boldsymbol{\mu}^{\mathrm{II}}$ and covariance matrix $\boldsymbol{\sigma}^{\mathrm{II}}$ for s_1 and s_2 for state II. Given an observation of s_1 and of s_2, the posterior probability of state II is

$$\begin{aligned} P(\mathrm{II}|s_1, s_2) &= \frac{P(s_1, s_2|\mathrm{II})P(\mathrm{II})}{P(s_1, s_2|\mathrm{I})P(\mathrm{I}) + P(s_1, s_2|\mathrm{II})P(\mathrm{II})} \\ &= \left(1 + \frac{P(s_1, s_2|\mathrm{I})P(\mathrm{I})}{P(s_1, s_2|\mathrm{II})P(\mathrm{II})}\right)^{-1} \end{aligned} \tag{2.23}$$

Inserting Eq. 2.22 in Eq. 2.23 gives

$$P(\mathrm{II}|s_1, s_2) = \left(1 + \sqrt{\frac{\det(\boldsymbol{\sigma}^{\mathrm{II}})}{\det(\boldsymbol{\sigma}^{\mathrm{I}})}} \times \frac{\exp(-\frac{1}{2}\sum_{i,j}(s_i - \mu_i^{\mathrm{I}})(\sigma^{\mathrm{I}})_{ij}^{-1}(s_j - \mu_j^{\mathrm{I}}))}{\exp(-\frac{1}{2}\sum_{i,j}(s_i - \mu_i^{\mathrm{II}})(\sigma^{\mathrm{II}})_{ij}^{-1}(s_j - \mu_j^{\mathrm{II}}))} \times \frac{1 - P(\mathrm{II})}{P(\mathrm{II})}\right)^{-1} \tag{2.24}$$

From the posterior surface $P(\mathrm{II}|s_1, s_2)$, we would like to recover the parameters of the classification problem: $P(\mathrm{II})$, $\boldsymbol{\mu}^{\mathrm{I}}$, $\boldsymbol{\sigma}^{\mathrm{I}}$, $\boldsymbol{\mu}^{\mathrm{II}}$, and $\boldsymbol{\sigma}^{\mathrm{II}}$. This recovery is degenerate — different sets of parameters can result in the same posterior surface.

With a little algebra, Eq. 2.24 can be reduced to the general form

$$P(\mathrm{II}|s_1, s_2) = \left[1 + \exp(c_0 + c_1 s_1 + c_2 s_2 + c_3 s_1 s_2 + c_4 s_1^2 + c_5 s_2^2)\right]^{-1} \tag{2.25}$$

where the c_i depend on the parameters $P(\mathrm{II})$, $\boldsymbol{\mu}^{\mathrm{I}}$, $\boldsymbol{\sigma}^{\mathrm{I}}$, $\boldsymbol{\mu}^{\mathrm{II}}$, and $\boldsymbol{\sigma}^{\mathrm{II}}$. Although these parameters have 11 degrees of freedom ($P(\mathrm{II})$, two each for vectors $\boldsymbol{\mu}^{\mathrm{I}}$ and $\boldsymbol{\mu}^{\mathrm{II}}$, and three each for the covariance matrices $\boldsymbol{\sigma}^{\mathrm{I}}$ and $\boldsymbol{\sigma}^{\mathrm{II}}$), the posterior surface only has 6 degrees of freedom. The parameters therefore have 5 unrecoverable degrees of freedom.

2.8.9 Fitting the Transcription Rate Surface from the *lac* Operon

To fit the Setty *et al.* data [120], we used Eq. 2.25, with s_1 corresponding to the logarithm of the IPTG concentration and s_2 corresponding to the logarithm of the cAMP concentration. As the base of the logarithm and a constant offset can be absorbed by the coefficients c_i, we chose to let $s_1 \in \{0, 1, \ldots, 5\}$ correspond to the six sample levels of IPTG and $s_2 \in \{0, 1, \ldots, 9\}$ correspond to the 10 sample levels of cAMP. We used a simplex search method (`fminsearch` in Matlab, the Mathworks, Massachusetts) to optimize the six parameters $c_0, c_1, \ldots, c_5$ so that the sum-squared error between Eq. 2.25 and the *lac* transcription data was minimized. We used multiple optimization runs and experimented with different initial conditions, but these factors seem to have little influence on the outcome of optimization. All or nearly all runs converged to essentially the same solution, which we therefore take to be close to optimal. The final parameters found were:

c_0	c_1	c_2	c_3	c_4	c_5
4.09	−1.88	0.15	−0.11	0.32	−0.06

which define the surface shown in Fig. 2-4B. There is not a unique two state, bivariate Gaussian discrimination problem corresponding to these parameters (as

described above). Of the many discrimination problem parameter sets consistent with the optimized c_i, we chose one by making the following assumptions:

$$(\boldsymbol{\sigma}^{\mathrm{I}})^{-1} = \begin{bmatrix} 0.4 & -c_3 \\ -c_3 & 0.3 \end{bmatrix} \tag{2.26}$$

$$(\boldsymbol{\sigma}^{\mathrm{II}})^{-1} = (\boldsymbol{\sigma}^{\mathrm{I}})^{-1} + \begin{bmatrix} 2c_4 & 0 \\ 0 & 2c_5 \end{bmatrix} \tag{2.27}$$

$$\boldsymbol{\mu}^{\mathrm{I}} = \begin{bmatrix} 1.5 \\ 3.5 \end{bmatrix} \tag{2.28}$$

$$\boldsymbol{\mu}^{\mathrm{II}} = \left(\begin{bmatrix} -c_1 \\ -c_2 \end{bmatrix} + (\boldsymbol{\sigma}^{\mathrm{I}})^{-1}\boldsymbol{\mu}^{\mathrm{I}} \right) \boldsymbol{\sigma}^{\mathrm{II}} \tag{2.29}$$

$$P(II) = (1 + \mathrm{e}^{z})^{-1} \tag{2.30}$$

where

$$z = c_0 + \frac{1}{2}(\boldsymbol{\mu}^{\mathrm{I}})^T(\boldsymbol{\sigma}^{\mathrm{I}})^{-1}\boldsymbol{\mu}^{\mathrm{I}} - \frac{1}{2}(\boldsymbol{\mu}^{\mathrm{II}})^T(\boldsymbol{\sigma}^{\mathrm{II}})^{-1}\boldsymbol{\mu}^{\mathrm{II}} + \frac{1}{2}\log\left[\frac{\det(\boldsymbol{\sigma}^{\mathrm{I}})}{\det(\boldsymbol{\sigma}^{\mathrm{II}})}\right] \tag{2.31}$$

which results in the distinct lognormal distributions in Fig. 2-4C. The five parameters unrecoverable from the posterior surface can be seen in our arbitrary choices for $\boldsymbol{\mu}^{\mathrm{I}}$, the diagonal elements of $(\boldsymbol{\sigma}^{\mathrm{I}})^{-1}$, and the off-diagonal elements (zero) added to $(\boldsymbol{\sigma}^{\mathrm{I}})^{-1}$ to make $(\boldsymbol{\sigma}^{\mathrm{II}})^{-1}$.

CHAPTER 3
Gene Networks in Atrial Fibrillation

In the previous chapter, we explored a transcription regulatory network model and uncovered functional properties. Here, we look at microarray data from two experimental models of atrial fibrillation and use different techniques to decipher the differences in gene regulation. By classifying the differentially expressed genes, we see if the histological observations can be explained at the transcriptional level. Towards the end of the chapter, we assemble a protein interaction network based on the differentially expressed genes and hypothesize on the mechanisms of pathology in one of the models.

3.1 Abstract

Gene-expression changes in atrial fibrillation patients reflect both underlying heart-disease substrates and changes because of atrial fibrillation-induced atrial-tachycardia remodeling. These are difficult to separate in clinical investigations. This study assessed time-dependent mRNA expression-changes in canine models of atrial-tachycardia remodeling and congestive heart failure. Five experimental groups (5 dogs/group) were submitted to atrial (ATP, 400 bpm for 24 hours, 1 week, or 6 weeks) or ventricular (VTP, 240 bpm for 24 hours or 2 weeks) tachypacing. The expression of ~21,700 transcripts was analyzed by microarray in isolated left-atrial cardiomyocytes and (for 18 genes) by real-time RT-PCR. Protein-expression changes were assessed by Western blot. In VTP, a large number of significant mRNA-expression changes occurred after both 24 hours (2,209) and 2 weeks (2,720). In ATP, fewer changes occurred at 24 hours (242) and fewer still (87) at 1 week, with no statistically significant alterations at 6 weeks. Expression changes in VTP varied over time in complex ways. Extracellular matrix-related transcripts were strongly upregulated by VTP consistent with its pathophysiology, with 8 collagen-genes upregulated >10-fold, fibrillin-1 8-fold and MMP2 4.5-fold at 2 weeks (time of fibrosis) but unchanged at 24 hours. Other extracellular matrix genes (eg, fibronectin, lysine oxidase-like 2) increased at both time-points (~10, ~5-fold respectively). In ATP, mRNA-changes almost exclusively represented downregulation and were quantitatively smaller. This study shows that VTP-induced congestive heart failure and ATP produce qualitatively different temporally-evolving patterns of gene-expression change, and that specific transcriptomal responses associated with atrial fibrillation versus underlying heart disease substrates must be considered in assessing gene-expression changes in man.

3.2 Introduction

Atrial fibrillation (AF) is the most common sustained cardiac rhythm disorder, and with the aging of the population both the prevalence and economic impact of AF are increasing progressively [135]. Although the mechanistic basis of AF remains incompletely understood, active research promises to provide new insights that may lead to improved therapeutic options [50, 136].

A variety of animal models have been used to assess AF pathophysiology under controlled conditions. Atrial tachyarrhythmias, including AF itself, alter atrial electrophysiology in ways that promote AF vulnerability [53, 59, 137]. Experimentally-induced congestive heart failure (CHF) also creates a substrate for AF maintenance, but by quite different mechanisms [54]. The atrial-tachycardia remodeling paradigm shows prominent changes in ion-channel function that lead to action-potential abbreviation and the promotion of atrial reentry [138, 139]. CHF-induced ionic-current changes do not promote reentry but may favor ectopic-impulse formation [140], and CHF-induced fibrosis promotes reentry by interfering with intra-atrial conduction [54].

The molecular basis of AF remains unclear. Gene microarray technology permits large-scale analysis of cardiac gene-expression changes and has been applied to compare AF patients with those in sinus rhythm. Expression profiling has pointed to several AF-related gene-expression changes [69, 71, 72, 141, 142], including alterations associated with oxidative stress [71], a ventricular-like expression signature [69] and changes in ion-transporters [142]. A limitation of this type of clinical gene-expression study is that it is very difficult to differentiate between AF-promoting changes caused by AF and those because of underlying cardiac disease. The analysis is further complicated by systematic inter-group differences in drug therapy, atrial size, and other cardiac variables.

Animal models of AF allow for greater control over study conditions and permit observations of the time course of any alterations. A human DNA microarray containing 6,035 cDNA probes applied to a porcine model of AF pointed to changes in myosin light chain-2 expression [70]. DNA microarrays with probes for canine-gene transcripts have recently become commercially-available. We designed the present study to analyze changes in canine cardiac gene-expression in two AF models: atrial-tachycardia remodeling induced by atrial tachypacing (ATP) and CHF-related remodeling produced by ventricular-tachypacing (VTP). Assessments were initially obtained at two time-points in each model: early after the onset of tachypacing (24 hours) and at a time of near steady-state remodeling (1 week for ATP [138], 2 weeks for VTP [66]). After initial studies showed that AF duration increases were smaller in 1-week ATP dogs versus 2-week VTP dogs, we added another group subjected to 6-week ATP.

3.3 Materials and Methods

3.3.1 Animal Model and Preparation

These methods followed previous publications [54, 59, 140, 143, 144]. All experiments were performed in male mongrel dogs weighing 25 to 32 kg. Animal handling was in accordance with the Guide for the Care and Use of Laboratory Animals published by the U.S. National Institutes of Health. In the initial series of experiments, 5 groups (n=5/group) were studied. Two groups were subjected to right ventricular tachypacing (VTP) for 24 hours or 2 weeks. Under sterile technique, a unipolar tined pacing lead (Medtronic) was inserted into the right-ventricular apex via the jugular vein under 1.5%-halothane anesthesia and

attached to a pacemaker in the neck programmed to 240 beats per minute. Two other groups were subjected to right-atrial tachypacing (ATP) for 24 hours and 1 week. Under sterile technique, a unipolar tined pacing lead (Medtronic) was inserted into the right atrium via the jugular vein under halothane anesthesia. This pacing lead was attached to a pacemaker implanted in the neck programmed to capture the atrium at 400 beats per min. Complete AV-block was induced by radiofrequency-ablation to prevent a ventricular tachyarrhythmic response to ATP. A ventricular demand pacemaker programmed to maintain the ventricular rhythm at $\geq$ 80 beats/min was implanted and connected to a unipolar lead inserted into the right ventricle via a jugular vein. A final group of VTP-sham control animals was handled identically to 24-hour VTP-dogs, but their pacemaker was not activated.

A second series of experiments was performed in 3 additional groups of dogs (n=5/group). Because 1-week ATP proved not to promote AF as much as 2-week VTP, we added a 6-week ATP group. We also tested for potential intervention-related differences by concurrently studying VTP and ATP shams. The VTP-shams were prepared as described above. The ATP-shams were prepared and handled identically to the 24-hour ATP-dogs, but their atrial pacemakers were not activated. VTP-shams were included in both experimental series because they were the primary control group for all analyses. For biochemical analyses in each series, experimental and sham animals were handled concurrently with DNA extraction and microarray processing performed on the same days with the same reagents and on the same batches of microarrays to minimize variability. After preparation periods, dogs were anesthetized (morphine, 2 mg/kg s.c.; α-chloralose, 120 mg/kg i.v. load; 29.25 mg/kg/hr maintenance infusion) and ventilated. In

vivo measurements were obtained and an isolated-cardiomyocyte preparation was snap-frozen for subsequent analysis.

All the animals received antibiotics: Penicillin G, Longisil, Vetoquinol Canada, 2 ml (containing penicillin G benzathine 300,000 IU + penicillin G procaine 300,000 IU) intra-muscular pre-operatively; and enrofloxacin, Baytril, Bayer, 150 mg oral/day for up to 5 days post-operatively. On study days, an ECG was recorded to confirm continued pacemaker-capture and the pacemaker was deactivated.

3.3.2 In Vivo Measurements and Cell Isolation

A median sternotomy was performed and Teflon-coated stainless steel electrodes were hooked into the RA appendage (RAA) for electrophysiological measurements. Atrial effective refractory period (AERP) was measured in the RAA with the extrastimulus technique at basic cycle lengths (BCLs) of 360, 300, 250, 200 and 150 ms, with 1 minute for steady-state conditions at each BCL. AF was induced with four times threshold intensity burst pacing (10 Hz, 5-10 s) to measure mean AF duration (DAF) in each dog as previously described [59, 54, 138]. For DAF $<$ 5 min, 15 measurements were performed; for DAF of 5-10 min, 10 measurements were performed; for DAF of 10-20 min, five measurements were performed; for DAF $>$20 min, three measurements were performed. Left-ventricular end-diastolic pressure was measured at the end of each experiment.

Animals were euthanized, hearts removed and placed in Tyrode solution (contents in mmol/L): NaCl 136.0, KCl 5.4, $MgCl_2$ 1.0, HEPES 5.0, Na_2HPO_4 3.3, glucose 10.0, $CaCl_2$ 2.0, pH 7.4 (adjusted with NaOH) equilibrated with 100% O_2 for dissection. The left atrium was perfused via the left circumflex coronary artery

and cell-isolation was performed via collagenase digestion according to previously-described methods [143, 144]. In brief, the coronary artery was cannulated and perfused with Tyrode solution at 37°C. All leaking arterial branches were ligated with silk thread to ensure adequate perfusion. The tissue was then perfused with nominally Ca^{2+} -free Tyrode solution for 15 minutes, followed by 40 minute-perfusion with the same solution supplemented with collagenase (0.4 mg/mL, CLSII, Worthington Biochemical) and 1% bovine serum albumin (Sigma-Aldrich). Cells were separated by gentle trituration with pipettes. Resuspended cells were filtered through a 200-μm sieve to remove tissue residues. Microscopic examination ensured a minimum of 80% rod-shaped cardiomyocytes. Cells were collected, washed and centrifuged at 1000 rpm for 3 minutes. The supernatant was removed and cardiomyocyte-enriched pellets frozen in liquid-N_2 were kept at -80°C until RNA extraction. Electrophysiological measurements were obtained in the RAA and cell isolation/biochemical analysis performed with LA tissue because in our previous experience tissue trauma caused by atrial manipulation for electrophysiological measurement can affect biochemical determinations.

3.3.3 RNA Extraction

For RNA extraction, cell pellets were immersed into Trizol (1 mL/100mg of pellet) and pulverised for 15 seconds with a Polytron at 12,000 rpm. Chloroform (100 μL/mL of Trizol) was added and samples were incubated on ice for 15 minutes. Samples were centrifuged at 8,000 rpm for 15 minutes at 4°C. The aqueous phase was transferred into new tubes and an equal volume of chloroform was added. Tubes were shaken vigorously and centrifuged at 8,000 rpm for 5 minutes at 4°C. The aqueous phase was transferred into new tubes and an equal

volume of isopropanol was added. Samples were incubated at -20°C for 45 minutes and centrifuged at 8,000 rpm for 5 minutes at 4°C. The supernatant was removed and pellets resuspended in 1.5 mL ethanol. The tubes were then centrifuged at 13,000 rpm for 5 minutes at 4°C. Pellets were resuspended in 70% ethanol and incubated overnight at -20°C. Samples were then centrifuged at 13,000 rpm for 5 minutes at 4°C. The supernatant was aspirated and pellets dried for 30 minutes. The pellets were then resuspended in DEPC water. RNA concentrations were quantified and assessed for purity by measuring optical density at 260 nm and 280 nm by running the samples on RNA 6000 nano-chips from Agilent with the Agilent 2100 Bioanalyzer (Model G2938B). The quality was also verified by running samples onto 2.5% agarose gels. Samples with OD ratio 260/280 nm >1.8 were selected for microarray processing.

3.3.4 Canine Genome Microarrays

The microarrays we used (Affymetrix GeneChip Canine Genome Array) are high-density oligonucleotide arrays (11-μm spots) containing 23,836 probe sets of 25-mer length probes designed to detect a total of 21,700 transcripts. Multiple ($\sim$11) pairs of probes are used to measure the level of transcription of each sequence represented on the array. The microarrays were synthesized by Affymetrix using photolithographic and combinatorial chemistry methods applied on 5" x 5" quartz wafers. The sequence information for the array includes public content from Genebank (release 137.0, August 2003), dbEST (October 2003), and proprietary beagle sequence content licensed from LION Bioscience AG. LION Bioscience sequence information was derived from sequences in cDNA libraries

for the following eleven tissues: testis, ovary, brain, embryo, liver, spleen, kidney, muscle, aorta, uterus, and jejunum.

The quality control of gene-chip arrays was monitored by Affymetrix via several control points including automated software tests during array design, tracking specific probe synthesis sequences during array synthesis and signal intensity tests with hybridization control sequences (*bioB*, *bioC*, *bioD*) and polyA probe sets (*dap*, *lys*, *phe*, and *thr*). Canine-specific housekeeping genes representing adrenergic receptors, glucose-6-phosphatase, and glyceraldehyde-3-phosphate were used as on-chip controls. The chip also contained 25-mer probes with exact sequence (perfect match) paired with a 25-mers probes containing a single point mutation (mismatch). The paired mismatch probe can be used to detect and eliminate false or contaminating fluorescence within that measurement.

3.3.5 RNA Processing on Arrays

From each sample, 10 μg of total RNA was used for the experiment. Affymetrix GeneChip®one-cycle target labeling and control reagents kit was used according to the protocol from Affymetrix (GeneChip®Analysis Technical Manual). The target cRNA derived from each sample was verified for quality on Agilent Bioanalyzer before fragmentation and 15 μg of fragmented cRNA was hybridized to the Affymetrix GeneChip®Canine Genome array. The chips were stained and washed using the GeneChip®Fluidics Station 450 and visualized on an Affymetrix GeneChip®Scanner 3000 according to Affymetrix protocol.

To synthesize first-strand cDNA, 20 μg of RNA was incubated with T7-T24 primers at 70°C for 10 minutes and reverse transcription was performed with

Superscript II reverse transcriptase and dNTPs at 42°C for 1 hour. The second-strand cDNA was synthesised with DNA ligase, DNA polymerase I and Rnase H in the presence of dNTPs. T4 DNA polymerase was then added to create blunt ends and the reaction was stopped using EDTA. Phenol extraction followed by ethanol precipitation was used to clean up cDNA by removing enzymes and excess dNTPs. Double-stranded cDNA was transcribed to labeled cRNA with T7 RNA polymerase in the presence of Biotin-labeling ribonucleotides, HY reaction buffer, and Rnase inhibitor mix. Free-labeled ribonucleotides were removed with Rneasy columns. Purified labeled cRNA concentrations were measured with spectrophotometry at 260 nm. cRNA was fragmented into a fragmentation buffer to obtain 100-bp length products. Products were controlled for quality with an Agilent Bioanalyser. Hybridization of 15 μg of fragmented cRNA to the probe arrays was performed in presence of Herring Sperm DNA, Control oligo B2. Hybridized target cRNA were stained with streptavidin phycoerythrin and arrays were scanned using a GeneArray Scanner at an excitation wavelength of 488 nm and emission wavelength of 570 nm.

3.3.6 Statistical Analysis of Microarray Data

The microarray expression data were analyzed using a combination of algorithms. We first applied the Invariant Set Normalization method [30] in dChip [29], which corrects for inter-array differences in average brightness. To integrate each gene's probe intensities into one value representative of gene expression, we used dChip to calculate the Model Based Expression Index [29]. dChip was set to use only the intensities of the perfect match probes as well as to detect single, probe, and array outliers. After formulating the gene-expression values, we used

Significance Analysis for Microarrays (SAM) [39] to detect differentially-expressed genes, accepting only genes with a q-value under one.

For genes without Affymetrix annotations, we used BLAST [145, 146] to find sequence homologies in mammals and chose only those with an E value under 10^{-4}. To enhance the annotations with functional information, we used Affymetrix's human to canine microarray comparisons to map canine genes to their human equivalents. With the human equivalents, we queried the Gene Ontology (GO) database for functional information. Genes not identified by this process were classified after identification by literature search.

3.3.7 Real-time RT-PCR

Microarray-based expression ratios were confirmed with real-time RT-PCR for 18 selected genes. First-strand cDNA was synthesized from 2 μg of total RNA using the High Capacity cDNA Archive Kit for RT-PCR (Applied Biosystems) for each group. On-line PCR was performed with FAM-labeled fluorogenic TaqMan probes and primers (Assay-by-design, Applied Biosystems) and TaqMan Universal Master Mix (Applied Biosystems). After 2 minutes at 50°C and 10 minutes at 95 °C, 40 amplification cycles (15 seconds at 95 °C and 1 minute at 60°C) were performed with the Gene Amp 5700 Sequence Detection System (Perkin-Elmer Biosystems). The fluorescence signals were normalized to the gene encoding 18S-ribosomal RNA and analysed with the comparative threshold cycle (Ct) relative-quantification method. For each sample from each dog, each gene was quantified in duplicate. Forward and reverse primers and TaqMan probe-sequences are provided online.

Data are expressed as means ± SEM. Comparisons among group means (Tables 3-1, 3-2, 3-3, 3-4, and 3-5) were performed with one-way analysis of variance (ANOVA) followed by Dunnett's test for individual-mean comparisons relative to control (VTP-sham). A two-tailed $P<.05$ was considered statistically significant.

3.3.8 Western Blot Analysis

The expression of selected genes was verified at the protein level by Western blot. Isolated-cardiomyocyte pellets were immersed in lysis buffer (10 mmol/L Tris-HCl, 0.32 mol/L sucrose, 5 mmol/L EDTA, 1% Triton X-100, 2 mmol/L DTT, 1 mmol/L phenymethylsulfonyl fluoride, 10 μg/mL leupeptin, 10 μg/mL pepstatin, 10 μg/mL aprotinin, 20 mmol/L NaF, 1 mmol/L Na_3VO_4) and pulverised with a Polytron at 10,000 rpm. The homogenates were then incubated on ice for 30 minutes, submitted to 3 freeze/thaw cycles and centrifuged at 13,000 rpm for 10 minutes. The supernatant was collected into new tubes. Protein concentrations were determined by Bradford assay. Equal amounts of cellular protein extracts (100 μ/sample) were separated by electrophoresis on SDS-polyacrylamide gels. Proteins were transferred to nitrocellulose membranes and incubated with 5% non-fat dry milk in TBST (TBS, pH 7.4 with 0.1% Tween-20) for 1.5-2 hours at room temperature. The antibodies used are listed online. Bands were quantified with QuantityOne software and calculated as a ratio over the corresponding VTP-sham sample in the same gel. Results were expressed relative to GAPDH band-intensity on the same samples. Data are expressed as means ±SEM. Comparisons among group means were performed with two-way analysis of variance

(ANOVA; group and stage as factors) followed by Bonferroni-adjusted t-tests for individual-mean comparisons for effects significant by ANOVA.

3.4 Results

3.4.1 Hemodynamics and Electrophysiology

Consistent with previous studies [54, 140], VTP-dogs showed increased left ventricular end-diastolic pressure and no significant AERP changes (Table 3-1). ATP-dogs were hemodynamically similar to sham controls, but as in previous work [53, 137, 59] showed substantial AERP decreases and loss of rate-adaptation at 1 week (Table 3-1). Atrial fibrillation duration increased progressively in both models, with changes reaching statistical significance at 2 weeks in VTP-dogs (Table 3-1) and at 6 weeks in ATP-dogs (Table 3-2). Although the ATP-induced AERP changes (decreased AERP and loss of AERP rate-adaptation) reached a maximum at 1 week (Table 3-1) and did not progress further at 6 weeks (Table 3-2), AF duration continued to increase between 1 and 6 weeks and statistically significant increases relative to baseline were achieved only at 6 weeks. There were no statistically significant differences between ATP-shams and VTP-shams (Table 3-2).

3.4.2 Microarray Findings

Figure 3-1 shows all mRNA-expression levels in the initial series of dogs, with the mean value for each transcript probeset plotted against mean VTP-sham expression. Values indicated by blue points are not significantly different from

Table 3–1: Hemodynamic and electrophysiological changes in the first series of dogs

	VTP-Sham	**24H VTP**	**2W VTP**	**24H ATP**	**1W ATP**
SBP, mm Hg					
Systolic	136 ± 7	109 ± 16	108 ± 15	133 ± 20	129 ± 10
Diastolic	90 ± 5	68 ± 9	70 ± 12	75 ± 14	77 ± 8
LVP, mm Hg					
End-diastolic	1 ± 1	5 ± 2	16 ± 5**	2 ± 2	2 ± 2
AERP, ms					
BCL					
360	122 ± 11	119 ± 8	136 ± 12	113 ± 18	70 ± 5**
300	125 ± 10	117 ± 9	133 ± 13	120 ± 12	71 ± 5**
250	124 ± 13	113 ± 11	129 ± 12	119 ± 7	76 ± 7**
200	117 ± 7	108 ± 10	121 ± 13	112 ± 5	75 ± 7**
150	98 ± 7	96 ± 10	102 ± 7	98 ± 8	76 ± 9*
DAF, s	39 ± 25	300 ± 360	837 ± 436*	15 ± 8	339 ± 384

SBP = systemic blood pressure; LVP = left-ventricular pressure;
BCL = basic cycle length; DAF = AF duration
*P < 0.05 , **P < 0.01 vs VTP-Sham

Table 3–2: Hemodynamic and electrophysiological changes in the second series of dogs

	VTP-Sham	ATP-Sham	6W ATP
SBP, mm Hg			
Systolic	133 ± 14	134 ± 11	121 ± 3
Diastolic	85 ± 11	74 ± 7	65 ± 1*
LVP, mm Hg			
End-diastolic	4 ± 2	4 ± 4	1 ± 2
AERP, ms			
BCL			
360	124 ± 4	120 ± 2	86 ± 12**
300	120 ± 3	121 ± 3	88 ± 10**
250	124 ± 13	113 ± 11	87 ± 9**
200	127 ± 10	121 ± 7	88 ± 9**
150	108 ± 7	102 ± 6	83 ± 8*
DAF, s	87 ± 39	21 ± 6	997 ± 324**
SBP = systemic blood pressure; LVP = left-ventricular pressure; BCL = basic cycle length; DAF = AF duration *P < 0.05 , **P < 0.01 vs VTP-Sham			

sham-values, whereas red points indicate statistically significant changes. In ATP-dogs, 242 probesets showed significant changes at 24 hours and 87 at 1 week. VTP-dogs showed significant changes for 2,209 probesets at 24 hours and 2,720 at 2 weeks. Corresponding results for the second series of dogs are shown in Figure 2. No statistically significant gene-expression differences were observed between 6-week ATP-dogs and VTP-shams, nor between ATP-shams and VTP-shams. Almost all (94%) of the significantly- changed genes in ATP-dogs lie below the black line of identity and are therefore under-expressed. For VTP-dogs, large numbers of genes lie on either side of the line of identity (53% underexpressed; 47% overexpressed).

Of the differentially-expressed genes in ATP-dogs, 55 of the downregulated and none of the upregulated values were common to both 24-hour and 1-week

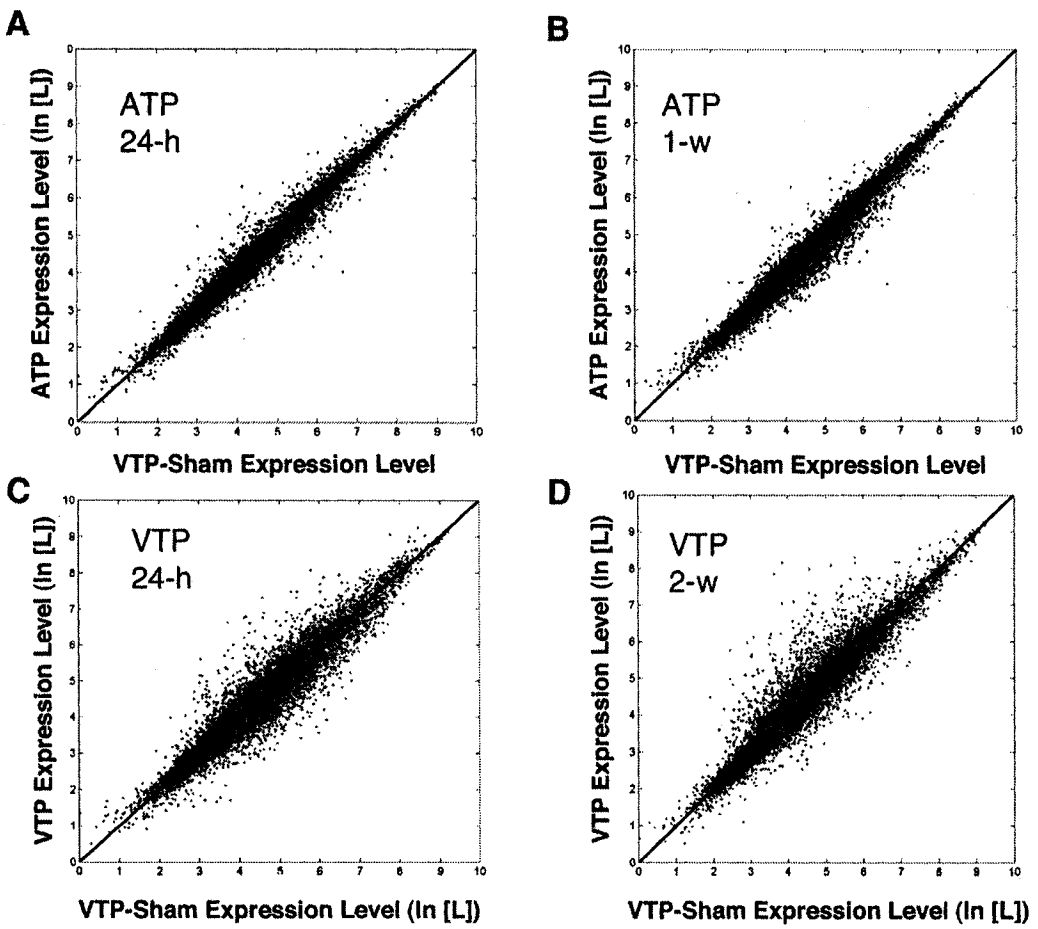


Figure 3–1: Overall changes in mRNA expression compared with sham in 24-hour ATP (**A**), 1-week ATP (**B**), 24-hour VTP (**C**) and 2-week VTP (**D**) samples. The absolute expression, in $\log_e(L)$, where L = mean sample-luminescence, for each intervention-group transcript probeset is plotted as a function of the corresponding value for the VTP-sham group and represented as a single point. Blue points are not significantly different from sham; red points are significantly different. If no changes in mRNA-expression occurred, all points would fall on the black line of identity.

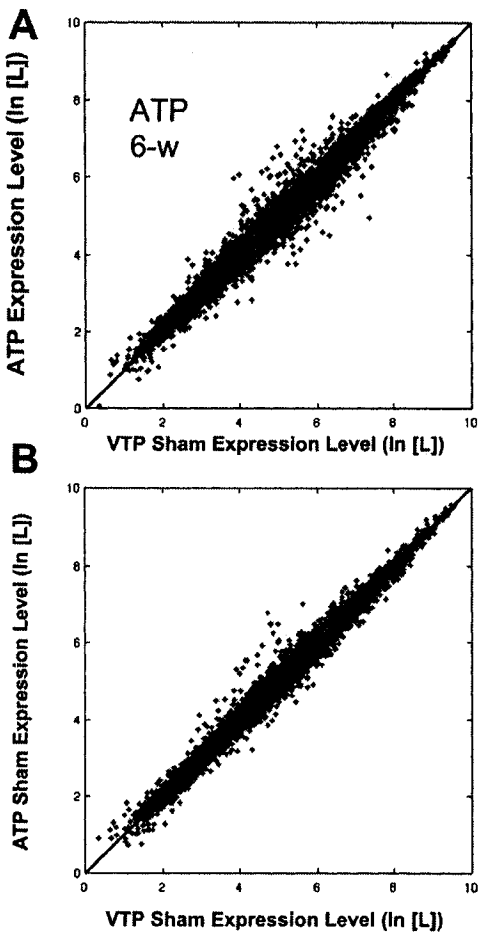


Figure 3–2: Overall mRNA-expression comparisons between (**A**) 6-week ATP vs VTP-sham (**B**), ATP-sham vs VTP-sham. Format as in Figure 3-1.

time-points. These 55 downregulated values make up more than 72% of the downregulated results at 1-week ATP; thus, most of the differentially-expressed genes at 1 week are also downregulated at 24 hours. The ATP-pattern indicates a stronger early response, with decreasing numbers of significantly-altered genes over time. Unlike ATP, the VTP model had 25% more significantly-altered expression values at the later time-point. There are 336 upregulated and 567 downregulated transcript probesets common to both time-points, constituting 41% and 33% of the significantly-altered 24-hour and 2-week values respectively. Thus, more than half of the genes differentially-expressed in VTP at each time point are unique to that time point. For the transcript probesets common to both time points, a statistically significant fraction (61%) was less altered at the 2-week time point.

Figure 3-3 shows the functional categories of genes that are significantly up- and downregulated in ATP and VTP-dogs. In ATP-dogs, the categories with the most genes altered are DNA/RNA synthesis/degradation and signal transduction. Almost all significantly-changed genes were downregulated. For all but the ribosomal-gene category, more genes were changed at 24 hours (black bars) than 1 week (white bars). The gray bars, representing genes changed at both 24 hours and one week, show that for most groups all genes changed at 1 week were also significantly altered at 24 hours. Several gene-categories (apoptosis, extracellular matrix (ECM), and transport) have no representation at 1 week. In contrast, for VTP-dogs most of the functional groups show unique 24- hour and 2-week responses, with overlap representing less than half the total response. Approximately as many values in each group represent upregulation as downregulation.

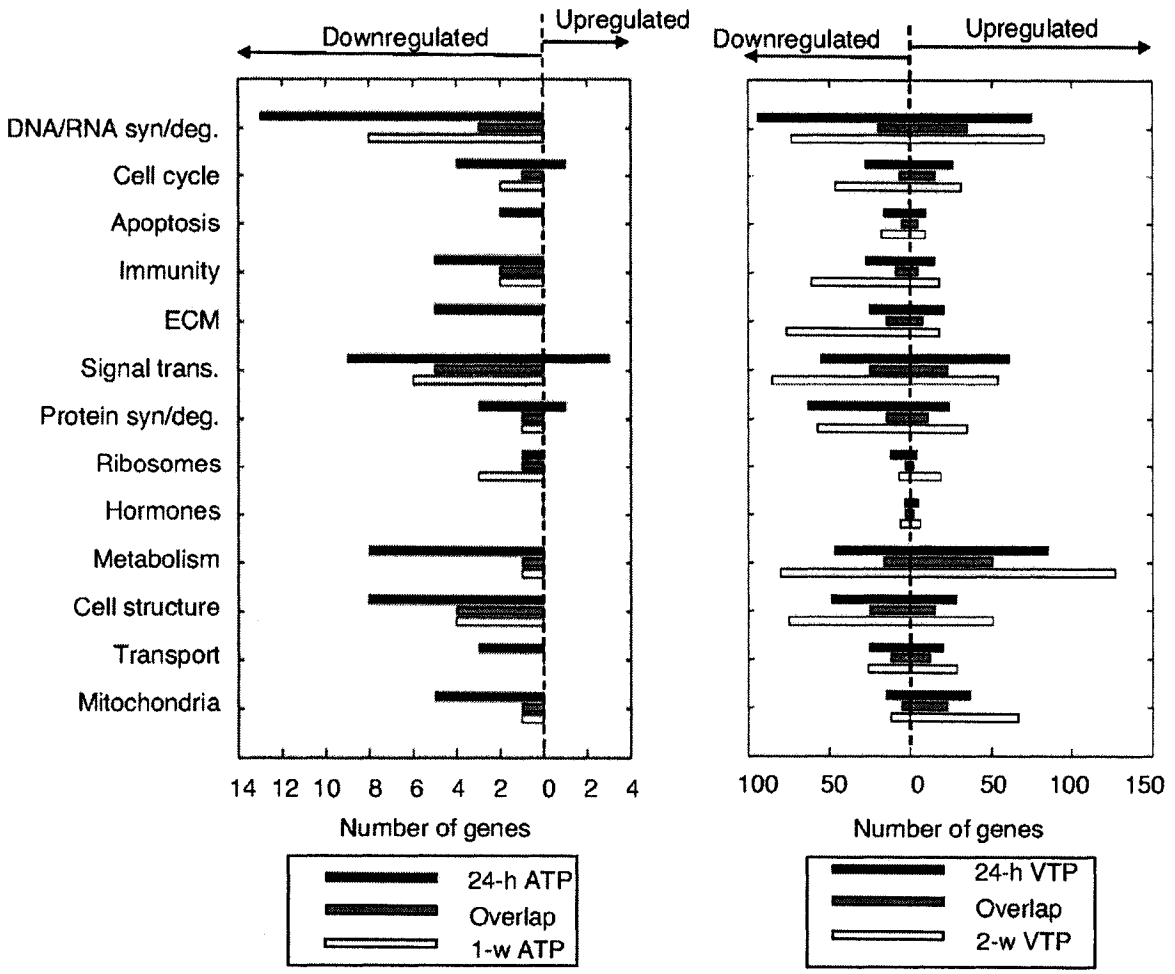


Figure 3–3: Number of transcript probesets in each functional group which was significantly down- or upregulated by atrial-tachycardia remodeling (left) or ventricular-tachycardia remodeling (right). "Overlap" refers to the number of values that were significantly affected in the same direction at both 24 hours and 1 week.

To evaluate whether gene-groups respond uniformly to each intervention or whether specific gene-groups change differentially, we calculated the percentage-change relative to sham in each dog for each significantly-altered transcript probeset and ranked all changes from largest to smallest, with 1 being the most-changed gene and the highest rank-number the least-changed gene. We then plotted for each gene-group the fraction of its ranks that fell within each cohort of genes as ranks increased by integer from 1 (the most changed expression-value) to the least-changed value. To assess statistical significance, we calculated the sum of ranks for each functional gene-group and then randomly reassigned expression ratios to groups. For each permutation, we calculated the sum of ranks for the random groups and compared them to the originals. This process was repeated 100,000 times, and those gene-groups whose sum of ranks were lower/higher than the random groups more than 97.5% of the time (2-tailed $P \leq .05$) were considered to have significantly larger/smaller expression-changes compared with overall behavior. Results following average behavior are shown by black lines; groups deviating significantly from average are shown by blue or red lines for larger or smaller changes respectively (Figure 3-4). For 24-hour VTP-dogs (Figure 3-4A), genes associated with metabolism, ECM, and cell structure/ mobility showed the largest changes. At 2-week VTP (Figure 3-4B), the same functional categories show larger-than-average changes, along with immunity/coagulation genes. At both time points, ribosome-associated genes showed smaller-than-average changes, but DNA/RNA synthesis/degradation genes only deviated from average at 2 weeks. The analyses are less clear for ATP-dogs (Figure 3-5), because of the much smaller number of significantly-changed genes. At 24 hours, immunity/coagulation and metabolism genes occupy significantly higher ranks than other groups, whereas at 1 week no groups deviate significantly from average responses.

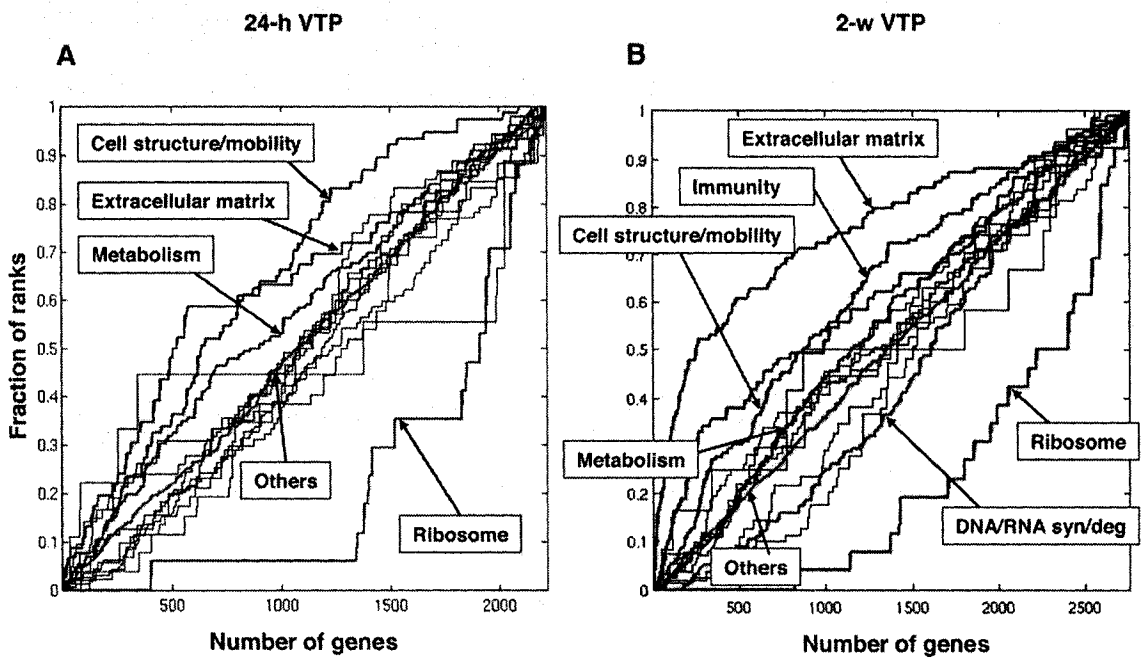


Figure 3–4: Cumulative fraction of ranks for each of the functional gene groups in VTP-dogs (for discussion of method, see text). Functional group curves in blue showed expression-ranks significantly greater than that of overall genome indicating larger-than-average changes; functional groups in red showed expression-ranks that were significantly less, indicating smaller-than-average changes.

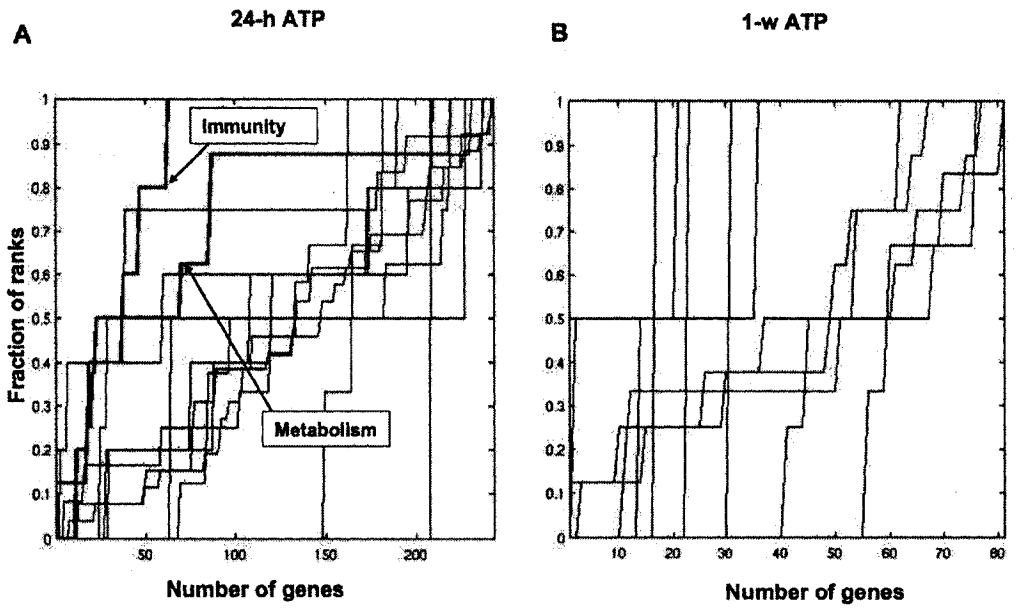


Figure 3–5: Cumulative fraction of ranks for each of the functional gene groups in ATP dogs (for discussion of method, see Results). Functional group curves in blue showed expression ranks significantly greater than those of overall genome; functional groups in red showed expression ranks that were significantly less.

3.4.3 Real-time RT-PCR Results

Figure 3-6 compares the expression levels of selected genes as determined by microarray and real-time RT-PCR methods in the first series of dogs. The genes were selected to include genes believed to be of pathophysiological significance, as well as genes with overexpression, underexpression and no apparent expression change. Overall, there was a strong linear correlation between results with the 2 methods (R=0.96). Because most of the values were concentrated within the 0 to 5-fold change range, this section of the graph is expanded at the left for better resolution. Detailed results are presented in Tables 3-3 and 3-4. There is generally close agreement between the independent determinations of mRNA expression

by the 2 methods. Of 36 sample-sets showing statistically significant changes by microarray, 22 (61%) show statistically significant changes of the same order and direction by RT-PCR. For the 14 sample-sets with significant changes by microarray and nonsignificant changes by RT-PCR, 11 (79%) show changes of the same direction and order with both methods.

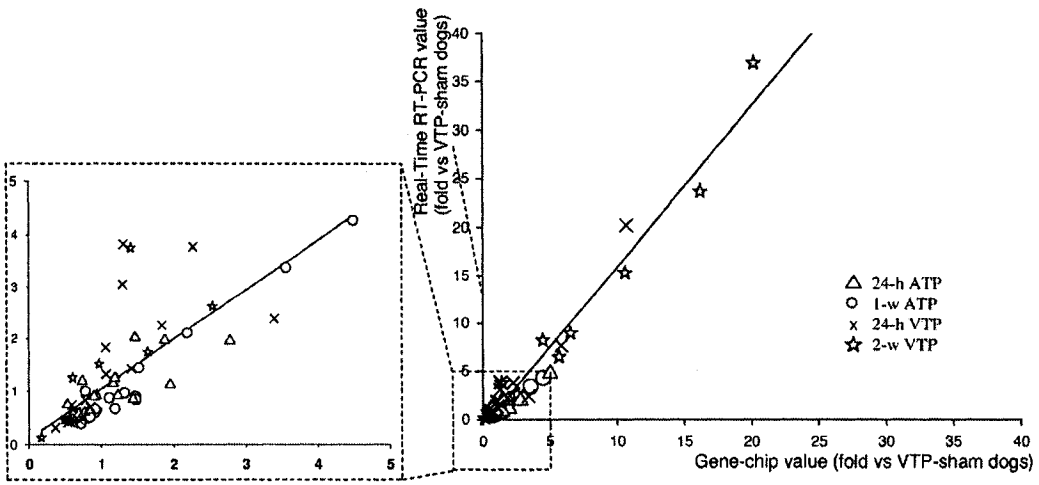


Figure 3–6: Correlation between expression changes by real-time PCR vs microarray. Regression lines are shown.

3.4.4 Western-blot Results

Figure 3-7 shows typical blots for the 9 gene products selected for Western-blot analysis. Table 3-5 presents the mean results of Western-blot analyses, along with the corresponding results from microarray analysis. Statistical congruence was observed for 28 of 36 sample sets: 24 sample-sets did not change significantly by either gene-chip or Western blot, and 4 sample-sets changed significantly in the same direction. For 5 of the remaining 8 sample-sets, changes were statistically significant for only 1 of gene-chip or Western blot, but were in the same

Table 3–3: RT-PCR confirmation for selected Affymetrix probesets: VTP

Gene (Affymetrix id)	24 hr VTP		2 wk VTP	
	RT-PCR	Microarray	RT-PCR	Microarray
Smooth muscle gamma actin (1586210)	2.39 ± 0.44	3.39 ± 0.34**	6.46 ± 3.60	5.66 ± 1.16
Skeletal myosin light chain 2 (1583107)	0.50 ± 0.03**	0.66 ± 0.11	0.51 ± 0.05**	0.66 ± 0.14**
Tissue inhibitor of metalloproteases 1 (1582383)	7.62 ± 2.00	5.85 ± 0.66**	8.91 ± 4.00*	6.53 ± 0.63*
Matrix metallopeptidase 2 (1582764)	1.33 ± 0.31	1.07 ± 0.38	8.17 ± 1.38**	4.46 ± 0.17**
Fibronectin (1582768)	20.19 ± 7.15*	10.70 ± 0.49**	23.67 ± 7.68*	16.20 ± 0.36**
Collagen alpha 1(III) chain precursor (1591762)	3.77 ± 0.98	2.27 ± 0.60	36.33 ± 7.83**	20.18 ± 0.17**
Cell cycle related kinase (1582412)	0.41 ± 0.06**	0.53 ± 0.20**	0.60 ± 0.09*	0.70 ± 0.15**
p53 (1582452)	1.83 ± 0.32	1.06 ± 0.10	1.52 ± 0.24	0.97 ± 0.11
Cathepsin L (1583225)	0.75 ± 0.16	0.59 ± 0.10**	1.26 ± 0.21	1.21 ± 0.12*
Von Willebrand factor (1582505)	1.43 ± 0.02	1.42 ± 0.38	2.63 ± 0.35**	2.54 ± 0.16**
Complement component C6 precursor (1585839)	3.05 ± 1.00	1.30 ± 0.39	15.27 ± 2.00**	10.59 ± 0.45**
PPAR gamma coactivator-1 (1596366)	0.47 ± 0.09*	0.53 ± 0.14*	0.44 ± 0.06**	0.57 ± 0.14*
Cytochrome P450c21 (1582564)	0.83 ± 0.14	0.58 ± 0.08**	1.25 ± 0.36	0.60 ± 0.21**
Isocitrate dehydrogenase subunit γ (1583519)	0.71 ± 0.06	0.85 ± 0.26	0.47 ± 0.07**	0.49 ± 0.18**
Clathrin heavy chain (1593830)	3.82 ± 1.17*	1.31 ± 0.22	3.76 ± 0.67*	1.41 ± 0.04**
KChIP2 (1582775)	0.31 ± 0.07**	0.36 ± 0.37**	0.13 ± 0.03**	0.17 ± 0.34**
Kv4.3 (1583046)	0.45 ± 0.07	0.70 ± 0.15**	0.66 ± 0.17	0.60 ± 0.09**
Cytochrome c oxidase subunit VIa (1583218)	2.26 ± 0.43*	1.84 ± 0.18**	1.75 ± 0.14	1.64 ± 0.06**

*$P < 0.05$, **$P < 0.01$, VTP = ventricular tachypacing model

Table 3–4: RT-PCR confirmation for selected Affymetrix probesets: ATP

Gene (Affymetrix id)	24 hr ATP		1 wk ATP	
	RT-PCR	Microarray	RT-PCR	Microarray
Smooth muscle gamma actin (1586210)	1.12 ± 0.21	1.95 ± 0.36	0.91 ± 0.23	1.47 ± 0.22
Skeletal myosin light chain 2 (1583107)	0.57 ± 0.07**	0.61 ± 0.06**	0.42 ± 0.05**	0.66 ± 0.15**
Tissue inhibitor of metalloproteases 1 (1582383)	1.97 ± 1.00	1.88 ± 0.51	0.85 ± 0.10	1.48 ± 0.48
Matrix metallopeptidase 2 (1582764)	0.92 ± 0.19	1.23 ± 0.34	1.45 ± 0.60	1.52 ± 0.40
Fibronectin (1582768)	4.75 ± 2.14	5.01 ± 0.60	3.38 ± 1.94	3.56 ± 0.76
Collagen alpha 1(III) chain precursor (1591762)	1.96 ± 1.11	2.78 ± 0.76	4.27 ± 1.16	4.49 ± 0.74
Cell cycle related kinase (1582412)	0.50 ± 0.32**	0.66 ± 0.19*	0.50 ± 0.08**	0.83 ± 0.11
p53 (1582452)	0.93 ± 0.18	0.93 ± 0.07	0.64 ± 0.10	0.93 ± 0.13
Cathepsin L (1583225)	0.60 ± 0.12	0.77 ± 0.14*	0.39 ± 0.09*	0.72 ± 0.12*
Von Willebrand factor (1582505)	1.15 ± 0.15	1.17 ± 0.13	0.97 ± 0.17	1.32 ± 0.06
Complement component C6 precursor (1585839)	2.02 ± 0.83	1.47 ± 0.50	2.11 ± 1.02	2.19 ± 0.90
PPAR gamma coactivator-1 (1596366)	0.64 ± 0.11	0.83 ± 0.29	0.60 ± 0.15	0.91 ± 0.24
Cytochrome P450c21 (1582564)	0.75 ± 0.10	0.53 ± 0.14**	0.46 ± 0.07	0.57 ± 0.19**
Isocitrate dehydrogenase subunit γ (1583519)	0.91 ± 0.17	0.90 ± 0.17	0.67 ± 0.11	0.92 ± 0.06
Clathrin heavy chain (1593830)	1.25 ± 0.33	1.19 ± 0.16	0.88 ± 0.26	1.11 ± 0.12
KChIP2 (1582775)	0.57 ± 0.17*	0.68 ± 0.26	0.52 ± 0.11*	0.85 ± 0.34
Kv4.3 (1583046)	1.19 ± 1.06	0.74 ± 0.09**	1.00 ± 0.14	0.78 ± 0.09*
Cytochrome c oxidase subunit VIa (1583218)	0.85 ± 0.21	1.44 ± 0.15*	0.67 ± 0.16	1.19 ± 0.19

*$P < 0.05$, **$P < 0.01$, ATP = atrial tachypacing model

quantitative direction. For 3 sample-sets (cathepsin L in 2-week VTP and cathepsin S in 24-hour and 2-week VTP dogs), statistically significant increases or decreases of the order of 30% to 50% were seen in one measurement and either directionally-discrepant or no change was observed with the other. There was thus good general agreement between changes in protein expression and changes in mRNA-expression, although some quantitative differences were clearly present (eg, for collagen-III, KChIP2 and IL1-RA).

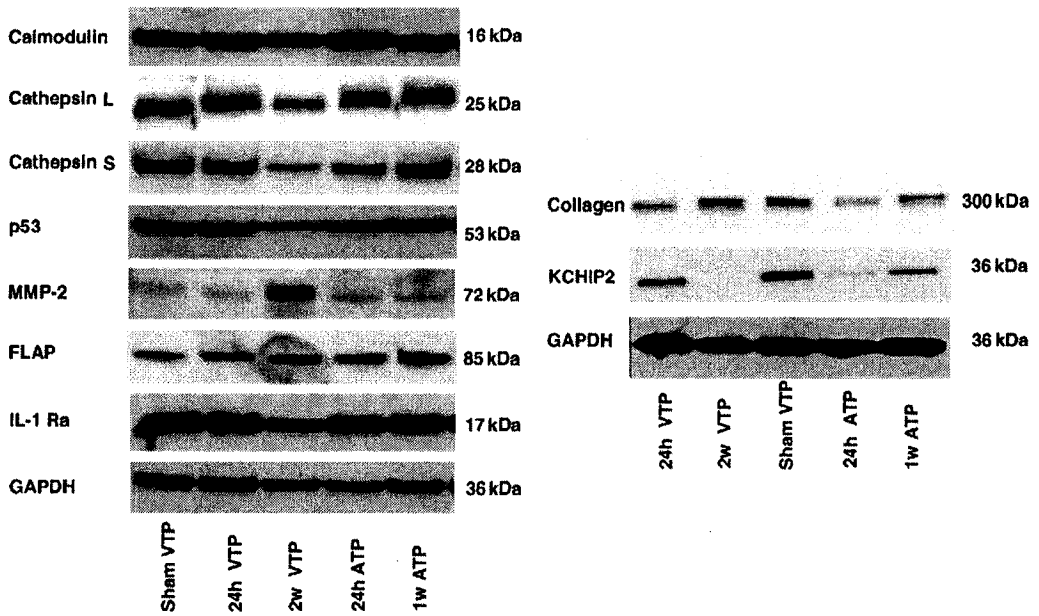


Figure 3–7: Examples of Western blots for 9 proteins studied to compare protein versus mRNA expression changes.

Table 3–5: Western confirmation for selected Affymetrix probesets

Protein	Gene ID	Group	Microarray	Western
Calmodulin	1583631	24h VTP	1.18	0.81 ± 0.08
		2w VTP	1.45	.56 ± 0.25
		24 ATP	1.15	1.02 ± 0.09
		1w ATP	1.56	1.15 ± 0.05
Cathepsin L	1583225	24h VTP	0.59*	0.89 ± 0.23
		2w VTP	1.21	0.63* ± 0.15
		24 ATP	0.77	0.93 ± 0.12
		1w ATP	0.72	0.95 ± 0.2
Cathepsin S	15996801	24h VTP	1.30*	1.01 ± 0.78
		2w VTP	1.49*	0.64 ± 1.10
		24 ATP	1.25	1.12 ± 0.34
		1w ATP	1.30	1.15 ± 0.17
p53	1582452	24h VTP	1.06	0.75 ± 0.15
		2w VTP	0.97	0.92 ± 0.29
		24 ATP	0.94	1.04 ± 0.35
		1w ATP	0.93	0.92 ± 0.28
Collagen III	1591762	24h VTP	2.27	0.98 ± 0.34
		2w VTP	20.18*	1.79** ± 0.29
		24 ATP	2.78	1.19 ± 0.37
		1w ATP	4.49	0.87 ± 0.27
Matrix metallopeptidase 2	1582764	24h VTP	1.07	1.39 ± 1.13
		2w VTP	4.46*	3.18** ± 2.40
		24 ATP	1.23	0.75 ± 0.17
		1w ATP	1.52	2.36 ± 0.77
FLAP	1601853	24h VTP	3.32*	1.29 ± 0.45
		2w VTP	4.70*	1.57 ± 1.45
		24 ATP	1.58	1.30 ± 0.08
		1w ATP	2.05	1.29 ± 0.64
Interleukin 1-RA	1582494	24h VTP	3.42	0.89 ± 0.14
		2w VTP	9.97	0.58 ± 0.18
		24 ATP	1.39	0.95 ± 0.20
		1w ATP	1.95	0.98 ± 0.27
KChip2	1582775	24h VTP	0.36*	0.24** ± 0.07
		2w VTP	0.17*	0.12** ± 0.01
		24 ATP	0.69	0.09** ± 0.01
		1w ATP	0.85	0.24** ± 0.03

VTP = ventricular tachypacing model, ATP = atrial tachypacing model,
*$P < 0.05$, **$P < 0.01$

3.5 Discussion

We have analyzed changes in canine atrial mRNA expression induced by atrial-tachycardia remodeling and ventricular-tachypacing induced heart failure over time. The results highlight major differences in the molecular basis of these two atrial arrhythmogenic-remodeling paradigms and indicate important time-dependent evolution of gene-expression changes.

3.5.1 Relationship to Previous Findings

Several gene-microarray studies have been performed in AF patients. Kim *et al.* found upregulation of pro-oxidant and downregulation of antioxidant genes [71]. Investigators subsequently found 33 genes with $> 50\%$ upregulation and 63 with $> 50\%$ downregulation [72, 141], with changes in genes related to cell signaling, inflammation, oxidation and cellular respiration [72]. Barth *et al.* found that the human atrial transcriptome changed to a ventricular-like pattern in AF-patients [69]. One limitation of these studies was the difference in heart disease between AF-patients (valve disease, often with cardiac hypertrophy, dilation and/or dysfunction) compared with sinus-rhythm controls (coronary-artery disease with well-preserved ventricular function). It is therefore difficult to separate changes because of AF from underlying disease-related remodeling. We recently analyzed AF-related transcriptome remodeling in heart-disease matched patients with AF versus sinus rhythm [68, 142]. Most gene-expression changes were attributable to underlying heart disease: $\sim 2/3$ of ion-channel gene-changes [142] and $>90\%$ in the complete transcriptome [68] occurred in both sinus-rhythm and AF patients.

Much less information is available from animal models of AF. Changes in cellular structure, metabolism, gene-expression regulation and differentiation genes

were observed in a goat model [147]. In a porcine atrial-tachypacing model, 387 genes were altered [70]. In neither model was ventricular-rate controlled, so a contribution of tachycardia-induced cardiomyopathy cannot be excluded.

Here, we used canine-specific microarrays to study changes with ~21,000 transcript probesets over time in ATP and VTP-dogs. Our results show striking differences in the quantity, magnitude and types of gene-expression changes induced by the two interventions. Whereas atrial-tachypaced dogs primarily showed decreasing numbers of transcript-expression changes over time, VTP-dogs displayed complex temporal evolution with some transcripts becoming less affected over time and others more affected. Particularly striking were changes in ECM-gene expression, with relatively small changes in most genes at 24 hours and very large changes at 2 weeks (eg, 8 collagen-genes upregulated >10-fold, fibrillin-1 8-fold and MMP2 4.5-fold). However, fibronectin was >10-fold and lysine oxidase-like (LOXL)-2 was ~5-fold upregulated at both 24-hour and 2-week time-points. ECM genes were virtually unchanged in ATP-dogs, with small (~20% to 30%) decreases in 3 collagen genes at 24 hours and no significant ECM-gene changes thereafter.

3.5.2 Relevance to Mechanisms of AF-related Remodeling

Ventricular tachypacing-induced CHF produces structural and ionic remodeling resembling the substrate for chronic AF in man [50, 54, 140]. Our results provide extensive new information about the nature and number of atrial gene-systems affected over time during the evolution of CHF. The prominence of changes in ECM genes, particularly those associated with collagen production, is consistent with the fibrosis that appears central to arrhythmogenesis [64, 67].

Collagen-associated mRNA expression is dramatically increased at 2-week VTP, when fibrosis approaches maximum, and is much less affected at 24 hours, when fibrosis has not yet appeared [66]. Early-phase reactive ECM genes, such as α1-antitrypsin and fibronectin (activity increased $\sim$5 and $\sim$10-fold at 24-hour VTP), may be involved in early changes leading to fibrosis. Altered regulation of genes involved in metabolism and cellular contraction are consistent with energy-saving adaptations.

To pursue the gene-response analysis, we considered genes whose protein-products interact with TGFβ, in view of evidence for a potentially-important role of TGFβ in AF-related fibrotic remodeling [65, 148]. We identified all proteins in the Human Protein Reference Database [149] that interact with TGFβ or a TGFβ-interacting protein. For 383 proteins identified, we used BLAST [145, 146] to find corresponding probesets on our microarray. Of 214 genes whose expression could be measured, 85 were differentially expressed in VTP-dogs (Fig. 3-8). Interestingly, connective-tissue growth factor (CTGF) gene-expression was enhanced by 24-hour VTP. CTGF is upregulated by angiotensin 2[150], TGFβ 1[151], or alterations in the cytoskeleton [152]. CTGF promotes fibrosis in pathological conditions by blocking a negative TGFβ-feedback loop mediated by Smad 7 signaling, allowing continued TGFβ-related activation [153]. The addition of CTGF to primary mesangial cells induces fibronectin production, cell migration, and cytoskeletal rearrangement [154]. Fibronectin expression is enhanced by 24-hour VTP and fibronectin interacts with a wide range of TGFβ-related products of genes altered in VTP (Fig. 3-8). Thus, CTGF is an interesting potential candidate for a significant role in VTP-related remodeling. Further exploration of this and other networks identified by the rich genomic data obtained in the present work is indicated, but goes beyond the scope of this study. The very large number of genes

affected by VTP points out the potential pitfalls in assessing a small number of selected genes without understanding the gene-response background against which such changes occur.

In contrast to VTP-remodeling, ATP-remodeling is associated with preserved tissue architecture and a predominance of ionic remodeling [54, 138, 139]. The virtual absence of ECM-gene changes in the ATP data set is consistent with this structurally-benign remodeling. We observed relatively limited changes in ion-channel genes, despite the importance of ion-channel alterations in ATP-induced AF promotion [50, 138, 139]. This may reflect important post-transcriptional mechanisms [57, 155], but likely also reflects relatively low-level expression of ion-channel genes that makes it difficult to differentiate ion-channel gene-expression changes from background noise in a pan-genomic microarray. To assess ion-channel subunit mRNA changes accurately requires specialized microarrays [142] that are presently unavailable for the dog. The decreasing number of gene-expression changes that occurred over time with ATP-remodeling suggests a time-related reduction in the stimuli for gene-expression change, consistent with atrial adaptation to the stress of ATP.

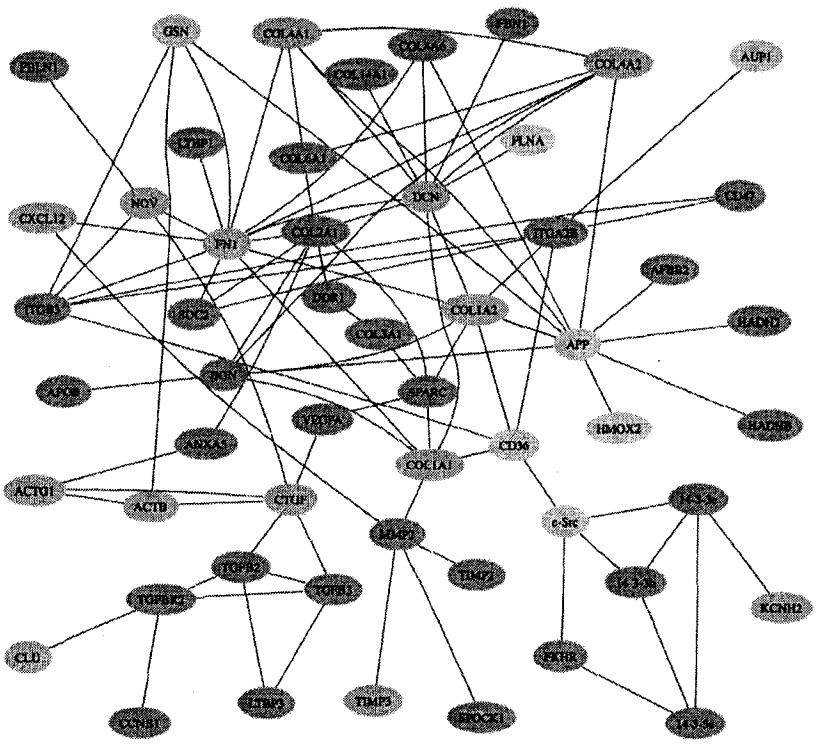
3.5.3 Potential Significance

Clinical gene-microarray studies in AF are limited by a range of factors (eg, underlying cardiac disease, drug therapy, duration of AF, etc) varying within the population. Animal models permit the assessment of transcriptomal changes under controlled duration and nature of atrial-remodeling stimuli. The present study is the first assessment of mRNA-expression remodeling in animal AF models

to use species-specific microarrays. In addition, it is the first to compare atrial-transcriptome remodeling because of atrial tachycardia per se with remodeling caused by an AF-promoting cardiac condition (CHF) and the first to study the evolution of gene-changes over time. Our results illustrate the importance of considering underlying disease-related gene-expression changes in AF populations. They also clearly contrast the relatively modest mRNA-expression changes caused by atrial tachycardia with the extensive alterations induced by CHF.

3.5.4 Potential Limitations

We selected analysis time-points based on evidence that early-phase VTP-induced atrial changes peak at 24 hours[66, 65], important atrial electrophysiological remodeling occurs with ATP at 1 week [138, 152] and with VTP at 2 weeks [65, 66]. Time-points additional to the ones we used might be interesting to examine. We do not know whether cardiac mRNA-level alterations were because of changes in synthesis or degradation. The studies necessary to resolve this question are not presently feasible at the scale that would be needed for the large number of genes we studied. We used an isolated-cardiomyocyte preparation for which we previously found very little contamination by other cell-types, but we cannot totally exclude contributions from noncardiomyocyte cell populations.



DCN	Decorin
CTGF	Connective tissue growth factor
FN1	Fibronectin 1
CLU	Clusterin
TIMP3	Tissue inhibitor of metalloproteinase 3
ACTB	Actin beta
ACTG1	Actin gamma 1
COL4A2	Collagen 4 alpha 2
COL1A1	Collagen 1 alpha 1
CXCL12	Chemokine ligand 12
KCNH2	Potassium channel, voltage-gated subfamily H member 2
GSN	Gelsolin
APP	Amyloid beta A4 protein
HMOX2	Heme oxygenase 2
AUP1	Ancient ubiquitous protein 1
FLNA	Filamin
COL1A2	Collagen 1 alpha 2
COL2A1	Collagen 2 alpha 1
DDR1	Discoidin domain receptor
HADHB	Hydrocyacyl dehydrogenase unit B
ITGA2B	Integrin alpha 2B
ITGB3	Integrin beta 3
VEGFA	Vascular endothelial growth factor A
TGFB2	Transforming growth factor beta 2
TGFBR2	Transforming growth factor receptor 2
TGFB3	Transforming growth factor beta 3
LTBP3	Latent TGF binding protein 3
FBN1	Fibulin 1
COL6A1	Collagen 6 alpha 1
HADH2	Hydroxyacyl CoA dehydrogenase 2
TIMP2	Tissue inhibitor of metalloproteinase 2
FKHR	Forkhead box O1A
APOB	Apolipoprotein B
14-3-3b	14-3-3 beta
14-3-3e	14-3-3 epsilon
14-3-3z	14-3-3 zeta
AFBB2	Amyloid beta A4 protein
CCNB1	Cyclin B1
FBLN1	Fibrillin
SDC2	Syndecan 2
SPARC	Osteonectin
COL4A6	Collagen 4 alpha 6
COL14A1	Collagen 14 alpha 1
LTBP1	Latent TGF binding protein 1
MMP2	Matrix metalloprotease 2
BGN	Biglycan
COL3A1	Collagen 3 alpha 1
ANXA5	Annexin 5

Figure 3–8: Proteins with differential mRNA expression in VTP dogs, that interact with TGFβ or a TGFβ-interacting protein, based on the Human Protein Reference Database. Nodes are color-coded to indicate time-points of differential expression: green is 24 hours only, red is 2 weeks only, and blue is both. For this representation, in cases where more than one probe-set was available for an individual transcript, we considered there to be a significant change if statistically significant alterations were detected by at least one probe-set. Connecting lines denote an interaction indicated in the Human Protein Reference Database.

CHAPTER 4
Problems and Properties

The microarray data analysis from Chapter 3 raised a number of issues with the current techniques of microarray analysis. These problems sparked an investigation of the unprocessed probe intensity values to better understand their properties. The results led to an alternative method for microarray analysis that effectively solved the encountered problems and is presented in Chapter 5. This section provides the background information and motivation for the new algorithm. Along with the material in Chapter 6, this may appear in a future manuscript extending the work of Chapter 5.

4.1 Problems

My analysis of microarray data raised many issues. While I offer some specific examples of problems, I found these same issues present when I analyzed other microarray data sets. Although less expensive and redesigned microarrays in combination with more extensive genomic annotations may circumvent these problems in the future, currently they are important concerns that can influence the results obtained and the conclusions drawn.

4.1.1 Problem 1: Lack of Clarity

As discussed in Chapter 1, microarray analysis is not standardized. It is a multi-step process with competing algorithms at each step and little agreement about which algorithm to use or, indeed, how many steps exist in the process. Analyzing the canine microarray data using different algorithms returned different results. Instead of using a combination of MBEI and SAM, we used RMA and SAM and found that there was agreement in only 50.9 % of the probesets selected in 2 week VTP. Also, we collaborated with a research group experienced in cDNA microarray analysis and discussed the analysis procedure. They introduced a scaling step before applying SAM as well as an outlier removal system in addition to the one used in MBEI. The ensuing discussion revealed we could not prove or disprove that their combination of methods improved accuracy without impairing sensitivity. This is both because there is a deficit in validation data sets and because many of the available algorithms, such as quantile normalization, affect the data in nonlinear ways that obscure the effects. Furthermore, even if a group of algorithms are agreed upon, algorithms have options that can produce different results. In the case of MBEI, as implemented in dChip [29], the user can choose

between a PM-only model or a PM-MM model, a log transform or not, background correction or not, and the removal of up to three different types of outliers [29]. All of these choices have underlying justifications, and all alter the probesets selected as differentially expressed [35]. Thus, the final results depend heavily on the preferences of the particular people analyzing the data.

4.1.2 Problem 2: Probeset Annotations

Even after finding which probesets are differentially expressed, there is still the problem of determining to which genes the probesets correspond. Affymetrix provides annotation files which contain information such as the gene title and transcript identification, but at the time of the analysis of Chapter 3, 96% of the probesets were without a gene title. The current update of the annotation file still has 46% of all probesets with an unknown target gene. While Affymetrix designs probesets to bind non-gene coding transcripts, we found that some of the unknown probeset targets could be identified using the genetic sequence alignment tool BLAST [145, 146] to find similar human or canine sequences. We used other transcript identifiers to determine the gene but with little success due to the canine genome's lack of annotation. Large scale use of BLAST, however, introduces new problems to the annotation process.

For a typical probeset target sequence, BLAST returns a list of alignments below a certain expectation value, akin to a p-value. Since the target sequences vary in size, a perfect alignment for short sequences can return the same expectation value as a suboptimal alignment for longer sequences. Gaps add extra complications because a gap in an area of the target sequence where probes do not bind is not as detrimental as a gap in an area probes specifically target. Another problem

is that aligned sequences can have different names but represent the same gene. Gene names depend on species and the context in which they were studied, so it is not uncommon to find a gene with ten different names. Finally, the BLAST alignment might return similar genes but different isoforms or transcript variants. It is hard to discern whether the probes of the probeset actually target a sequence that is different in the two transcript variants or not. Due to all of these confounding factors, the process of matching probesets with the best annotation based on the BLAST alignment needs human supervision. This not only introduces human error and bias but limits the number of probesets that can be annotated in a given time.

4.1.3 Problem 3: Probesets in Excess and Dearth

The complications with annotating probesets also limits knowing exactly which genes the microarray measured. Because humans need to supervise the annotation process and there are usually several times more genes found unchanged than changed, there is a lack of annotation information for the unchanged probesets. When researchers do not see certain genes on the list of those differentially expressed, it is unclear as to whether those genes were measured and found unchanged or if there were no corresponding probes on the chip. Thus, when researchers try to understand the genetic mechanisms which govern their experiment, they cannot distinguish whether pathways act independently of a certain gene or not.

After annotating the probesets on the Canine microarray, we found that multiple probesets shared the same annotation. For example, there are at least four probesets annotated as fibrillin 1: 1584905_at, 1589076_at, 1595549_at, and 1595550_at. In the 2 week VTP group, three of the four probesets were found

upregulated with only 1595550_at being the exception. To complicate matters, the original annotations listed only 1595549_at, and 1595550_at as fibrillin 1, while the other two had no definition. Without doing any sequence analysis to identify probesets with no definition, we would have been left with one of two probesets claiming fibrillin 1 changed. Currently, there is no way to determine whether one probeset was a false positive or the other a false negative without doing an experiment. Even with the more extensive annotations, we are still left with conflicting reports as to whether or not fibrillin 1 changed. This poses problems for other probesets found to be changed because there is a dearth of information concerning how many different probesets each gene has.

4.2 Properties

We investigated the properties of probes in a controlled experiment where we knew exactly what was on the microarray. The Latin square data set [46] is a collection of 42 HGU-133A Affymetrix microarrays representing 14 experimental groups each with 3 replicates. Each microrray has 42 genes spiked in at known concentrations ranging from 0 to 512 picomolar, incremented by powers of two and following a Latin square experimental design. The data is publically available [46] and provided for algorithm testing.

4.2.1 Performance

Because of the Latin square setup, we have the intensity values of 498 probes over a range of 14 different concentrations. If the intensity distribution for each concentration is distinct then not only can the individual probes be used to detect

differential expression, but they can also be used to infer the concentration. The histograms, however, show that there is substantial overlap in probe intensity distributions for different concentrations such that an intensity of 1,000 could correspond to multiple concentrations.

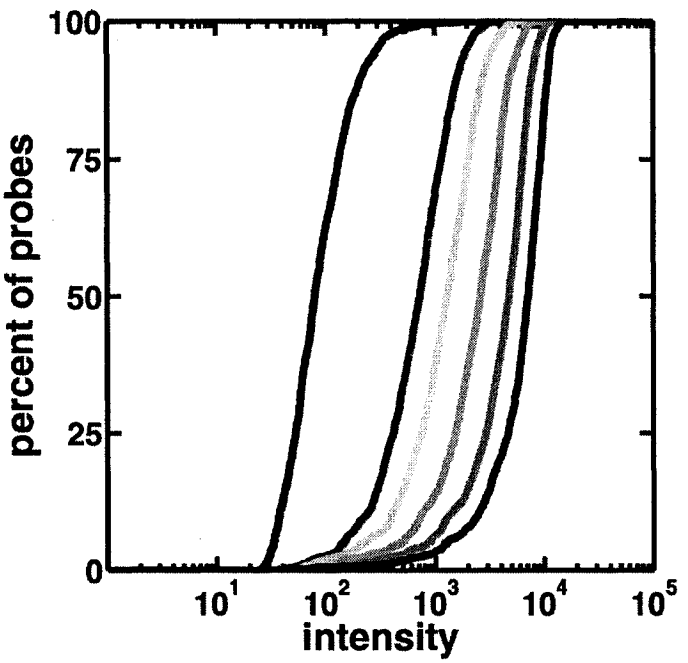


Figure 4–1: The cumulative distribution function for probes measuring genes at a concentration of 0 (black), 32 (red), 64 (yellow), 128 (cyan), 256 (green), and 512 (blue) picomolar.

The overlap in the intensity distributions does not report how individual probes react to increasing concentrations. It is possible that a population of probes do not respond to increasing concentration and the observed overlap is caused by their fluctuating intensity values. We tracked the intensities of individual probes as concentration increased and recorded the percentage that increased. For the lowest

concentrations around .125 picomolar, just over 50 percent of the probes increased in intensity with a two fold increase in concentration. Once the concentrations rise above 1 picomolar, the probes detected fold changes of 2 or greater over 85 percent of the time. Thus, a majority of the probes respond to increasing concentration and the overlap in intensities depends on the degree in which the probes increase.

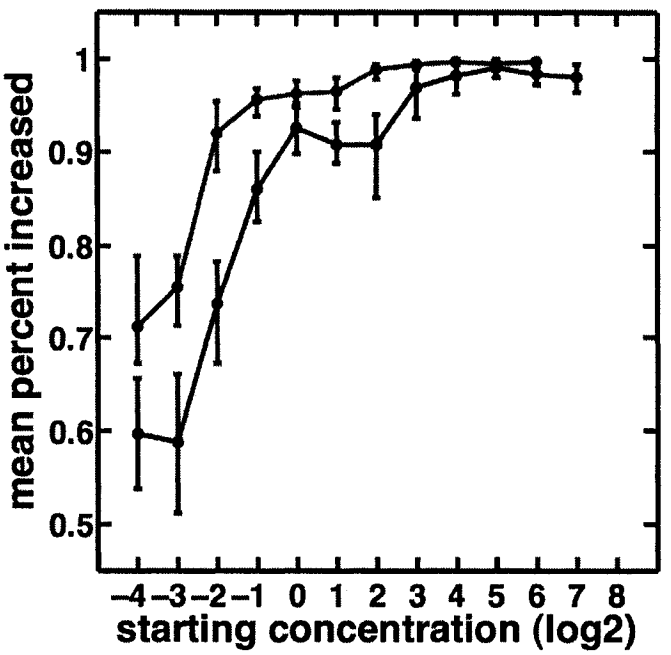


Figure 4–2: The percent of probes that increased in intensity between fold changes of 2 (blue) and 4 (red). Each point represents the percent of probes that increased using the average intensities of a set of replicates. The error bars show the best and worst possible percentage had the best or worst array been used to represent the intensities of a set of replicates.

4.2.2 Noise

Besides inter-probe variation there can also be variation caused by the microarrays [17, 29]. We wanted to determine the magnitude of the noise between replicates and compare it to the inter-probe variation. Figure 4-3 shows that although the variation between replicates increases with increasing concentration, the inter-probe variation is much greater. Even though the variation of individual probes between replicates might be low, there are reports of systematic scanning effects that make some microarrays have higher intensities than others. To observe this, we compared the intensity of all the probes not spiked-in between microarrays. If there were no systematic fluctuations most of the comparisons would result in 50 percent of the probes being higher with a small standard deviation. Without any normalization more than half of the comparisons resulted in 70 percent or more of the probes having higher intensity. With normalizations such as those reviewed in the Chapter 1, over 90 percent of the comparisons were below the 60 percent brightness bias. While quantile normalization [17] made the most improvement, the invariant set method used in MBEI [30] or a simple median scaling eliminated most of the bias, as well.

4.2.3 Normality

Although previous work has shown that expression indices are not normally distributed, we wondered if an individual probe measuring the same concentration has intensities that are normally distributed. To test for normality, we applied the Lilliefors test (`lillietest` in Matlab, The Mathworks, Massachusetts) to every probe in the Latin square setup. For the 22,000 probesets not spiked in we used

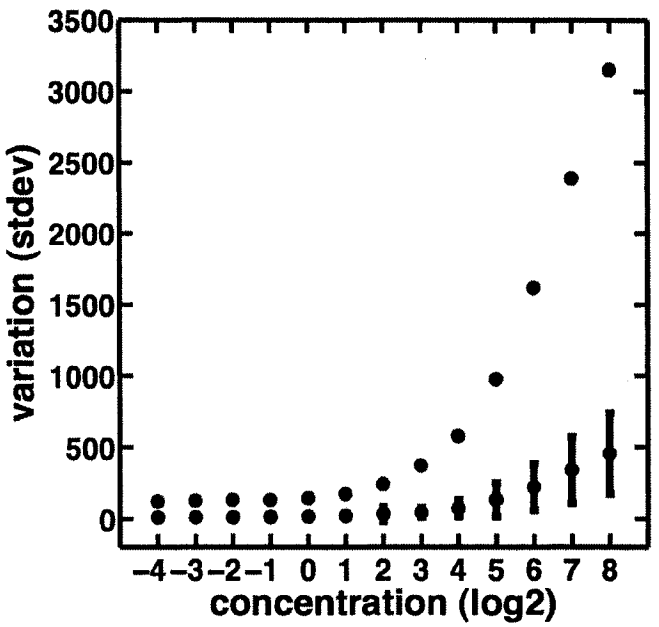


Figure 4–3: The variation in replicate intensities (blue) and probe intensities (red) versus concentration. We show the average standard deviation of probe intensities within replicates measuring the same concentration. The error bars correspond to the standard deviation of this variation. This stands in contrast to the standard deviation of all probe intensities measuring a particular concentration (red). While the replicate variation increases with respect to concentration, so does the mean probe intensity to the extent that the coefficient of variation is approximately constant at $\sim$.1.

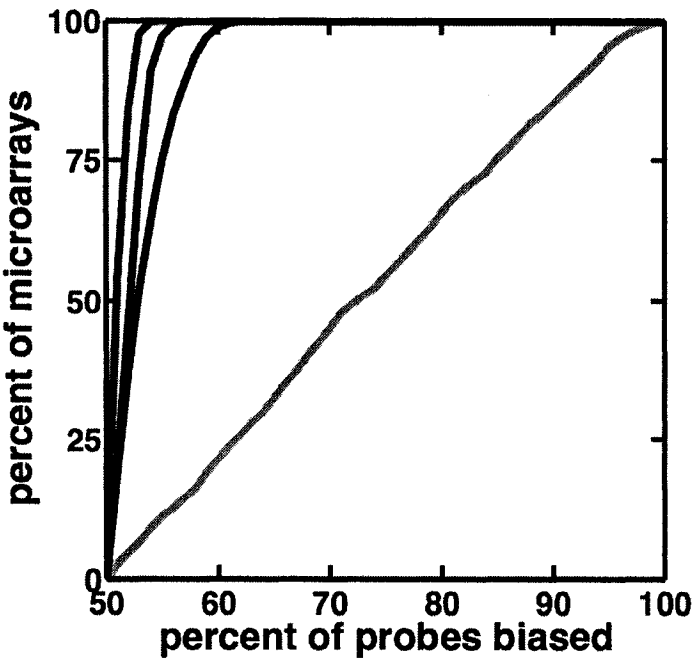


Figure 4–4: The cumulative distribution function of the number of comparisons resulting in more than half of the probes with increased intensities for different normalizations. We compared every chip to every other chip and calculated the percentage of probes measuring 0 picomolar that increased and decreased. We took whichever percent was larger, so if chip 1 had 40% of its probes larger than chip 2, we took a value of 60%, which is just equivalent to swapping chip 1 and chip 2 in the comparison. This removed the bias caused by ordering the microarrays in the comparison. The results with no normalization (green) show much higher scanning effect bias as compared to quantile normalization (red), invariant set normalization (blue), and median scaling (black).

their values for all 42 microarrays and found no evidence to reject the hypothesis that probe intensities are normally distributed. This subset represents probesets that should be measuring background which is not a good representative of a real microarray experiment. We performed the same test on every probe in each condition of the canine microarray data from Chapter 3 and found similar results. We also tested every probe in a microarray experiment with 12 replicates from the McGill University and Genome Quebec Innovation Center and found no evidence to contradict the assumption of normality.

CHAPTER 5
Evaluation Theory

We generalize the problem of microarray analysis to a broad question of how to combine the scores from judges evaluating different options. In this context, we find that the distribution of the judges' scores determines whether a sum rule or majority rule is better. We then apply this general result to microarray analysis and propose a new algorithm, which we compare with popular alternatives. We extend our analysis to consider the costs of an evaluation and how this influences the optimal number of judges.

5.1 Abstract

I am quite prepared to be told, with regard to the cases I have here proposed, 'Oh, that is an extreme case: it could never really happen!' Now I have observed that this answer is always given instantly, with perfect confidence, and without any examination of the details of the proposed case. It must therefore rest on some general principle: the mental process being probably something like this – 'I have formed a theory. This case contradicts my theory. Therefore this is an extreme case, and would never occur in practice.' C. L. Dodgson (1876) [156]

Evaluators must determine a relative ranking between two or more competing options or competitors in settings as diverse as: academic selection committees [156, 157], social policies [158], elections [156, 157, 158, 159, 160], figure skating competitions [161, 162], and gene expression in microarrays [14, 18, 29, 39]. Since evaluators are not perfect, it is common to combine several independent judgments to reach a final decision [163, 164, 165, 166]. In a sum rule method the numerical ratings of several independent judges are added together, but such a method gives a single outlying judge disproportionate weight. Alternatively, in a majority rule procedure the option favoured by the most judges wins. For situations involving more than two competitors, however, majority rule algorithms may lead to paradoxes in which the ranking has cycles rather than a clear ordering [156, 157, 158, 159]. Here, we show that the accuracy of evaluations depends on both the number of judges and the distribution of their scores: if the mean values of the judges' scores are tightly distributed then a sum rule is best, but if they are broadly distributed a majority rule is preferred. By using a cost-benefit analysis,

we can determine the optimal number of judges and the value placed on decisions. We apply these results to derive an alternative algorithm for evaluating gene expression using microarrays and show that it outperforms those most commonly used [14, 18, 29, 39]. These results can be used to determine optimal judging procedures in a wide range of circumstances.

5.2 Figure Skating and the Voting Paradox

Figure skating, with its many competitors and judges, serves as a paradigmatic example of the challenges in evaluation. Well-publicized controversies led the International Skating Union to switch from a majority rule based ranking in 2002 to a new system based on the sums of judges' scores in 2006 [161, 162]. To illustrate the problems, we show the score-sheets for two competitors in the 2006 US Junior Figure Skating Championship [167] (Fig. 5-1a). Although six of the nine judges preferred Competitor LD, Competitor KW had a higher average score and was ranked above Competitor LD. The discrepancy arises because the sum rule permits evaluators who give larger differences in their scores between options to have more power in determining the final result.

When there are more than two competitors, other complications can surface. In the hypothetical example shown in Fig. 5-1b, the sum rule cannot produce a ranking because the average score of each competitor is the same. Had any judge scored one competitor a fraction of a point higher, that competitor would have been the winner. Applying the majority rule, the scores of each judge can be converted into a rank ordering from 1 to 5, where 1 is highest. Here, the ranking matrix is a Latin square, i.e. no two judges agree about the ranking of

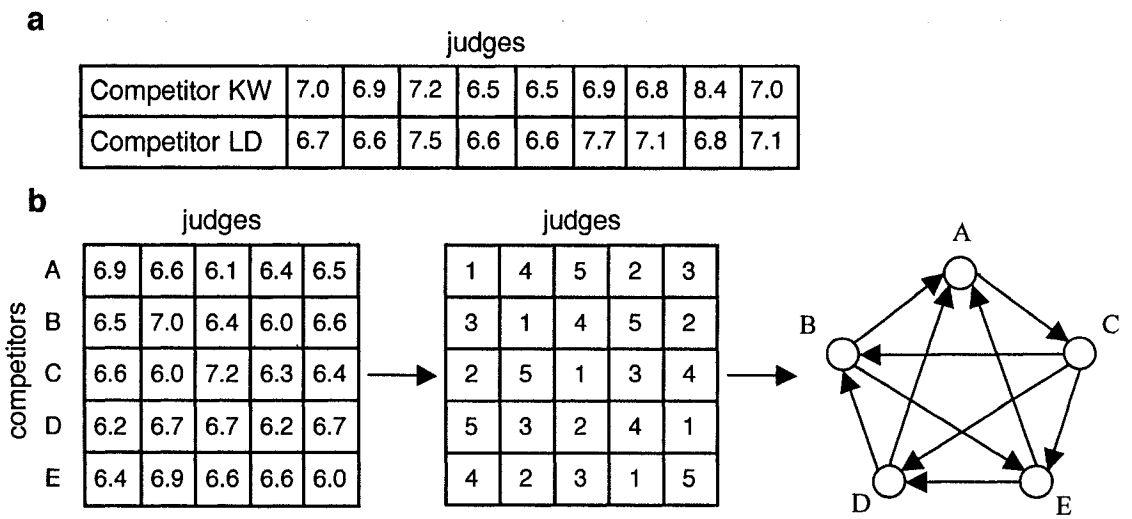


Figure 5–1: The voting paradox. **a.** Two competitors and their scores taken from the 2006 U.S. Junior Olympic Championship [167]. The sum and majority rules lead to different rankings. **b.** A hypothetical example with 5 judges and 5 competitors. The sum of the scores of all competitors is the same. Using the majority rule, any score-sheet becomes a table of rankings. Based on this ranking, we can generate a directed graph where the nodes are the competitors. If more judges rank competitor A better than B, then A beats B and there is an arrow from A to B. The resulting directed graph has a cycle through all the nodes, and a clear ranking cannot be found.

any competitor. By comparing competitors one by one, we can represent a score-sheet as a directed graph. Using results from the theory of tournaments [160], we have shown (see Appendix C) that a Latin square score-sheet produces a directed graph with a cycle through all the nodes. Thus, there is no clear winner. This is an example of the well-known voting paradox that emerges from majority voting between several potential options [156, 157, 158, 159, 160].

5.3 Distribution of Microarray Probe Intensities

The controversies and difficulties in ranking figure skaters highlight the need for a systematic analysis of evaluation methods. Before determining the best way to combine scores from different judges, we should consider the collective distribution of their scores. In situations such as figure skating, the distribution of the scores from different judges is comparatively narrow, primarily due to the guidelines for awarding points (Fig. 5-2a). In other cases the distributions can be comparatively broad, such as gene expression microarrays. Microarrays survey the expression of thousands of genes using DNA probes. Each gene's mRNA transcript has a set of probes designed to bind it specifically in different regions. Samples of mRNA are labelled so that the amount bound to each probe on the microarray can be measured. Here, the evaluators are the probes and the scores are the measured fluorescent intensities. The distribution of fluorescent intensity from different probes all exposed to the same concentration of mRNA is approximately log normally distributed (Fig. 5-2b). These broad distributions partly result from differences in the binding affinities and conformations of individual probes.

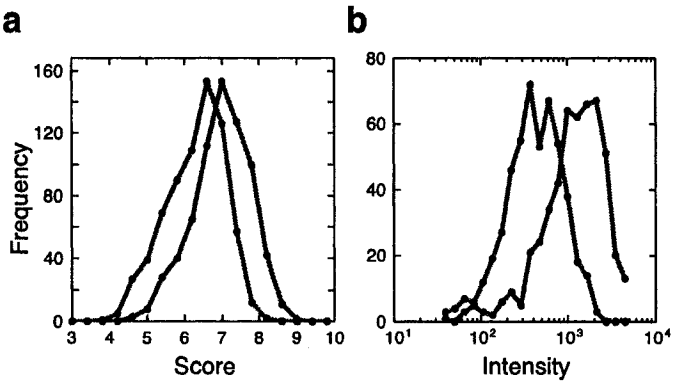


Figure 5–2: Distributions of evaluators' scores. **a.** Distribution of scores for competitors ranked 1-5 (red) and those ranked 6-10 (blue) at the US Junior Figure Skating Championship [167]. The range of the scores is between 3 and 10 due to scoring guidelines. The overlap shows that competitors ranked 6-10 can be given scores as high as those ranked 1-5. **b.** Actual distributions of the fluorescence from microarray probes for mRNA at concentrations of 16 pM (blue) and 64 pM (red) from the Latin square dataset [46]. The distributions are approximately log normal and cover a broad range due to the fluctuations in probe binding, scanning effects, and other biases.

5.4 Analysis of Evaluation: Accuracy and Cost

To see how the distribution of scores from different judges influences the final decision, we examine the accuracy of evaluations amalgamated from individual judges. We first assume that there are two competitors A and B, where A is better than B. An individual judge's scores for A and B will be drawn from two different normal distributions, such that the mean score for A will be higher (See Appendix A). The amount of overlap between the two distributions indicates the inaccuracy of the evaluation, quantifying how often B is incorrectly given a higher score than A. For simplicity, each judge will have the same accuracy, say 65% or 70%. The

total accuracy of the panel of judges, therefore, will depend on the number of judges and whether their mean scores come from a tight or broad distribution, modelled by a normal or a log normal distribution, respectively. Comparing the panel's accuracy as a function of the number of judges, the sum rule is superior to the majority rule when the judges' mean scores are tightly distributed (Fig. 5-3a). The majority rule is superior, however, when the judges' mean scores are broadly distributed. Furthermore, when the judges' mean scores are broadly distributed, the majority rule of judges with 65% accuracy can even surpass the accuracy of the sum rule based on judges with 70% accuracy.

Although increasing the number of judges leads to improved accuracy, in practical circumstances it also increases the cost. We can compute the optimal number of judges if we know the ratio between the cost per error and the cost per judge. To illustrate this, if the cost per error is 75 times the cost per judge, then the total cost has a minimum at five to nine judges depending on the distribution of the judges' scores and the method for combining them (Fig. 5-3b). If the ratio of the cost per error to the cost per judge increases, so would the optimal number of judges.

5.5 Majority Rule Algorithm for Microarray Analysis

Because microarray probe intensities are broadly distributed (Fig. 5-2b), majority rule algorithms should outperform sum rules in detecting differential gene expression. Although there is considerable research on evaluating gene expression with microarrays, the two main algorithms [14], Model Based Expression Index (MBEI) [29] and the Robust Multichip Average (RMA) [18], both use a weighted

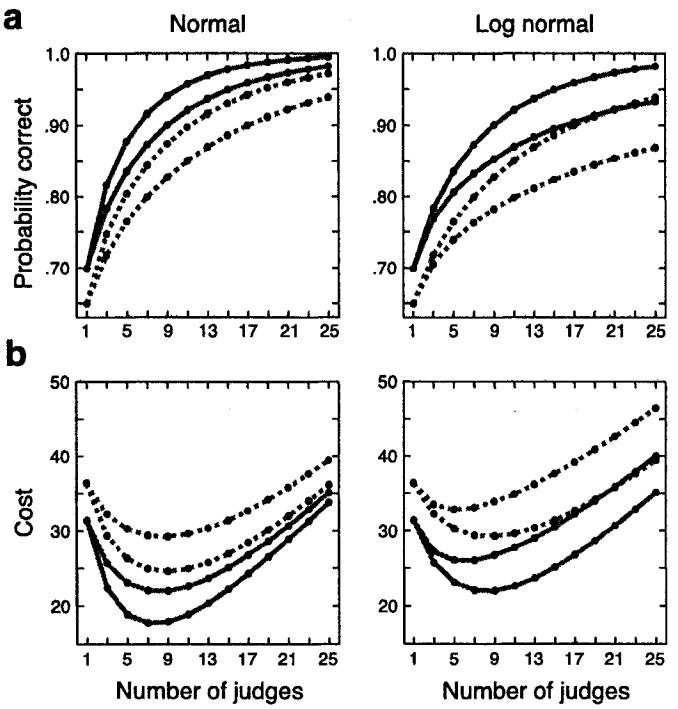


Figure 5–3: Comparison of the sum rule and majority rule for different distributions of the mean scores of the judges. **a.** The accuracy of a panel of judges employing a sum rule (red) or a majority rule (blue) plotted versus the number of judges. The judges' mean scores are sampled from either a normal distribution with mean 700 and standard deviation 70 (left) or a log normal distribution with mean 700 and standard deviation 1400 (right). Each data point represents the mean accuracy from 10,000 different samples of judges' scores. Individual judges with accuracies of 70% (solid lines) are plotted along with judges with accuracies of 65% (dashed line). At both the 65% and 70% level of individual judge accuracy, a nonparametric sign test confirms the differences in majority rule and sum rule performance are statistically significant with a p-value $< .01$ for all cases of more than one judge. **b.** Cost functions versus the number of judges. Assuming the cost function $C_{\text{total}} = C_{\text{error}} P_{\text{error}} + C_{\text{judge}} n_{\text{judge}}$, where C_{error} is the cost per error, P_{error} is the probability of an error (1- probability correct), C_{judge} is the cost per judge, and n_{judge} is the number of judges. The ratio of the cost of an error to the cost of a judge is 75. The same labelling scheme as used in **a.** applies.

sum to combine the probe intensities for each gene. A statistical test like Significance Algorithm for Microarrays (SAM) [39] is then applied to these values to detect statistically significant changes in each gene.

To evaluate whether a majority rule based algorithm would outperform current gene expression evaluation methods, we developed a new algorithm to detect differentially expressed genes. Assuming intensities from individual probes are normally distributed, we compare each probe between experimental groups using a t-test with a $p<0.01$ as the cut-off for calling a probe changed. We then count the number of probes for each gene that changed and use the binomial function to compute the probability that a gene with n probes would have k changed. Because of this emphasis on the individual probes and their equal role in determining differential expression, we call the algorithm Probe By Probe (PBP). We compare the results of PBP to the results obtained using the MBEI-SAM and RMA-SAM methods using a publicly available dataset [46] from Affymetrix in which a collection of genes are spiked in at different concentrations. We compute the receiver operator characteristic curve [168] for each algorithm (Fig. 5-4) to show the tradeoffs between selecting true positives while excluding false positives. Our algorithm outperforms both RMA-SAM and MBEI-SAM in the area of the fewest false positives.

Since the PBP algorithm focuses on individual probes, it facilitates the identification of faulty probes and could lead to improved gene expression analysis. Furthermore, both RMA and MBEI use fitting routines to estimate the weights or binding affinities of individual probes. Currently, adding or subtracting probes for genes requires recomputing the weights for several, if not all, genes and performing the statistical tests again. In contrast, changing the composition of a gene's probe

set in PBP only requires recalculating the binomial function for that gene. This flexibility should help investigators to interpret their microarray experiments.

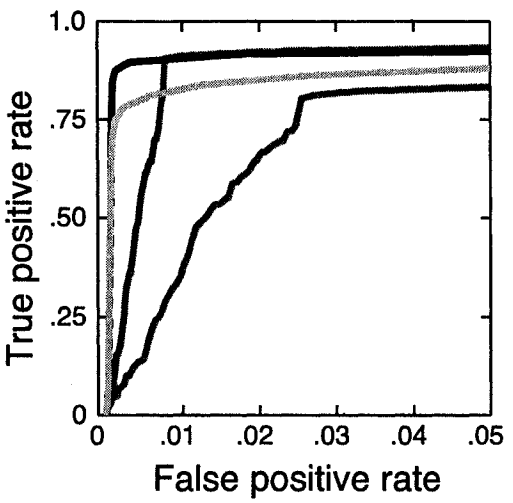


Figure 5–4: Receiver operating characteristic curves comparing the Probe By Probe (green) method, based on a majority rule algorithm, with current methods RMA-SAM (red) and MBEI-SAM (blue), both based on a sum rule algorithm. Each point is a different cut-off, q-values for the methods using SAM and probabilities for PBP. The receiver operating characteristic curves show that PBP affords the highest levels of true positives for the lowest levels of false positives. PBP with the RMA normalization to correct for differences in microarray brightness is plotted as a black line, outperforming the PBP without any initial normalization.

5.6 Cost Analysis for Figure Skating

If we apply the cost-benefit analysis to evaluations, we can gain information on the implied costs of decisions. For example, by comparing the accuracy of judges in evaluating the top competitor at every stage in the 2006 US Junior Figure Skating Championship [167], we estimate the accuracy of an individual judge to be about 76% (See Appendix B). Assuming that the competition choice of

using nine judges is optimal, then the implied ratio of cost of an error in evaluation to cost of a judge is 100 – 152. Using the majority rule with such conditions, the expected accuracy of this evaluation would be about 95%. This analysis can also be applied to microarrays to determine the optimal number of probes for each gene. Genes with less accurate probe sets will need more probes than those with reliable probe sets. Furthermore, adjusting probe set numbers can ensure that each gene is as likely to be detected as another. Besides figure skating and microarrays, this cost-benefit analysis has a wide gamut of applications including grant reviews, boxing matches, and United States Supreme Court decisions (See Appendix B and Table 5-1).

5.7 Discussion

Here, we present a framework that offers a theoretical foundation for deciding on the optimal numbers of judges and the best ways to combine their scores. We used this framework to show that in figure skating a sum rule works better, validating the recent rule changes. We also applied it to the selection of differential gene expression in microarrays to suggest an alternative approach. By focusing on the distribution of judges' scores, we have been able to compare different methods of evaluation and suggest circumstances in biotechnology in which current methods may be improved.

5.8 Acknowledgement

We thank NSERC, MITACS, and CIHR for financial support. We thank Drs. P. Swain and E. Cooper for helpful conversations.

5.9 Appendix A: Evaluation Accuracy

5.9.1 Individual Judge Accuracy

We assume that there are two options A and B, where A is better than B. An individual judge's accuracy is, therefore, the percentage of time A is given the higher score. If a judge scores options A and B according to the distributions D_A and D_B then the overlap between these distributions determines this error (Fig. 5-5).

$$P_{\text{correct}} = 1 - P_{\text{error}} = 1 - \int_{-\infty}^{\infty} D_{\text{B}}(x) \left(\int_{-\infty}^{x} D_{\text{A}}(y) dy \right) dx \tag{5.1}$$

We assume D_{A} and D_{B} are normal distributions so equation (5.1) becomes:

$$P_{\text{correct}} = \frac{1}{2} - \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\text{B}}} \exp\left(\frac{-(x-\mu_{\text{B}})^2}{2\sigma_{\text{B}}^2} \right) \text{erf}\left(\frac{x-\mu_{\text{A}}}{\sigma_{\text{A}}\sqrt{2}} \right) dx \tag{5.2}$$

Throughout our work we assume that the individual judges have equal accuracy. To ensure this in our calculations, after we sampled the mean of a judge's score for option B, from either the tight (normal) or broad (log normal) distribution, we then calculated the mean and standard deviation for option A to keep P_{correct} the same. If two normal distributions have the same coefficient of variation, $\frac{\sigma}{\mu}$, and the means have a fixed ratio, $\frac{\mu_{\text{A}}}{\mu_{\text{B}}}$, then the value of P_{correct} is constant.

5.9.2 Sum Rule

If each judge scores according to a normal distribution then the total score is also a normal distribution with a mean and variance equal to the sum of the means and variances of the individual judges. Thus, the total score for an option with n judges would be:

$$D_{\text{totA}} = \frac{1}{\sqrt{2\pi}\sigma_{\text{t}}} \exp\left(\frac{-(x-\mu_{\text{t}})^2}{2\sigma_{\text{t}}^2}\right), \tag{5.3}$$
$$\sigma_{\text{t}} = \sqrt{\sigma_{\text{A1}}^2 + \sigma_{\text{A2}}^2 + \ldots + \sigma_{\text{An}}^2},$$
$$\mu_{\text{t}} = \mu_{\text{A1}} + \mu_{\text{A2}} + \ldots + \mu_{\text{An}}$$

Once we have the means of the individual judges, we can compute the distribution of the sum of their scores using equation (5.3) and calculate the probability correct using equation (5.1).

5.9.3 Majority Rule Accuracy Calculations

To calculate the accuracy of an evaluation made employing the majority rule, we only need to consider the probability an individual judge is correct and the total number of judges. When all judges have an equal probability of being correct, p, and there are n judges (n is odd) the probability of making a correct decision is a sum of binomial probabilities.

$$P_{\text{panel\_correct}} = \sum_{i=\frac{n+1}{2}}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{5.4}$$

We compare this to the results obtained from using the sum rule where the p from equation (5.4) is the same as P_{correct} from equation (5.1).

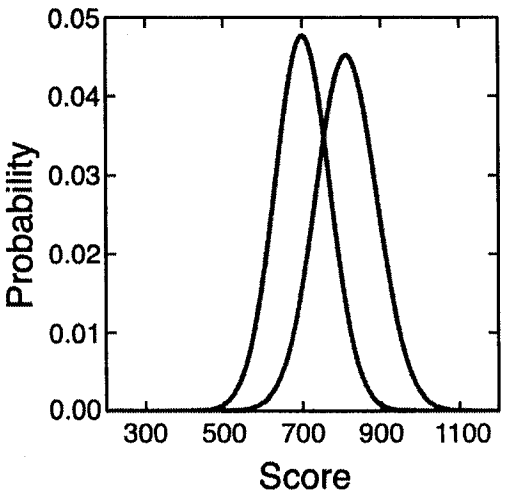


Figure 5–5: Sample score distributions for a judge with 70% accuracy. Option A (red) has a higher mean and standard deviation than option B (blue). The degree of overlap between the two quantifies the error. Judges with higher accuracy have scoring distributions farther apart.

5.10 Appendix B: Cost-benefit Analysis

5.10.1 Cost-benefit Analysis: Figure Skating

For the cost-benefit analysis, we use a cost function that relates the cost per judge, C_{judge}, to the cost of making an incorrect decision, C_{error}.

$$C_{\text{tot}} = n_{\text{judge}}\, C_{\text{judge}} + C_{\text{error}}\, P_{\text{error}}(n_{\text{judge}}) \tag{5.5}$$

The cost of an error is multiplied by the probability of making an error, P_{error}, which depends only on the number of judges n_{judge}. Although P_{error} also depends on the accuracy of the individual judges and the type of rule used, we consider those fixed for the cost-benefit analysis.

For figure skating, we assume the competition uses a majority rule. To estimate an individual judge's accuracy, p, we used data from the 2006 U.S. Junior Figure Skating Championships [167]. We assume each judge is equally accurate and then count the number of times a judge failed to rank the top competitor of a stage number one in that stage. This serves as an upper limit to the accuracy since it assumes the number one competitor was in fact the best. Based on this, we estimate an individual judge accuracy of 76%. If the figure skating competition's choice of nine judges is optimal, we can estimate the ratio of cost per error to cost per judge by creating two inequalities.

$$\begin{aligned} 9\ C_{\text{judge}} + C_{\text{error}}\ P_{\text{error}}(9) &\leq 7\ C_{\text{judge}} + C_{\text{error}}\ P_{\text{error}}(7) \\ 9\ C_{\text{judge}} + C_{\text{error}}\ P_{\text{error}}(9) &\leq 11\ C_{\text{judge}} + C_{\text{error}}\ P_{\text{error}}(11) \end{aligned} \tag{5.6}$$

Simplification can define the range of any parameter.

$$\frac{P_{\text{error}}(9) - P_{\text{error}}(11)}{2} \leq \frac{C_{\text{judge}}}{C_{\text{error}}} \leq \frac{P_{\text{error}}(7) - P_{\text{error}}(9)}{2} \tag{5.7}$$

Substituting in the 76% judge accuracy and using the majority rule as our $P_{\text{error}}(n_{\text{judge}})$ function we calculate the range for $\frac{C_{\text{error}}}{C_{\text{judge}}}$ to be between 100 and 152.

5.10.2 Cost-benefit Analysis: Other Examples

We applied the cost-benefit analysis to three other examples: boxing, grant reviews, and the U.S. Supreme Court (Table 5-1). In each case, we calculated the shaded quantity to demonstrate the implied costs or accuracies.

We estimated the accuracy of the U.S. Supreme Court and World Boxing Association judges using the same approach as we did for figure skating. We used all court cases from the 2005-2006 term for the Supreme Court [169] and all World

Boxing Association boxing bouts in 2006 for the boxing judges [170]. In the case of the Supreme Court, the accuracy estimate shows little change when using cases from the 2004-2005 term, dropping from 83.7% to 80.6%.

To estimate the cost per judge in both grant reviews and the U.S. Supreme Court, we considered the cost per judge per review or case. Since Canadian Institutes of Health Research (C.I.H.R.) grant reviewers typically review 30-70 grants in two days [171], we assumed $ 1,000 in travel expenses and divided it over 50 grants to obtain 20 dollars per grant. Supreme Court justices have an average yearly salary of $ 203,000 [172] and typically hear oral arguments on 100 cases a term [173], which means the cost per judge per case is $2,030.

The cost per error of a grant review is assumed to be the value of the grant, $100,000 in the case of a C.I.H.R. operating grant [171]. For the boxing example, we assume a prize of $100,000 for winning the bout. The actual amount varies depending on the promotion, the weight class, and the organization. In both cases, the prize is considered the cost of an error since the wrong person receives it.

Since both the Supreme Court and boxing bouts use majority rules, the $P_{\text{error}}(n_{\text{judge}})$ functions are simply $1 - P_{\text{panel\_correct}}$ from equation (5.4).

In the boxing example, we assume the choice of three judges is optimal and use the inequality approach, see equation (5.6), to determine the range for a judge's salary. We use the same approach to estimate the cost of the U.S. Supreme court making a wrong decision on a case. Grant reviews, however, are not as straightforward since they employ a sum rule rather than a majority rule. For our $P_{error}(n_{\text{judge}})$ function we constructed a scenario with two competing grants, as in Fig. 5-3. Not only did the judges score each grant according to a normal distribution, but their means were sampled from a tight normal distribution. We chose this model because grant reviewers have guidelines requiring them to give

scores from 0 to 5, similar to the figure skating example. Using this model and the assumption all judges have equal accuracy, we were able to estimate an individual judge's accuracy using equation (5.6).

In compiling the information in Table 5-1, we had to make a number of assumptions that are open to debate. For example, if we had assumed that the cost per grant reviewer was a higher figure, say $200, then the implied accuracy would shift to 74-75.5%. Given the cost of either $20 or $200 per reviewer per grant, the implied accuracy for grant reviewers is considerably lower than the estimated accuracy for judges at a boxing match. If reviewer accuracy could be improved by training (or if it is higher than the implied accuracies given here) then an optimal evaluation procedure would reduce the number of grant reviewers on a typical panel. This would have beneficial effects of reducing the loads on reviewers and streamlining reviewing procedures. In contrast, given that the actual monetary value of many U.S. Supreme Court hearings may often be much higher than the $496,760-1,011,400 range, a larger number of judges might be desirable. By making explicit the relationships between accuracy of judges, the cost per judge, and the cost per error, the methods reported here may help policy makers optimize decision making processes.

Table 5–1: Cost-benefit analysis applications

Event	Number of judges	Cost per judge(USD)	Judge accuracy	Cost of an error (USD)
Boxing match	3	305-2138	95%	100,000
Grant review	10	20 per review	80-82%	100,000
U.S. Supreme Court	9	2,030 per case	83.7 %	496,760-1,011,400

5.11 Appendix C: Latin Square Tournaments

5.11.1 Introduction

A competition involving multiple competitors can be represented by a directed graph (digraph) with each vertex representing a different competitor. If competitor i beats competitor j, then there is directed edge from vertex v_i to vertex v_j. Assuming that ties are not allowed, if each competitor plays each other competitor exactly once, then in the associated digraph each vertex will be connected by a uniquely oriented edge to each other vertex of the graph. Such digraphs, called *tournaments*, have been the subject of a great many studies, for example see Harary and Moser [160] or Thomassen [174] for reviews and references to early work. Tournament digraphs can arise in a wide range of different circumstances. Paradoxical situations such as the "voting paradox", in which there is no clear winner but rather a cycle in the tournament digraph are well known [158].

We consider a competition in which there are n competitors and n judges. Each judge assigns a numerical rating of $1, 2, \ldots, n$ to each competitor. If all judges were perfect, the rating for each competitor from each judge would be identical. We assume a different case, in which the judges are mediocre, so that no two judges assign the same rating to any competitor. Typical score-sheets are shown in Table 5-2 for $n = 7$, where the rows represent competitors and the columns represent judges. Grids such as these are Latin squares and have been popularized recently by the Sudoku craze, since a Sudoku is a Latin square where $n = 9$. A *Latin square score-sheet* is an $n \times n$ matrix in which each entry is an integer $1, 2, \ldots, n$, and in which all the integers $1, 2, \ldots, n$ are contained exactly once in each row and in each column.

Table 5–2: Latin square score-sheet with $n = 3$

	J_1	J_2	J_3
v_1	1	2	3
v_2	2	3	1
v_3	3	1	2

Table 5–3: Latin square score-sheet with $n = 7$ with $s_1 = 1$ and $s_7 = 5$

	J_1	J_2	J_3	J_4	J_5	J_6	J_7
v_1	1	2	3	4	5	6	7
v_2	3	7	5	1	6	2	4
v_3	4	5	6	2	1	7	3
v_4	5	6	4	7	3	1	2
v_5	6	3	7	5	2	4	1
v_6	2	4	1	6	7	3	5
v_7	7	1	2	3	4	5	6

The average rating of each competitor is exactly the same. We assume that competitor i beats or dominates competitor j if more judges rate competitor i better than competitor j. Provided the numbers of competitors and judges is an odd number, there will be a clear winner between each pair of competitors. Consequently, a tournament digraph can be constructed. We call the digraph constructed based on a Latin square score-sheet a *Latin square tournament.* Figure 5-6 shows the digraph associated with the Latin square score-sheet in Table 5-2. It contains a Hamiltonian cycle, passing through each of the vertices. Thus, since all competitors lie on a cycle, it is impossible to determine which competitor is best in a Latin square tournament.

5.11.2 Basic Properties of Tournaments

For completeness, we present basic definitions and properties of tournaments, but refer the reader to [160, 174] and references therein for a more thorough treatment. A directed graph, or *digraph*, consists of a finite set $V = v_1, v_2, \ldots, v_n$

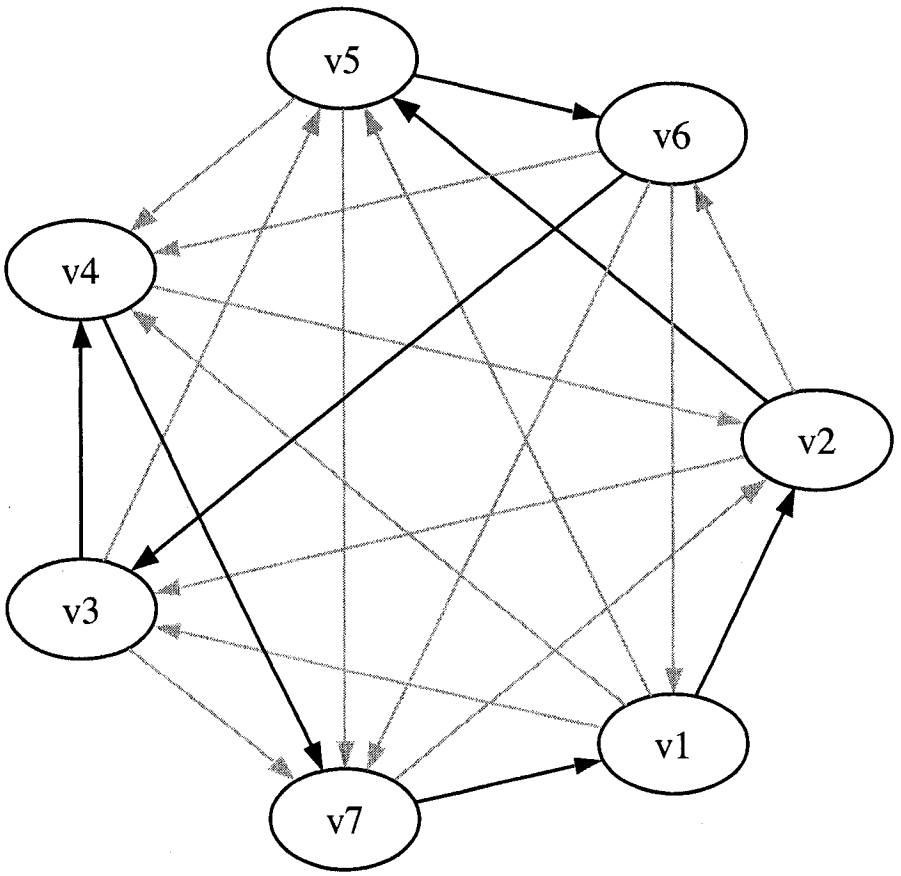


Figure 5–6: Digraph for Table 5-3

of vertices, and a finite set of directed edges. Each edge is directed from a vertex v_i to a second different vertex v_j. A *tournament* is a digraph in which there is a single uniquely directed edge between every pair of vertices. If an edge is directed from v_i to v_j we say that v_i dominates v_j.

The *outdegree* (resp. *indegree*) of a vertex is the number of edges directed outwards from (resp. inwards towards) it. The *score*, s_i of vertex v_i is equal to its outdegree. Without loss of generality we can label the vertices to generate a *score sequence* in which $s_1 \leq s_2 \leq \cdots \leq s_n$.

Property 1. It follows from the above definitions that the number of edges in a tournament is equal to $\frac{1}{2}n(n-1) = S(n)$. Likewise the sums of the outdegrees (or indegrees) of the vertices in a tournament is equal to $S(n)$.

A digraph is *strongly connected* if for each pair of two different vertices v_i and v_j there is a directed path that starts at v_i and ends at v_j. A path that starts and ends on the same vertex is called a *cycle.* A *Hamiltonian cycle* passes once through every vertex of the graph.

We use the following two theorems based on Harary and Moser [160].

Theorem 1 *A strongly connected tournament has a Hamiltonian cycle.*

Theorem 2 *Let T be a tournament with a score sequence $s_1 \leq s_2 \leq \cdots \leq s_n$. Then T is strongly connected if and only if*

$$\sum_{i=1}^{n} s_i = \frac{1}{2}n(n-1) = S(n), \tag{5.8}$$

and the following inequalities hold for every positive integer $p < n$,

$$\sum_{i=1}^{i=p} s_i > S(p). \tag{5.9}$$

5.11.3 Basic Properties of Latin Square Tournaments

In order to analyze Latin square tournaments, we define the *dominance.* Assume the *ratings* of vertex v_i by judge j are given by the vector $R_{i,j}$, where $j = 1, 2, \ldots, n.$. The *dominance* $D_{i,k}$ of vertex v_i over vertex v_k is

$$D_{i,k} = \sum_{j=1}^{n} H(R_{i,j} - R_{k,j}), \tag{5.10}$$

where the Heaviside step function $H(x) = 1$ for $x > 0$, and $H(x) = 0$ for $x \leq 0$. Clearly, $D_{i,k} + D_{k,i} = n$.

Property 2. Since in each column of the Latin square score-sheet, each integer $1, 2, \ldots, n$ appears exactly once, it follows that, for each i,

$$\sum_{k=1;k\neq i}^{n} D_{i,k} = S(n). \tag{5.11}$$

Thus, $S(n)$ also represents the total number of times that the ratings of any given vertex will dominate the ratings of all the other vertices.

We now determine the maximum score for any vertex in a Latin square tournament. In order for v_i to dominate v_k, we must have $D_{i,k} \geq \frac{(n+1)}{2}$. Thus, the maximum score for any vertex is the largest L for which

$$L\frac{(n+1)}{2} \leq S(n),$$

so that $L = n - 2$. This leads to the following property.

Property 3. In a Latin square tournament the maximum outdegree for a vertex is $n - 2$ and the minimum outdegree for a vertex is 1.

We now give an example to help fix ideas, and to present a special case of the general result in the next section.

Example 3 *The Latin square tournament in which $s_1 = 1$ contains a Hamiltonian cycle.*

Table 5-2 gives a Latin square score-sheet with $s_1 = 1$, $n = 7$. The associated tournament digraph for Table 5-2 is Fig. 5-6.

For the case of arbitrary n, we suppose that v_1 is dominated by $v_2, v_3, \ldots, v_{n-1}$ with $D_{1,k} = \frac{n-1}{2}$, for $k = 2, 3, \ldots, n-1$, and that v_1 dominates v_n. Consequently,

$$\sum_{k=2}^{n-1} D_{1,k} = \frac{1}{2}(n-2)(n-1),$$

and $D_{1,n} = (n-1)$. The rating n in v_n (e.g. $R_{n,1}$ in Table 5-2) must always dominate. Further $H(R_{n,k} - R_{j,k}) = H(R_{1,k} - R_{j,k})$ for $j = 2, 3, \ldots, n-1$ and $k = 2, 3, \ldots, n$. Since $D_{1,k} = \frac{n-1}{2}$ for $k = 2, 3, \ldots, n-1$, $D_{n,k} = \frac{n+1}{2}$ for $k = 2, 3, \ldots, n-1$. Thus, v_n will have a score of $n-2$. For any two vertices, v_j and v_m different from v_1 and v_n, there is a path $v_j \rightarrow v_1 \rightarrow v_n \rightarrow v_m$. Further, $v_1 \rightarrow v_n$, and there is a path from $v_n \rightarrow v_k \rightarrow v_1$, where $k \neq 1$ and $k \neq n$. Since the tournament graph is strongly connected from Theorem 1, the graph contains a Hamiltonian cycle, Fig. 5-6.

5.11.4 Hamiltonian Cycles in Latin Square Tournaments

We now present our main result.

Theorem 4 *All Latin square tournaments contain a Hamiltonian cycle.*

Assume a tournament of n vertices generated by a Latin square score-sheet with n judges. If the inequality in Eq. (5.9) is true for all $p < n$, our result is established. We establish the proof by contradiction.

Assume the inequality in Eq. (5.9) is not true for some particular $p < n$, so that

$$\sum_{i=1}^{p} s_i = S(p). \tag{5.12}$$

Call P the set of vertices $v_1, v_2, \ldots, v_p$, and call M the set of $m = n - p$ vertices, $v_{p+1}, v_{p+2}, \ldots, v_n$, not in P. Any edge between a vertex in M and a vertex in P must be directed from the vertex in M to the vertex in P. Consequently, $s_i > s_j$

where $v_i \in M$ and $v_j \in P$. Therefore, the first p entries in the score sequence $s_1 \leq s_2 \leq \cdots \leq s_n$ are in P and the remainder are in M.

Since no vertex in P can dominate a vertex in M, for any vertex $v_j \in P$ and $v_k \in M$,

$$\max \sum_{k=p+1}^{n} D_{j,k} = \frac{1}{2}(n-1)m.$$

Since

$$\sum_{j=1}^{p}\sum_{k=1}^{p} D_{j,k} = nS(p),$$

we find

$$\max \sum_{j=1}^{p}\sum_{k=1}^{n} D_{j,k} = nS(p) + \frac{1}{2}(n-1)pm. \tag{5.13}$$

However, we know from Property 1 that

$$\sum_{j=1}^{p}\sum_{k=1}^{n} D_{j,k} = pS(n). \tag{5.14}$$

In order for Eq. (5.13) and Eq. (5.14) to both be satisfied we must have

$$\max \sum_{j=1}^{p}\sum_{k=1}^{n} D_{j,k} - \sum_{j=1}^{p}\sum_{k=1}^{n} D_{j,m} \geq 0. \tag{5.15}$$

However, this cannot be since the difference in Eq. (5.15) is equal to $-m$. Consequently, Eq. (5.12) cannot be true, and by Theorem 2, Latin square tournaments must have a Hamiltonian circuit.

CHAPTER 6
Algorithm Applications

Chapter 4 exposed several problems encountered during microarray data analysis, and Chapter 5 introduced a new algorithm, PBP, for analyzing such data. In this chapter, I explore how the new algorithm addresses the previous problems. I then apply the new algorithm to the canine data from Chapter 3 to see how it handles data with biological variability. After comparing my new results to the RT-PCR confirmations, I further extend the analysis to search for regulated pathways. This is not a complete overhaul of the earlier analysis, but rather serves to demonstrate the effectiveness and application of PBP. This chapter represents a preliminary analysis of the original data and opens the door to future work and deeper analysis.

6.1 Solutions to Microarray Analysis Problems

The algorithm PBP presented in the previous chapter outperformed RMA-SAM and MBEI-SAM on the Latin square data set [46]. This validation data set, however, only had 42 genes present out of a total ~22,000 probesets. It could be that other validation data sets would show different results. Beyond performance, however, PBP successfully solves each of the problems posed in Chapter 4, which current techniques do not.

6.1.1 Solution 1: Clarity

Compared to other microarray algorithms, the PBP method handles the data in a simple and transparent manner. Using either original or normalized intensity values, PBP determines which probes changed and calculates the probability that a gene would have so many changed probes. Unlike RMA and MBEI, which compress probes into one value based on a particular model, the only transformation in PBP is the normalization step done beforehand. If a gene is found unchanged in an MBEI or RMA method, it could be due to the computation of the expression index, the selected options, or the differential gene selection algorithm. If, on the other hand, a gene is found unchanged using PBP then simply not enough probes were found changed. Investigating these probes may reveal poor hybridization properties, such as self-binding, or perhaps that the wrong reference sequence was used to generate the probes. Tracking down the unchanged probes in PBP could lead to better probe selection in future microarrays.

PBP also improves the transparency of gene selection by increasing flexibility. In PBP, it is easy to see the effects of removing or adding a probe because such

manipulations only affect the probability calculation for one gene. RMA and MBEI methods currently do not make such probe adjustments easily: the probeset mapping file would need to be modified and both the fitting routines and the differential gene selection algorithm would need to be rerun. The flexibility of PBP also makes results easy to adapt to future modifications of reference sequences. For example, if a gene is later found to have a different sequence such that some of the original probes no longer match, the only results that change in PBP is the probability calculation for that gene. This problem is not trivial since sequence databases are very dynamic. They contain a mixture of old and new sequences of different reliability, with more and more sequences being uploaded [175]. Thus, PBP improves the clarity of gene selection and offers a unique flexibility that makes microarray results adaptable to new information.

6.1.2 Solution 2: Annotations

The PBP method can use the original probeset assignments and, therefore, have the same problems as other algorithms. On the other hand, we propose that the probeset assignments be disregarded and the probes mapped to a set of known transcripts like NCBI's refseq database [175], which contains a curated list of non-redundant transcripts. As far as we know, no other algorithm ignores the probeset designations used by Affymetrix. This is surprising because one study showed that up to 37% of the probes on an Affymetrix microarray can be reassigned and up to ~10% detect multiple transcripts [176]. For each transcript, we find the matching Affymetrix probes and create a probeset this way. Because the transcripts are known, there is no need to do further analysis or BLAST [145, 146] alignment. This mapping also makes it clear whether the microarray can distinguish different

isoforms or transcript variants: if the transcript variants have the same sets of probes, they cannot be distinguished. This method does produce the problem that some probes will not correspond to a unique gene or transcript. This, however, is not a new problem as it exists using any other algorithm; it is just made explicit using PBP instead of incorporated into a weighted sum as in RMA or MBEI. By making it more explicit, any substantial issues presented by non-unique probes may prompt Affymetrix to replace them with more specific probes. In sum, PBP removes the task of identifying the targets of probesets and solves associated annotation problems.

6.1.3 Solution 3: Multiple Probesets

Using reference sequences and mapping Affymetrix probes to them also resolves the problem of multiple probesets. If multiple probesets correspond to the same transcript, then in PBP all the probes from these four probesets will be used in one decision. In the case of fibrillin 1 presented in Chapter 4, three of the four probesets were found upregulated and one was found unchanged. Mapping the probes to fibrillin 1, we found that all of the probes for the three upregulated probesets targeted fibrillin 1 and were found statistically upregulated. The probes from the unchanged probeset, annotated by Affymetrix as fibrillin 1, did not target the reference sequence. So, instead of conflicting results, we found 33 of 33 probes upregulated. Moreover, the reference sequence matching ensures we know exactly which genes can be quantified using the microarray and which cannot. To our knowledge, other algorithms do not address the problem of multiple probes in any stage of the gene selection process.

6.2 Application of PBP to the Canine Data

Instead of using PBP on more validation data sets with a small number of transcripts present, we used it to analyze the microarray data from Chapter 3. Because this data set has large numbers of transcripts present and has biological variability both within replicates and between groups, we thought this would serve as an excellent opportunity to assess PBP's performance in a "realistic" setting.

6.2.1 Mapping Probes to Genes

To test PBP on the canine microarray data, we first abandoned the Affymetrix probeset assignments. Instead, we mapped the probes to a set of reference sequences to avoid both the multiple probeset dilemma and the annotation complications. We chose the NCBI canine refseq database which contains a collection of 33,644 curated, nonredundant transcripts [175]. Because it is curated, it supposedly offers reliability over coverage. Although similar mappings have been performed before using BLAST [176], we looked for only perfect alignments between probes and reference sequences, forbidding any gaps or mistakes. Of the total 263,234 probes on the canine microarray, ~69 % did not match a transcript due to a combination of factors including differences in the reference sequences used by Affymetrix, missing sequences from the NCBI refseq database, and probes designed to target expressed sequences rather than genes [176]. From the mapping, we discovered that over 30,000 probes matched more than one transcript. Closer inspection revealed that some of these probes matched sequences common to transcript variants, confirming the earlier doubt that Affymetrix probesets could not distinguish transcript variants [35, 176].

Another striking discovery was that Affymetrix designs its probesets so that over 99% of its probesets have 11-15 probes while over 50% of the transcripts with probes have more than 15 probes or less than 11 probes. This means that genes do not have equal numbers of probes even though the probesets would give that impression. From the work in Chapter 5, the accuracy of a group of probes depends on both their number and their individual accuracies. Therefore, two genes might have an equal level of detection despite the fact that one has 16 mediocre probes while the other has 7 good probes. Microarray analysis algorithms do not correct for the possibility that some genes may be more detectable than others. PBP does not have a suitable way of solving this problem, but because its use can identify habitually problematic probes it can lead to better probe design which can help to address this issue.

After mapping the probes to the transcripts, we used the MBEI normalization [30] on the raw probe intensities to remove a source of variability from the comparison. We applied PBP and selected genes changed with a q-value of 1%, as was done earlier, using Hochberg's procedure [177]. The results shown in Table 6-1 reveal a qualitative agreement with the analysis of the VTP data from before. The ATP data, on the other hand, no longer shows an adaptation. Using an RMA normalization or no normalization at all did not change this result. To see if this is caused by our mapping, we used PBP with the original Affymetrix probeset assignments. In conjunction with either no normalization or the RMA normalization, we still find that 1 week ATP has more probesets changed than 24 hour ATP. In fact, the only time PBP reports the adaptation is when the original probeset assignments are used with the MBEI normalization, which represents the situation closest to the original analysis. It is difficult to validate whether such an adaptation occurs, but we can check a selection of genes with the RT-PCR data.

Table 6–1: The number of changed genes in each experimental condition using PBP

Condition	PBP # up	PBP # down
24 hr ATP	395	317
1 wk ATP	378	946
24 hr VTP	1756	1522
2 wk VTP	1721	1683

6.2.2 RT-PCR Comparison

We used the 18 genes from the RT-PCR results of Chapter 3 to compare PBP to the MBEI-SAM approach used earlier. These 18 genes were chosen from the probesets used in the MBEI-SAM analysis, and the RT-PCR probes were constructed based on the target sequences of those probesets. We used the RT-PCR forward and reverse primer sequences along with the probe sequences to find the matching transcripts in the canine NCBI refseq database. In several cases, the RT-PCR primers and probes matched multiple reference sequences, primarily because the reference sequences include transcript variants as well as the primers and probes target regions shared by all transcript variants. For the comparisons using PBP, these transcript variants often shared the same probes and so did not give different results. The original RT-PCR table labelled genes changed at 1 and 5% significance. We performed the comparison using a q-value of 5% as the cutoff for all genes, although a 1% q-value would not have changed the results. The table shows that PBP had slightly better agreement than the original MBEI-SAM method. Furthermore, when we used the probesets from the original analysis as opposed to the mapping of the reference sequence, we found that PBP does even better. Thus, PBP offers better validation with the RT-PCR results than

the MBEI-SAM combination. This demonstrates that PBP can perform well in experiments with biological variability and with thousands of genes present in a sample.

Table 6–2: RT-PCR comparison between MBEI-SAM and PBP

Algorithm	Found	RT-PCR	
		Changed	Unchanged
MBEI-SAM	Changed	21	15
	Unchanged	5	31
PBP	Changed	21	12
	Unchanged	5	34

6.2.3 Higher Level Analysis

The use of reference sequences to map probesets not only bypasses annotation issues but facilitates higher level analysis. In comparison to the canine proteome, the human proteome is more fully annotated. Databases like the Human Protein Reference Database [149] contain information on over 25,000 proteins including interactions, domains, and functions. Moreover, there exists annotated pathways which show how human proteins interact in response to certain stimuli. The HPRD database has 20 pathways with lists of genes that are found differentially expressed when the pathways are utilized. To access this information, we need a way to match canine genes to human genes. Fortunately, NCBI has a Homologene database [178] which contains clusters of homologous transcripts. We were able to match 50% of the changed transcripts with human equivalents.

With the list of differentially expressed genes for the pathways, we could use a hyper-geometric distribution to calculate the probability of finding x genes from a comparison like 1 week ATP in the pathway's list. Based on this comparison,

only two pathways were found significant (p-value $< .05$) and only in the 24 hour VTP model. Interestingly, one of these pathways was the TNF-alpha signaling pathway, which utilizes MAP kinase and capase pathways to cause apoptosis [149]. This is encouraging because apoptosis was only found in the VTP canine model and only at the early stage, around 24 hours (see Chapter 1). The other pathway found changed at 24 hour VTP was epidermal growth factor receptor signaling, which activates Ras/Raf/MAP kinase signaling. The 24 hour VTP model shows increased activity of MAP kinase and other signaling molecules, like ERK, in contrast to the 2 week VTP model. This pathway analysis is therefore in concordance with the known biology and also reveals targets for future research.

This pathway analysis is not fully developed but demonstrates what can be done with an algorithm, like PBP, that solves the annotation problems. In Chapter 3, the classification process was tedious, requiring the use of BLAST on thousands of sequences along with human supervision to interpret the results. Many of the protein databases, like HPRD [149], use the NCBI reference sequence identification [175]. This makes it straightforward to link the microarray results with the information in the protein databases. Besides interaction information, the protein databases have annotations with information like functional classification and cellular localization. Thus, the process of classifying the genes can be automated. In future work, we could look at the differentially regulated classes of genes to better understand which functional mechanisms characterize diseases. Similarly, we could look for differentially regulated domains on proteins. Preliminary work uncovered cytoskeletal structure motifs upregulated in the VTP models. Although this higher level analysis has been done before [179], PBP can increase the reliability and speed of analysis by avoiding the difficulty of annotations.

CHAPTER 7
Conclusions

In Chapter 2, I demonstrated that genetic networks can perform Bayesian inference to make decisions in noisy environments. This presents a new paradigm for understanding the design of genetic networks and explains previously puzzling experimental data. Parameter sensitivity analysis showed that this system can easily adapt to new inference problems. I also found that the type of inference performed dictates whether a repressor or an activator is preferable. Thus, the classification problem can determine the type of genetic regulation. Future work could investigate how genetic inference modules can be linked together to classify complex environments.

Moving from theoretical models of genetic networks to *in vivo* networks, I analyzed microarray data by comparing two canine models of atrial fibrillation and their evolution over two time points. I showed that the number of changed genes in the ventricular tachypacing model is an order of magnitude greater than the number found in the atrial tachypacing model. The atrial tachypacing model had a large amount of overlap in changed genes between the 24 hours and 1 week time points, showing an adaptation. An additional experimental group of atrial tachypacing at 6 weeks found no genes changed and confirmed this hypothesis. In contrast, the ventricular tachypacing model showed similar numbers of genes differentially expressed at both time points with little overlap, pointing to separate mechanisms evolving over time. I found that, of the differentially expressed genes, the extracellular matrix and cell structure genes showed the greatest change in the

ventricular tachypacing model, thus matching histology with gene expression. A protein interaction diagram based on the microarray results highlighted connective tissue growth factor as a putative mediator responsible for the fibrosis found in the ventricular tachypacing model.

This work is important because it shows that gene expression is a major determinant in the ultimate pathology of a common cardiac arrhythmia. The adaptation in the atrial tachypacing model appears at the genetic level and explains why the phenotype can disappear after tachypacing has been discontinued. With this work, the fibrosis in the ventricular tachypacing model can be traced to increased genetic regulation of extracellular matrix components. Furthermore, the early (24 hour) and late (2 week) responses are large in magnitude and employ different mechanisms. By investigating the protein interaction network in this experiment, I have found future targets for study that may explain what triggers these different mechanisms.

Based on problems encountered in the analysis of the canine microarray data (see Chapter 4), I proposed a new algorithm for detecting differential gene expression. This algorithm outperformed existing methods on a validation data set (Chapter 5). To understand why this algorithm does better on microarray data, I generalized the problem of detecting differential expression to an evaluation problem. Within this wider context, I found that the distribution of the judges' scores determines which algorithm is better: sum rule or majority rule. I then applied a cost analysis to these evaluation procedures and showed that the ratio of a cost of a wrong decision to the cost of a judge determines the optimal number of judges. These results have implications for improving judging procedures in a vast array of fields including grant reviews and figure skating.

In Chapter 5, I explored the simple case where all judges were equally accurate. In future work, it would be interesting to see how these results are influenced by panels of mixed accuracy. There may be a relationship between the decision rule and the distribution of the accuracies of judges. Moreover, I found that increasing the number of equally accurate judges improves the panel's accuracy but more complex behavior can occur when less accurate judges are added. In some cases, adding less accurate judges detracts from the panel's accuracy while in other cases it improved the decision. This tradeoff between the benefit of the experts and the benefit of the many is worth future study.

Finally, I reanalyzed the canine data from Chapter 3 using the algorithm from Chapter 5. The new algorithm solved problems found in other microarray analysis methods such as unclear transformations and annotation complications. I remapped the probes to known sequences and found that it outperformed the MBEI-SAM method used in Chapter 3 on the RT-PCR confirmations. I extended the analysis further and found two pathways activated in the 24 hour VTP condition. Thus, I showed that PBP can function on "real" microarray data and outperform existing methods. Furthermore, PBP permits higher level analysis that can select pathways for future experiments. This work is important because it introduces an effective algorithm that is easy to understand. PBP has the potential to improve microarray technology by returning to the level of individual probes. It also encourages experimentalists to participate in the analysis of microarray data, which may improve the understanding of microarray results.

REFERENCES

[1] E. S. Lander. Array of hope. *Nat Genet*, 21(1 Suppl):3–4, Jan 1999.

[2] D. E. Bassett, M. B. Eisen, and M. S. Boguski. Gene expression informatics–it's all in your mine. *Nat Genet*, 21(1 Suppl):51–55, Jan 1999.

[3] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.

[4] E. S. Mansfield, J. M. Worley, S. E. McKenzie, S. Surrey, E. Rappaport, and P. Fortina. Nucleic acid detection using non-radioactive labelling methods. *Mol Cell Probes*, 9(3):145–156, Jun 1995.

[5] P. C. Hayes, C. R. Wolf, and J. D. Hayes. Blotting techniques for the study of DNA, RNA, and proteins. *BMJ*, 299(6705):965–968, Oct 1989.

[6] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21(1 Suppl):33–37, Jan 1999.

[7] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, Jan 1999.

[8] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

[10] C. Debouck and P. N. Goodfellow. DNA microarrays in drug discovery and development. *Nat Genet*, 21(1 Suppl):48–50, Jan 1999.

[11] M. J. Marton, J. L. DeRisi, H. A. Bennett, V. R. Iyer, M. R. Meyer, C. J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, D. E. Bassett, L. H. Hartwell, P. O. Brown, and S. H. Friend. Drug target validation and

identification of secondary drug target effects using DNA microarrays. *Nat Med*, 4(11):1293–1301, Nov 1998.

[12] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

[13] E. Han, Y. Wu, R. McCarter, J. F. Nelson, A. Richardson, and S. G. Hilsenbeck. Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *J Gerontol A Biol Sci Med Sci*, 59(4):306–315, Apr 2004.

[14] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, Jan 2006.

[15] R. D. Canales, Y. L., J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol*, 24(9):1115–1122, Sep 2006.

[16] J. Seo and E. P. Hoffman. Probe set algorithms: is there a rational best bet? *BMC Bioinformatics*, 7:395, 2006.

[17] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[18] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[19] F. Naef, D. A. Lim, N. Patil, and M. Magnasco. DNA hybridization to mismatched templates: a chip study. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65(4 Pt 1):040902, Apr 2002.

[20] F. F. Millenaar, J. Okyere, S. T. May, M. van Zanten, L. A. C. J. Voesenek, and A. J. M. Peeters. How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7:137, 2006.

[21] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada. The Affymetrix GeneChip platform: an overview. *Methods Enzymol*, 410:3–28, 2006.

[22] Affymetrix. *GeneChip Expression Analysis: Technical Manual*, 2004.

[23] S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol*, 6(2):R16, 2005.

[24] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[25] R. Hoffmann, T. Seidl, and M. Dugas. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol*, 3(7):RESEARCH0033, Jun 2002.

[26] B. Harr and C. Schltterer. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res*, 34(2):e8, 2006.

[27] L. Qin, R. P. Beyer, F. N. Hudson, N. J. Linford, D. E. Morris, and K. F. Kerr. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics*, 7:23, 2006.

[28] Affymetrix. *Statistical Algorithms Description Document*, 2002.

[29] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31–36, 2001.

[30] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*, 2:0032.1–0032.11, 2001.

[31] A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol*, 2(12):RESEARCH0055, 2001.

[32] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*, Suppl 37:120–125, 2001.

[33] J. H. Do and D. Choi. Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol Cells*, 22(3):254–261, Dec 2006.

[34] Affymetrix. *GeneChip Expression Analysis: Data Analysis Fundamentals*, 2004.

[35] L. Xu, G. A. Maresh, J. Giardina, and S. H. Pincus. Comparison of different microarray data analysis programs and description of a database for microarray data management. *DNA Cell Biol*, 23(10):643–651, Oct 2004.

[36] S. Draghici. Statistical intelligence: effective analysis of high-density microarray data. *Drug Discov Today*, 7(11 Suppl):S55–S63, Jun 2002.

[37] D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet*, 32 Suppl:502–508, Dec 2002.

[38] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends Genet*, 18(5):265–271, May 2002.

[39] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98:5116–5121, 2001.

[40] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell, and J. K. Lee. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 19(15):1945–1951, Oct 2003.

[41] W. Liu, R. Li, J. Z. Sun, J. Wang, J. Tsai, W. Wen, A. Kohlmann, and P. M. Williams. PQN and DQN: algorithms for expression microarrays. *J Theor Biol*, 243(2):273–278, Nov 2006.

[42] L. Zhou and D. M. Rocke. An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics*, 21(21):3983–3989, Nov 2005.

[43] S. C. Geller, J. P. Gregg, P. Hagerman, and D. M. Rocke. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19(14):1817–1823, Sep 2003.

[44] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, Jun 2001.

[45] A.K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green. BGX: a fully Bayesian integrated approach to the analysis of Affymetrix geneChip data. *Biostatistics*, 6(3):349–373, Jul 2005.

[46] Affymetrix Latin Square Data. `www.affymetrix.com/support/technical/sample_data/datasets.affx` (accessed Oct. 11, 2006).

[47] K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K. R. Cho, T. J. Giordano, S. B. Gruber, E. R. Fearon, J. M. G. Taylor, and S. Hanash. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics*, 6:26, 2005.

[48] W. J. Lemon, J. J. T. Palatini, R. Krahe, and F. A. Wright. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, 18(11):1470–1476, Nov 2002.

[49] E. J. Benjamin, P. A. Wolf, R. B. D'Agostino, H. Silbershatz, W. B. Kannel, and D. Levy. Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation*, 98(10):946–952, Sep 1998.

[50] S. Nattel. New ideas about atrial fibrillation 50 years on. *Nature*, 415(6868):219–226, Jan 2002.

[51] S. Nattel and L. H. Opie. Controversies in atrial fibrillation. *Lancet*, 367(9506):262–272, Jan 2006.

[52] S. Nattel, A. Shiroshita-Takeshita, B. J. J. M. Brundel, and L. Rivard. Mechanisms of atrial fibrillation: lessons from animal models. *Prog Cardiovasc Dis*, 48(1):9–28, 2005.

[53] M. C. Wijffels, C. J. Kirchhof, R. Dorland, and M. A. Allessie. Atrial fibrillation begets atrial fibrillation. A study in awake chronically instrumented goats. *Circulation*, 92(7):1954–1968, Oct 1995.

[54] D. Li, S. Fareh, T. K. Leung, and S. Nattel. Promotion of atrial fibrillation by heart failure in dogs: atrial remodeling of a different sort. *Circulation*, 100(1):87–95, Jul 1999.

[55] B. J. J. M. Brundel, R. H. Henning, H. H. Kampinga, I. C. Van Gelder, and H. J. G. M. Crijns. Molecular mechanisms of remodeling in human atrial fibrillation. *Cardiovasc Res*, 54(2):315–324, May 2002.

[56] L. Yue, P. Melnyk, R. Gaspo, Z. Wang, and S. Nattel. Molecular mechanisms underlying ionic remodeling in a dog model of atrial fibrillation. *Circ Res*, 84(7):776–784, Apr 1999.

[57] B.J. Brundel, J. Ausma, I.C. van Gelder, J.J. Van der Want, W.H. van Gilst, H.J. Crijns, and R.H. Henning. Activation of proteolysis by calpains and structural changes in human paroxysmal and persistent atrial fibrillation. *Cardiovascular Research*, 54:380–389, 2002.

[58] C. A. Morillo, G. J. Klein, D. L. Jones, and C. M. Guiraudon. Chronic rapid atrial pacing. Structural, functional, and electrophysiological characteristics of a new model of sustained atrial fibrillation. *Circulation*, 91(5):1588–1595, Mar 1995.

[59] R. Gaspo, R. F. Bosch, M. Talajic, and S. Nattel. Functional mechanisms underlying tachycardia-induced sustained atrial fibrillation in a chronic dog model. *Circulation*, 96(11):4027–4035, Dec 1997.

[60] J. R. Wilson, P. Douglas, W. F. Hickey, V. Lanoce, N. Ferraro, A. Muhammad, and N. Reichek. Experimental congestive heart failure produced by rapid ventricular pacing in the dog: cardiac effects. *Circulation*, 75(4):857–867, Apr 1987.

[61] P. W. Armstrong, T. P. Stopps, S. E. Ford, and A. J. de Bold. Rapid ventricular pacing in the dog: pathophysiologic studies of heart failure. *Circulation*, 74(5):1075–1084, Nov 1986.

[62] S. Nattel, A. Shiroshita-Takeshita, S. Cardin, and P. Pelletier. Mechanisms of atrial remodeling and clinical relevance. *Curr Opin Cardiol*, 20(1):21–25, Jan 2005.

[63] T. Cha, J. R. Ehrlich, L. Zhang, and S. Nattel. Atrial ionic remodeling induced by atrial tachycardia in the presence of congestive heart failure. *Circulation*, 110(12):1520–1526, Sep 2004.

[64] T. Cha, J. R. Ehrlich, L. Zhang, Y. Shi, J. Tardif, T. K. Leung, and S. Nattel. Dissociation between ionic remodeling and ability to sustain atrial fibrillation during recovery from experimental congestive heart failure. *Circulation*, 109(3):412–418, Jan 2004.

[65] N. Hanna, S. Cardin, T. Leung, and S. Nattel. Differences in atrial versus ventricular remodeling in dogs with ventricular tachypacing-induced congestive heart failure. *Cardiovasc Res*, 63(2):236–244, Aug 2004.

[66] S. Cardin, D. Li, N. Thorin-Trescases, T. Leung, E. Thorin, and S. Nattel. Evolution of the atrial fibrillation substrate in experimental congestive heart

failure: angiotensin-dependent and -independent pathways. *Cardiovasc Res*, 60(2):315–325, Nov 2003.

[67] K. Shinagawa, Y. Shi, J. Tardif, T. Leung, and S. Nattel. Dynamic nature of atrial fibrillation substrate during development and reversal of heart failure in dogs. *Circulation*, 105(22):2672–2678, Jun 2002.

[68] G. Lamirault, N. Gaborit, N. Le Meur, C. Chevalier, G. Lande, S. Demolombe, D. Escande, S. Nattel, J. J. Lger, and M. Steenman. Gene expression profile associated with chronic atrial fibrillation and underlying valvular heart disease in man. *J Mol Cell Cardiol*, 40(1):173–184, Jan 2006.

[69] A.S. Barth, S. Merk, E. Arnoldi, L. Zwermann, P. Kloos, M. Gebauer, K. Steinmeyer, M. Bleich, S. Kaab, M. Hinterseer, H. Kartmann, E. Kreuzer, M. Dugas, G. Steinbeck, and M. Nabauer. Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature. *Circulation Research*, 96:1022–1029, 2005.

[70] L. Lai, J. Lin, C. Lin, H. Yeh, Y. Tsay, C. Lee, H. Lee, Z. Chang, J. Hwang, M. Su, Y. Tseng, and S. K. S. Huang. Functional genomic study on atrial fibrillation using cDNA microarray and two-dimensional protein electrophoresis techniques and identification of the myosin regulatory light chain isoform reprogramming in atrial fibrillation. *J Cardiovasc Electrophysiol*, 15(2):214–223, Feb 2004.

[71] Y. H. Kim, D. S. Lim, J. H. Lee, D. Lim, W. J. Shim, Y. M. Ro, G. H. Park, K. G. Becker, Y. S. Cho-Chung, and M. Kim. Gene expression profiling of oxidative stress on atrial fibrillation in humans. *Exp Mol Med*, 35(5):336–349, Oct 2003.

[72] R. Ohki, K. Yamamoto, S. Ueno, H. Mano, Y. Misawa, K. Fuse, U. Ikeda, and K. Shimada. Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. *Int J Cardiol*, 102(2):233–238, Jul 2005.

[73] N. Kim, Y. Ahn, S. K. Oh, J. K. Cho, H. W. Park, Y. Kim, M. H. Hong, K. I. Nam, W. J. Park, M. H. Jeong, B. H. Ahn, J. B. Choi, H. Kook, J. C. Park, J. Jeong, and J. C. Kang. Altered patterns of gene expression in response to chronic atrial fibrillation. *Int Heart J*, 46(3):383–395, May 2005.

[74] M. Ptashne. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci*, 30(6):275–279, Jun 2005.

[75] S. Istrail and E. H. Davidson. Logic functions of the genomic cis-regulatory code. *Proc Natl Acad Sci U S A*, 102(14):4954–4959, Apr 2005.

[76] M. Lewis. The lac repressor. *C R Biol*, 328(6):521–548, Jun 2005.

[77] C. J. Wilson, H. Zhan, L. Swint-Kruse, and K. S. Matthews. The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cell Mol Life Sci*, 64(1):3–16, Jan 2007.

[78] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

[79] M. Santillan and M. C. Mackey. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. *Biophys J*, 86(3):1282–1292, Mar 2004.

[80] N. Yildirim and M. C. Mackey. Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophys J*, 84(5):2841–2851, May 2003.

[81] P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks–methods, recent results, and future directions. *Bull Math Biol*, 62(2):247–292, Mar 2000.

[82] H. Bolouri and E. H. Davidson. Modeling transcriptional regulatory networks. *Bioessays*, 24(12):1118–1129, Dec 2002.

[83] P. Wong, S. Gladney, and J. D. Keasling. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol Prog*, 13(2):132–143, 1997.

[84] J. M. G. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–989, Aug 2003.

[85] M. Santillan and M. C. Mackey. Why the lysogenic state of phage lambda is so stable: a mathematical modeling approach. *Biophys J*, 86(1 Pt 1):75–84, Jan 2004.

[86] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–1133, Feb 1982.

[87] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks–II. Immunity control in bacteriophage lambda. *Bull Math Biol*, 57(2):277–297, Mar 1995.

[88] R. Edwards and L. Glass. Combinatorial explosion in model gene networks. *Chaos*, 10(3):691–704, Sep 2000.

[89] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, 22(3):437–467, Mar 1969.

[90] L. Glass. Classification of biological networks by their qualitative dynamics. *J Theor Biol*, 54(1):85–107, Oct 1975.

[91] P. S. Swain. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol*, 344(4):965–976, Dec 2004.

[92] J. J. Tyson, K. C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol*, 15(2):221–231, Apr 2003.

[93] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. Van Oudenaarden. Multistability in the lactose utilization network of Escherichia coli. *Nature*, 427(6976):737–740, Feb 2004.

[94] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–443, Nov 2004.

[95] J. Monod, J. Wyman, and J. P. Changeux. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12:88–118, 1965.

[96] J. Changeux and S. J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, Jun 2005.

[97] D. E. Koshland, G. Nmethy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, Jan 1966.

[98] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–464, Jun 2005.

[99] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, Aug 2002.

[100] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1):69–73, May 2002.

[101] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, Apr 2003.

[102] T. S. Gardner and J. J. Collins. Neutralizing noise in gene networks. *Nature*, 405(6786):520–521, Jun 2000.

[103] M. Thattai and A. van Oudenaarden. Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530, May 2004.

[104] D. Orrell and H. Bolouri. Control of internal and external noise in genetic regulatory networks. *J Theor Biol*, 230(3):301–312, Oct 2004.

[105] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, Jun 2000.

[106] C. V. Rao, D. M. Wolf, and A. P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237, Nov 2002.

[107] N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387(6636):913–917, Jun 1997.

[108] M. Thattai and A. van Oudenaarden. Attenuation of noise in ultrasensitive signaling cascades. *Biophys J*, 82(6):2943–2950, Jun 2002.

[109] P. B. Detwiler, S. Ramanathan, A. Sengupta, and B. I. Shraiman. Engineering aspects of enzymatic signal transduction: photoreceptors in the retina. *Biophys J*, 79(6):2801–2817, Dec 2000.

[110] S. Hooshangi, S. Thiberge, and R. Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc Natl Acad Sci U S A*, 102(10):3581–3586, Mar 2005.

[111] L. H. Hartwell and T. A. Weinert. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246(4930):629–634, Nov 1989.

[112] P. S. Swain and E. D. Siggia. The role of proofreading in signal transduction specificity. *Biophys J*, 82(6):2928–2933, Jun 2002.

[113] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat. Genet.*, 31:69–73, 2002.

[114] J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811–1814, 2004.

[115] D. J. C. Mackay. *Information theory, inference, and learning algorithms.* Cambridge University Press, New York, New York, 2003.

[116] E. Dekel and U. Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436:588–592, 2005.

[117] M. A. Savageau. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of Escherichia coli. *Genetics.*, 149:1677–1691, 1998.

[118] A. L. Koch. The protein burden of lac operon products. *J. Mol. Evol.*, 19:455–462, 1983.

[119] R. T. Cox. Probability, frequency, and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1946.

[120] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci. U. S. A.*, 100:7702–7707, 2003.

[121] M. Ptashne and A. Gann. *Genes and Signals.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002.

[122] R. S. Makman and E. W. Sutherland. Adenosine 3',5'-phosphate in Escherichia coli. *J. Biol. Chem.*, 240:1309–1314, 1965.

[123] R. Thomas. Boolean formalization of genetic control circuits. *J. Theor Biol.*, 42:563–585, 1973.

[124] L. Glass. Combinatorial and topological methods in non-linear chemical kinetics. *J. Chem. Phys.*, 63:1325–1335, 1975.

[125] C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.

[126] J. Mukhopadhyay, R. Sur, and P. Parrack. Functional roles of the two cyclic AMP-dependent forms of cyclic AMP receptor protein from Escherichia coli. *FEBS. Lett.*, 453:215–218, 1999.

[127] A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proc. Nat. Acad. Sci. U. S. A.*, 43:553–566, 1957.

[128] J. E. Ferrell. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.*, 14:140–148, 2002.

[129] M. A. Shea and G. K. Ackers. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.*, 181:211–230, 1985.

[130] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[131] K. S. Brown and Sethna J. P. Statistical mechanics approaches to models with many poorly known parameters. *Phys. Rev. E*, 68:21904, 2003.

[132] G. Schwartz. Estimating the dimensions of a model. *Ann. Stat.*, 6:461–464, 1978.

[133] J. Stelling, E. D. Gilles, and F. J. Doyle. Robustness properties of circadian clock architectures. *Proc. Natl. Acad. Sci. U. S. A.*, 101:13210–13215, 2004.

[134] M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A*, 104:1876–1889, 2000.

[135] J.S. Steinberg. Atrial fibrillation: an emerging epidemic? *Heart*, 90:239–240, 2004.

[136] M.A. Allessie, P.A. Boyden, A.J. Camm, A.G. Kleber, M.J. Lab, M.J. Legato, M.R. Rosen, P.J. Schwartz, P.M. Spooner, D.R. Van Wagoner, and A.L. Waldo. Pathophysiology and prevention of atrial fibrillation. *Circulation*, 103:769–777, 2001.

[137] A. Elvan, K. Wylie, and D.P. Zipes. Pacing-induced chronic atrial fibrillation impairs sinus node function in dogs. electrophysiological remodeling. *Circulation*, 94:2953–2960, 1996.

[138] L. Yue, J. Feng, R. Gaspo, G.R. Li, Z. Wang, and S. Nattel. Ionic remodeling underlying action potential changes in a canine model of atrial fibrillation. *Circulation Research*, 81:512–525, 1997.

[139] D. Dobrev and U. Ravens. Remodeling of cardiomyocyte ion channels in human atrial fibrillation. *Basic Res Cardiol*, 98:137–148, 2003.

[140] D. Li, P. Melnyk, J. Feng, Z. Wang, K. Petrecca, A. Shrier, and S. Nattel. Effects of experimental heart failure on atrial cellular and ionic electrophysiology. *Circulation*, 101:2631–2638, 2000.

[141] R. Ohki, K. Yamamoto, S. Ueno, H. Mano, Y. Misawa, K. Fuse, U. Ikeda, and K. Shimada. Transcriptional profile of genes induced in human atrial myocardium with pressure overload. *Int J Cardiol*, 96(3):381–387, Sep 2004.

[142] N. Gaborit, M. Steenman, G. Lamirault, N. Le Meur, S. Le Bouter, G. Lande, J. Leger, F. Charpentier, T. Christ, D. Dobrev, D. Escande, S. Nattel, and S. Demolombe. Human atrial ion channel and transporter subunit gene-expression remodeling associated with valvular heart disease and atrial fibrillation. *Circulation*, 112:471–497, 2005.

[143] L. Yue, J. Feng, G.R. Li, and S. Nattel. Transient outward and delayed rectifier currents in canine atrium: properties and role of isolation methods. *Am J Physiol*, 270:H2157–H2168, 1996.

[144] D. Li, L. Zhang, J. Kneller, and S. Nattel. Potential ionic mechanism for repolarization differences between canine right and left atrium. *Circ Res*, 88:1168–1175, 2001.

[145] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[146] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[147] V.L. Thijssen, H.M. van der Velden, E.P. van Ankeren, J. Ausma, M.A. Allessie, M. Borgers, G.J. van Eys, and H.J. Jongsma. Analysis of altered gene expression during sustained atrial fibrillation in the goat. *Cardiovasc Res*, 54:427–437, 2002.

[148] S. Verheule, T. Sato, T. 4th Everett, D. Otten, M. Rubart-von der Lohe, H.O. Nakajima, H. Nakajima, L.J. Field, and J.E. Olgin. Increased vulnerability to atrial fibrillation in transgenic mice with selective atrial fibrosis caused by overexpression of TGF-beta1. *Circ Res*, 94:1458–1465, 2004.

[149] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T.K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H.N. Shivashankar, B.P. Rashmi, M.A. Ramya, Z. Zhao, K.N. Chandrika, N. Padma, H.C. Harsha, A.J. Yatish, M.P. Kavitha, M. Menezes, D.R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S.K. Anand, V. Madavan, A. Joseph, G.W. Wong, W.P. Schiemann, S.N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G.C. Blobe, C.V. Dang, J.G. Garcia, J. Pevsner, O.N. Jensen, P. Roepstorff, K.S. Deshpande, A.M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363–2371, 2003.

[150] B. Liu, J. Yu, L. Taylor, X. Zhou, and P. Polgar. Microarray and phosphokinase screenings leading to studies on ERK and JNK regulation of connective tissue growth factor expression by angiotensin II 1a and bradykinin B2 receptors in Rat1 fibroblasts. *J Cell Biochem*, 97:1104–1120, 2006.

[151] J.J. Mulsow, R.W. Watson, J.M. Fitzpatrick, and P.R. O'Connell. Transforming growth factor-beta promotes pro-fibrotic behavior by serosal fibroblasts via PKC and ERK1/2 mitogen activated protein kinase cell signaling. *Ann Surg*, 2005:880–887, 2005.

[152] C. Ott, D. Iwanciw, A. Graness, K. Giehl, and M. Goppelt-Struebe. Modulation of the expression of connective tissue growth factor by alterations of the cytoskeleton. *J Biol Chem*, 278:44305–44311, 2003.

[153] N.A. Wahab, B.S. Weston, and R.M. Mason. Modulation of the TGF-beta/Smad signaling pathway in mesangial cells by CTGF/CCN2. *Exp Cell Res*, 307:305–314, 2005.

[154] J.K. Crean, D. Finlay, M. Murphy, C. Moss, C. Godson, F. Martin, and H.R. Brady. The role of p42/44 mapk and protein kinase b in connective tissue growth factor induced extracellular matrix protein production, cell migration, and actin cytoskeletal rearrangement in human mesangial cells. *J Biol Chem*, 277:44187–44194, 2002.

[155] T. Christ, P. Boknik, S. Wohrl, E. Wettwer, E.M. Graf, R.F. Bosch, M. Knaut, W. Schmitz, U. Ravens, and D. Dobrev. L-type ca2+ current downregulation in chronic human atrial fibrillation associated with increased activity of protein phosphatases. *Circulation*, 110:2651–2657, 2004.

[156] C. L. Dodgson. *The Pamphlets of Lewis Carroll, Vol. 3. The Political Pamphlets and Letters of Charles Lutwidge Dodgson and Related Pieces. A Mathematical Approach. Reprint of 'A method of taking votes on more than two issues' by C. L. Dodgson (1876).* University Press of Virginia, Charlottesville, 2001.

[157] R.M. May. Electoral procedures: Preference and paradox. *Nature*, 303:16–17, 1983.

[158] K. J. Arrow. *Social Choice and Individual Values.* Yale University Press, New Haven, 1963.

[159] Marquis de Condorcet, J.-A.-N. de C. *Essai sur l'application de l'analyse a la probabilit des decisions redues a la pluralit de voix.* De L'Imprimerie Royale, Paris, 1785.

[160] F. Harary and L. Moser. The theory of round robin tournaments. *Am. Math. Monthly*, 73:231–246, 1966.

[161] International Skating Union. `http://www.isu.org` (accessed Oct. 11, 2006).

[162] C. Seife. New skating system fails virtual replay. *Science*, 299:651, 2003.

[163] J. Surowiecki. *The Wisdom of Crowds.* Doubleday, New York, 2004.

[164] R. Hastie and T. Kameda. The robust beauty of majority rules in group decisions. *Psychol. Rev.*, 112:494–508, 2005.

[165] R. D. Sorkin, R. West, and D. E. Robinson. Group performance depends on the majority rule. *Psychol. Sci.*, 9:456–463, 1998.

[166] R. P. Larrick and J. B. Soll. Intuitions about combining opinions: Misappreciation of the averaging principle. *Manage. Sci.*, 52:111–127, 2006.

[167] U.S. Figure Skating. `http://www.usfigureskating.org/event_details.asp?id=26227` (accessed Oct. 11, 2006).

[168] J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[169] Northwestern University's Medill Journalism Supreme Court listings. `http://docket.medill.northwestern.edu/archives/002315.php` (accessed Nov. 9, 2006).

[170] World Boxing Association. `http://www.wbaonline.com/ratings/results/default.asp?` (accessed Nov. 6, 2006).

[171] Canadian Institutes for Health Research. `http://www.cihr-irsc.gc.ca/e/23467.html#7` (accessed Nov. 6, 2006).

[172] Office of Personnel Management. `http://www.opm.gov/news_events/congress/testimony/5_16_2006.asp` (accessed Nov. 9, 2006).

[173] U.S. Supreme Court. `http://www.supremecourtus.gov/about/about.html` (accessed Nov. 6, 2006).

[174] C. Thomassen. Hamiltonian-connected tournaments. *J. Combinatorial Theory Ser. B*, 28:142 –163, 1980.

[175] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–D504, Jan 2005.

[176] J. Harbig, R. Sprinkle, and S. A. Enkemann. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*, 33(3):e31, 2005.

[177] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful appraoch to multiple testing. *J R Statist Soc B*, 57:289–300, 1995.

[178] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35(Database issue):D5–12, Jan 2007.

[179] R. K. Curtis, M. Oresic, and A. Vidal-Puig. Pathways to the analysis of microarray data. *Trends Biotechnol*, 23(8):429–435, Aug 2005.