DNA binding economies

by

Luis Pérez-Breva

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2007

© Massachusetts Institute of Technology 2007. All rights reserved.
DNA binding economies

by

Luis Pérez-Breva

Submitted to the Department of Electrical Engineering and Computer Science on August 31st, 2007, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

This thesis develops a new scalable modeling framework at the interface of game theory and machine learning to recover economic structures from limited slices of data. Inference using economic models has broad applicability in machine learning. Economic structures underlie a surprisingly broad array of problems including signaling and molecular control in biology, drug development, neural structures, distributed control, recommender problems, social networking, as well as market dynamics.

We demonstrate the framework with an application to genetic regulation. Genetic regulation determines how DNA is read and interpreted, is responsible for cell specialization, reaction to drugs, metabolism, etc. Improved understanding of regulation has potential to impact research on genetic diseases including cancer.

Genetic regulation relies on coordinate binding of regulators along DNA. Understanding how binding arrangements are achieved and their effect on regulation is challenging since it is not always possible to study regulatory processes in isolation. Indeed, observing the action of regulators is an experimental and computational challenge. We need causal genome-wide models that can work with existing high-throughput observations.

We abstract DNA binding as an economy and develop fast algorithms to predict average binding arrangements as competitive equilibria. The framework supports viewing regulation as a succession of regulatory states. We complete the framework with algorithms to infer causal structure from high-throughput observations. Learning here deviates from work in learning in games, it is closer to the economic theory of revealed preferences. Our algorithms predict the effect of experimental perturbations and can be used to refine experimental hypotheses.

We show that the economic approach reproduces known behavior of a genetic switch (λ-phage), and that it can complete the map of coordinate binding in yeast.

Thesis Supervisor: Tommi Jaakkola
Title: Associate Professor
Acknowledgments

I first arrived at MIT in 1998 as a MS in Chemical Engineering, for a one-year research stay. I have since lived in 3 countries, I joined a startup in Silicon Valley, learned about innovation in business, got a Master degree in quantum physics and completed a PhD thesis at the interface of computation, economic theory and molecular biology. This is, arguably, a non obvious combination of backgrounds. It has, nonetheless, been an incredible prelude to the journey that starts at the end of this PhD. I am indebted to many people that have made this prelude all the more interesting.

I would not have gotten here if back in 2004, Tommi Jaakkola had not appreciated some potential in my colorful background. I am indebted to him for giving me the chance and support to undertake the cross-disciplinary research that has come to fruition in this thesis. Tommi’s enthusiasm was instrumental to look beyond the risks of engaging into cross-disciplinary work, and appreciate the subtleties shared by different fields. In hindsight, through many conversations ranging from caffeine content in coffee to abstracting scientific intuitions mathematically, Tommi has taught me how to weave the pieces of my background together formally. He has improved me as a researcher. I look forward to our collaboration to extend this work to new challenges in computational biology, medicine, and a variety of fields.

I am grateful to my committee, Asuman Ozdaglar and Tomaso Poggio, for finding time in July and August to look over the thesis draft and manage to clear their schedule to make my thesis dissertation happen. Theirs and Manolis Kellis insightful feedback helped me improve many fine points in this thesis.

I had the honor to interact with Luis E. Ortiz while he was a postdoctoral lecturer at MIT. He patiently taught me everything I know about game theory and economics. Many of the ideas in this thesis stem from our conversations, this thesis would not have been possible without his advice and collaboration. I look forward to continuing working with him in the future.

I am most grateful to the “Fundación Rafael del Pino” for their financial and moral support during the last two years of my PhD.
My PhD journey started in February 2001, and has had many episodes. At some point in 2003 I decided I wanted to use my PhD as an opportunity to learn more about what it takes to do research, understanding how to find funds, contrast my ideas with people beyond the research world, learn about international research, dare to explore new research avenues and collaborations, and, in hindsight, learn to be wrong most of the time. I have succeeded in most of these goals, I am grateful to MIT for encouraging my craving for exploration. None of this would have been possible without the mentoring of Tomaso Poggio. Tomaso Poggio gave me my first opportunity at MIT in 1998, convinced me to apply for the PhD program, and took me as a student in the Center for Biological and Computational Learning at the beginning of my PhD; he understood my leaving the group when in 2003 I decided to set on my exploration path, and after my return to MIT in 2004, he has been an invaluable source of advice and wisdom. His patience and mentoring have been an inspiration to me both in research and in the interaction with students.

In my years at MIT I have had the chance to interact with various MIT faculty members. I have struggled to find time to collaborate with Gerald Sussman in an inspiring project at the interface of computation and quantum mechanics, and I have learnt from his passion for science. I look forward to this collaboration as a long term project. I had the privilege to teach next to Paul Penfield Jr. and Seth Lloyd in a course that glues together information theory and the always elusive concept of entropy. I enjoyed every minute of my experience teaching Artificial Intelligence next to Patrick Winston. While not part of the MIT faculty, Marilyn A. Pierce from EECS headquarters has probably been the single most important reference to navigate around MIT and feel at home. Thank you, Marilyn.

Many other MIT faculty have generously shared their time with me. These include Victor Zue, Eric Grimson, Paul Schechter, and Bill Freeman, that were always extremely supportive and helped me get over my own mistakes. Nearing the end of my PhD, Ken Zolot and the faculty from the Deshpande Center of Technological Innovation opened my eyes to yet another facet of MIT at the interface of research and innovation. I am grateful to them for the many interesting conversations and for
letting me be a small part of it.

I should not like to forget Dr Enric Julia, director of Institut Químic de Sarrià, my Spanish alma mater, for he was instrumental in my decision to pursue the MIT experience back in 1998.

My MIT life has had a never ending supply of new friends that have enriched my PhD experience. There are simply too many good things to say about having Merche Balcells as a friend. Eduardo Torres-Jara has listened to my every random comment provided coffee kept coming. Gadi Geiger deserves special mention for throughout the years he has become the patient (sadistic?) single recipient of my every complaint about grad student life. Many more helped me through the PhD, listing them all would be impossible. I try to assemble their names in random order: Monday Okoro, Claire Monteleoni, Jason Rennie, John Barnett, Amir Globerson, Xavier Carreras, Antonio Torralba, Aude Oliva, Toni Valero, Stéphanie Molinard, Antoine F. Ezzat, Adlar Kim, Gabriel Kreimann, Sanmay Das, Stan Bileschi, Thomas Serre, Giorgos Zacharia, Jake Bouvrie, Theodoros Evgeniou, Marypat Fitzgerald, Bob Dressler, Manlio Allegra, Martin Feuerstein, Alex Rodriguez, Laura Martinez de Guerenu, Elsa Vila-Artadi, Albert Biosca, Israel Ruiz.

Above all I thank my parents and brothers, because they were the first to encourage me to come to MIT. Their support throughout these years of exploration has been unconditioned, even if in many cases (I dare confess now) I was not all that certain of the outcome.

My wife, Marta, has been next to me throughout my entire MIT experience; first as my long-distance girlfriend, and, happily for me, as my wife since 2005. I cannot imagine the suffering she has endured all these years seeing me work late nights and following closely all the ups and downs in my exploration, that were many. I am most impressed by her resilience remaining at my side, for while I have occasionally harvested the rewards of my exploration, she has had to endure most of the process as an spectator. I cannot but admire her for who she is, and thank her for her unconditional love. To her, this thesis and my life are dedicated.
## Contents

1 Introduction .................................................................................. 19
   1.1 Motivation .............................................................................. 19
   1.2 Approach ................................................................................ 21
     1.2.1 Evaluation ...................................................................... 23
   1.3 Contributions ......................................................................... 25
   1.4 Road Map ............................................................................... 27

2 Background: Genetic Regulation, DNA Binding, Economies .......... 29
   2.1 Biological Background: Genetic Regulation .......................... 30
      2.1.1 Transcriptional Regulation ........................................... 32
      2.1.2 DNA Binding as Transport and Binding ....................... 35
   2.2 Computational methods in Binding and Regulation ............... 36
      2.2.1 Mechanistic Approaches .............................................. 37
      2.2.2 Statistical and Probabilistic Approaches ....................... 38
      2.2.3 This Work .................................................................... 39
   2.3 Game Theory and Economic Theory .................................... 40

I Mathematical Abstraction ............................................................... 46

3 DNA binding – Economic Abstraction .......................................... 47
   3.1 Motivation .............................................................................. 48
   3.2 DNA Binding Abstraction ..................................................... 50
      3.2.1 Agents .......................................................................... 51
3.2.2 Protein-Agent: Transport Mechanism ........................................... 51
3.2.3 Site-Agent: the Binding Subproblem ........................................... 52
3.2.4 Competitive Equilibrium and Steady State Behavior ................. 52
3.3 Three DNA Binding Economies ......................................................... 53
3.4 A Site-Specific Economy ................................................................. 54
  3.4.1 Assumptions ................................................................. 54
  3.4.2 Utility Functions ............................................................. 55
  3.4.3 Fixed Point ................................................................. 59
3.5 A Promoter-Specific Economy ......................................................... 60
  3.5.1 Assumptions ................................................................. 61
  3.5.2 Formal Description .......................................................... 61
  3.5.3 Utility Functions ............................................................. 62
  3.5.4 Fixed Point ................................................................. 63
  3.5.5 Considerations about the Promoter Specific Economy .......... 64
3.6 A Nucleus-Specific Economy .......................................................... 65
  3.6.1 Assumptions ................................................................. 65
  3.6.2 Formal Description .......................................................... 66
  3.6.3 Utility functions ............................................................. 67
  3.6.4 Fixed Point ................................................................. 68
  3.6.5 Considerations about the Nucleus Specific Binding economy .. 69
3.7 Fixed Point Equations or Economy? ............................................... 69
  3.7.1 Economic Interpretation ...................................................... 70
  3.7.2 Advantage over Traditional Reaction-Equation models ......... 71
3.8 Summary ................................................................................. 73
  3.8.1 Conceptual Mapping ......................................................... 73
  3.8.2 Conventions ................................................................. 73

4 Predictions of the Economies ....................................................... 75
  4.1 Site Specific Economy ............................................................... 76
    4.1.1 Equilibrium Existence ..................................................... 76
5 Learning the Economic Structure

5.1 Previous Work on the Theory of Learning Applied to Game Theory and Economics
   5.1.1 Player Learning
   5.1.2 Revealed Preferences

5.2 Site-Specific DNA Binding Economy
   5.2.1 Set Up
   5.2.2 Algorithm

5.3 Nucleus-Specific DNA Binding Economy
   5.3.1 Set Up
   5.3.2 Algorithm

II Experimental results

6 Small Scale Experiments
   6.1 Forward Predictions with Parameters from Experiments
      6.1.1 Overview of λ-phage Infection
      6.1.2 Parameters for the λ-phage Site Specific Economy
      6.1.3 Reproducing λ-phage Lysogeny
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1.4 Empirical Complexity Analysis of the Site Specific Economy</td>
<td>121</td>
</tr>
<tr>
<td>6.1.5 Toy example: Promoter Specific Economy</td>
<td>123</td>
</tr>
<tr>
<td>6.2 Learning the Parameters</td>
<td>125</td>
</tr>
<tr>
<td>6.2.1 Learning a Site Specific Economy for ( \lambda )-phage Lysogeny</td>
<td>125</td>
</tr>
<tr>
<td>7 Large Scale Experiments</td>
<td>131</td>
</tr>
<tr>
<td>7.1 Scaling up: Learning a Nucleus Specific Economy for Yeast</td>
<td>132</td>
</tr>
<tr>
<td>7.2 Results</td>
<td>132</td>
</tr>
<tr>
<td>7.2.1 Reduced Set of Genes and Proteins</td>
<td>132</td>
</tr>
<tr>
<td>7.2.2 Extended Set of Genes and Proteins</td>
<td>134</td>
</tr>
<tr>
<td>III Conclusions</td>
<td>145</td>
</tr>
<tr>
<td>8 Extensions</td>
<td>147</td>
</tr>
<tr>
<td>8.1 Extensions Within Biology</td>
<td>148</td>
</tr>
<tr>
<td>8.1.1 Tissue Regulation and Connectivity Map</td>
<td>148</td>
</tr>
<tr>
<td>8.1.2 Augmenting the Economies with Sequence Information</td>
<td>149</td>
</tr>
<tr>
<td>8.1.3 Theoretical Extensions</td>
<td>151</td>
</tr>
<tr>
<td>8.1.4 Experimental Extensions</td>
<td>152</td>
</tr>
<tr>
<td>8.2 Extensions to Other Fields</td>
<td>152</td>
</tr>
<tr>
<td>9 Discussion</td>
<td>155</td>
</tr>
<tr>
<td>9.1 Mathematical Abstraction</td>
<td>155</td>
</tr>
<tr>
<td>9.2 Experimental Results</td>
<td>156</td>
</tr>
<tr>
<td>9.2.1 Small Scale Results</td>
<td>156</td>
</tr>
<tr>
<td>9.2.2 Large Scale Results</td>
<td>157</td>
</tr>
<tr>
<td>9.3 Advantages of the Economic Abstraction</td>
<td>159</td>
</tr>
<tr>
<td>9.3.1 Generalization to Other Problems</td>
<td>161</td>
</tr>
<tr>
<td>9.4 Summary of Contributions</td>
<td>161</td>
</tr>
</tbody>
</table>
## List of Figures

1-1 Depiction of the question addressed by DNA binding economies ... 21
2-1 Central Dogma of Molecular Biology ................................. 31
2-2 Watson’s version of the central dogma ............................... 31
2-3 Illustration of a genetic circuit ........................................ 38
3-1 Predictions of the DNA binding economies ............................ 48
3-2 Representation of the competitive equilibrium of a site specific economy 55
3-3 Bipartite Structure of the site specific DNA binding economy .............................. 58
3-4 Representation of the competitive equilibrium of a promoter specific economy ............................................................. 62
3-5 Representation of the competitive equilibrium of a nucleus specific economy ............................................................. 66
4-1 Exploiting the bipartite structure of the site specific DNA binding economy ............................................................. 78
4-2 Graphical illustration of the induction step in the proof of Theorem 4 80
5-1 Illustration of the learning problem ........................................ 93
6-1 Architecture of the λ-phage switch ...................................... 114
6-2 Levels of abstraction of the economy for λ-phage and the corresponding genetic circuit ............................................................. 116
6-3 Predicted behavior of λ-phage and comparison to stochastic simulation 119
6-4 Block diagram for predicting binding profiles in the site specific λ-phage economy .................................................. 120
6-5 Causal predictions with a DNA binding economy. λ-phage knockout ................................................................. 121
6-6 Empirical complexity analysis for the site specific economy .................................................................................. 122
6-7 Predictions on a promoter specific economy ................................................................................................. 124
6-8 Prediction of interdependence between DNA subsystems using a promoter specific economy ................................. 125
6-9 Analysis of bounding algorithm for the promoter specific economy ....................................................................... 126
6-10 Simulated λ-phage microarray measurements. ........................................................................................................ 128
6-11 Predicted binding in λ-phage with learnt parameters .......................................................................................... 130
6-12 Predicted transport in λ-phage with learn parameters ........................................................................................ 130
7-1 Structure of the nucleus specific in a representative sample selection of genes and regulators ............................... 133
7-2 Representation of the structure learnt Using the nucleus specific economy on an extended selection of genes and regulators ........................................................................................................ 139
7-3 Network Representation of a cluster of genes and regulators ................................................................................... 140
7-4 Predicted pair-wise complexes binding .................................................................................................................. 141
7-5 Ste12-Tec1 bindings across sites for each condition .................................................................................................. 142
7-6 Ste12-Tec1 bindings ..................................................................................................................................................... 142
7-7 Yap6-Cin5 bindings across sites for each condition .................................................................................................. 143
7-8 Yap6-Cin5 bindings ..................................................................................................................................................... 143
7-9 Skn7-Cin5 bindings across sites for each condition .................................................................................................. 144
7-10 Skn7-Cin5 bindings .................................................................................................................................................... 144
C-1 Block diagram for predicting binding profiles in the site specific λ-phage economy .................................................. 186
List of Tables

2.1 Game theory example: prisoner’s dilemma ................................. 42

3.1 Conceptual mapping from biology to economic theory. The mapping of protein-DNA binding onto an abstract economy requires the definition of the utilities of each type of agent and a proof that an equilibrium exists and can be computed. ......................................................... 73

3.2 Notation ................................................................. 74

5.1 Notation ................................................................. 104

6.1 Qualitative affinities (lysogeny). ................................................. 115

6.2 (left) Distribution of proteins near the sites. The numbers represent the fraction of proteins allocated to the sites. (right) predicted occupancy of site \( j \) by protein \( i \), can be interpreted as a probability of binding at each site ......................................................... 118

6.3 Parameter values learnt using the initialization method described in appendix C. Values and standard deviations result from averaging the parameters learnt from ten different instances of simulated microarray measurements ................................................. 129

6.4 Parameter values learnt when the learning algorithm is fed with the physical constants given in section 6.1.2 with 10% additive noise. Values and standard deviations result from averaging the parameters learnt from ten different instances of simulated microarray measurements . 129

7.1 Function associated with genes in small subsample ....................... 135
Chapter 1

Introduction

1.1 Motivation

Genetic regulation determines how DNA is read and interpreted, and has a key role in determining the behavior of an organism and the way it responds to external stimuli. Genetic regulation is responsible for cell specialization, individual reaction to certain drugs, metabolism, etc. An improved understanding of genetic regulation has the potential to impact research on genetic diseases including cancer.

The first component of genetic regulation is transcriptional regulation. Transcriptional regulation relies on the coordinate action of regulators, that bind specific regions of DNA and alter the way DNA is read. At any given time, the arrangement of regulators along DNA helps explain which genes are being expressed. Hence, each binding arrangement along DNA is an indication of the processes that are active in the cell.

However, understanding how each binding arrangement is achieved and how it affects transcription is challenging since it is not always possible to study regulatory processes in isolation (v. Herman et al. (1993)). The action of regulators is driven by mechanisms at different scales. Globally, understanding how regulators reach their target sites (transport mechanism) often requires accounting for interactions with other regulatory processes resulting in increasingly complex models. Locally, the broad range of biological mechanisms that describe transcription initiation poses a
challenge for detailed mechanistic models. For instance, regulators exhibit coordina-
tion and compensating effects, and the function of each regulator may be affected by
local structural changes induced by other regulators; furthermore, because regulators
themselves are encoded in DNA, and their production is regulated, regulation often
involves feedback loops. Finally, obtaining precise measurements of the activity of
each regulator is both an experimental and a computational challenge.

Control theory abstractions of regulation (Jacob and Monod, 1961; McAdams and
Arkin, 1998) emphasize that unlike other control systems, genetic regulation is highly
decentralized. Regulation arises from local adherence of each regulatory subsystem
to the laws of physics, chemistry, and molecular biology; and results from protein
availability and the local composition of the cell.

While it is sometimes possible to accurately explain the binding arrangement
of regulators locally, explaining the global arrangement of regulators in response to
experimental perturbations is very difficult. We need causal models that easily scale
genome-wide and can work with existing high-throughput observations.

Current models of transcriptional regulation, range from simulation to statisti-
cal methods. Simulation models (Samoilov et al., 2002; Gillespie, 1976, 1977, 2001;
Arkin et al., 1998; Shea and Ackers, 1985) can successfully predict dynamic behavior
if given an accurate representation of the problem (e.g., genetic circuits (Monod and
Jacob, 1961; Davidson et al., 2002; McAdams and Arkin, 1998)); however they do
not scale easily and require prior knowledge about the underlying process. Statistical
and probabilistic methods, on the other hand, provide accurate genome-wide descrip-
tions of experiments and can help uncover dependencies between parts of the system
without extensive prior information (Harbison et al., 2004; Bar-Joseph et al., 2003;
de Jong, 2002). However, their ability to extrapolate is limited.

This thesis lies in between mechanistic and statistical methods and defines a new
class of models that are scalable and able to uncover causal structure from high-
throughput observations. Our models of DNA binding predict average binding ar-
rangements that enable us to abstract genetic regulation as a sequence of regulatory
states. They are geared towards capturing interaction between subsystems genome-
wide. Moreover, the new modeling approach has potential to impact experiment
design. Models extract structure and parameters from high-throughput observations;
predictions to experimental perturbations can then be used to fine-tune hypotheses
efficiently using our models in a trial and error manner.

The economic models and methods we develop illustrate a broader family of scal-
able methods for recovering economic structures from narrow, possibly noisy, slices of
data. Specifically, they are particularly apt to address problems in which observations
about an aggregate behavior are known to emerge from the action of individuals.

1.2 Approach

Figure 1-1 depicts the problem we seek to address. Each binding arrangement is
seen as an instance of a competitive equilibrium in an economy defined by the pref-
erences of individual proteins and DNA sites. To emphasize that each arrangement
indicates different regulatory activity we sometimes refer to binding arrangements as
regulatory states. By analogy with chemical equilibrium, the competitive equilibrium
summarizes a steady state behavior for the whole of DNA. It can be used to determine
average allocation of regulators given the availability of proteins in the nucleus.

Figure 1-1: DNA binding regulators (proteins) modify the way DNA is read. Our models
address the following question. Given the availability of regulators (caricatured as free-shape
bodies above,) we ask what their expected occupancy along DNA (caricatured as a cylinder
above.) The presence of bound proteins along DNA in average defines a regulatory state.
Like concentrations at chemical equilibrium, the expected occupancy can be interpreted
as a dynamic steady state. Processes affecting protein availability (i.e., degradation and
protein production) are externalized. The goal is to predict the allocation shown at the
right with the inputs from the left. To introduce learning we associate the allocation with
observations and try to solve the inverse problem of determining the economic structure
that gave rise to such observations.
Game theory and economic models (von Neumann and Morgenstern, 1944; Nash, 1951; Fudenberg and Tirole, 1991; Arrow and Debreu, 1954) provide an appropriate framework for understanding (and searching for) the context and constraints associated with interacting regulatory processes. The theory of abstract economies provides a framework to analyze the binding arrangement of regulators directly as a competitive equilibrium, independently of the way the arrangement is reached. Like the law of mass action for chemical reactions, competitive equilibrium can provide causally meaningful predictions in response to experimental perturbations (gene knockouts, changing concentration of regulator, etc) in the form of binding arrangements. Thus competitive equilibrium models regulation independently of the reaction kinetics.

The key hypothesis is that given fixed concentrations of proteins in the nucleus there is an average arrangement of regulators, and it can be interpreted as an average over a population or a time average. The arrangement defines the system state genome-wide.

Our models view regulators and DNA sites as agents in an economy. The physical, biological and chemical parameters specific to the way each agent interacts with the others define the structure of the economy. The laws of physics and chemistry provide the “preferences” of each agent in the economy. In the economic approach these preferences lead to predictions about average arrangement of proteins along DNA given the available protein concentrations. We view these arrangements as defining regulatory states. The predicted regulatory states can be combined with kinetic studies to study evolution of regulatory processes as a sequence of regulatory states. Finally, the economic approach can be used to predict the effect of experimental perturbations on the regulatory state of the cell.

Often the physical and chemical quantities that define the preferences are unknown. By developing methods for learning economic models we hope to infer causal structure of regulation from noisy observations of regulatory states such as microarray observations. Our approach to learning deviates substantially from work at the interface of machine learning and game theory and is closer in spirit to the theory of revealed preferences in economics.
Our approach enables four new key predictions: predictions to guide experiment design, functional identification of complexes and joint binding events, a causal representation of binding as a network of binding constants, and new measures of similarity to discover the function of genes and proteins according to the predicted binding structure.

The new modeling approach is scalable, yet provides meaningful causal predictions about average allocation. We show how to learn the causal structure of the economy from limited observations and provide fast algorithms that predict average binding arrangements genome-wide as a function of the availability of regulators. The application of the economic approach to DNA binding illustrates key features of a new modeling framework. In particular, the economic formalism casts biological concepts in computational terms and captures the problem structure in terms of explicit player preferences.

Compared to traditional approaches to model DNA binding, the economic approach has two key advantages owing to the transport mechanism and competitive equilibrium:

- The transport mechanism enables us to model the problem of DNA binding at different scales, and study interactions between processes mediated by protein availability (i.e., resource allocation of reagents in excess and propagation of limiting reagent constraints).

- The notion of competitive equilibrium propagates the constraints imposed by each agent globally, while remaining at a convenient level of abstraction. A global energy function might be able to capture the same level of detail in the interactions but only at the expense of introducing sufficient interaction terms, and thus increasing the complexity of the representation of the problem.

### 1.2.1 Evaluation

We validate our approach by comparing our predictions with known biological results. We demonstrate that the DNA binding economies reproduce a genetic switch in ac-
cordance with the qualitative and quantitative literature on the subject. The genetic switch is the viral infection of E. Coli by the λ-phage virus (Ptashne, 1987). The second main experimental result shows that the economic approach can complete the map of coordinate binding of regulators in yeast (Harbison et al., 2004).
1.3 Contributions

This thesis has key explicit contributions to computational biology. The application to computational biology illustrates a new modelling framework with broad applicability in machine learning. We overview the contributions to computational biology and the key distinctive features of the new modelling framework.

Computational biology. We develop a new abstraction of genetic regulation to infer causal regulatory structure from narrow observations of binding arrangements, and predict regulatory states genome-wide in response to experimental perturbations and as a function of protein availability.

- We define three classes of economies that differ by the scope of the transport mechanism, and explore the trade-off between purely competitive processes and processes that involve explicit coordination. We prove that an equilibrium in pure strategies exists for any setting of the parameters of the economies.

- We map each economy to a fixed point equation and prove monotonicity properties. We also indicate a connection with the theory of supermodular games.

- We develop fast iterative algorithms based on iterated dominance to predict regulatory states genome-wide. Each iteration of our algorithms is polynomial in the number of sites and proteins. Convergence rate of the iterative process depends, however, on the values of the parameters and the specificity of the transport mechanism. The algorithms provided in this thesis can be seen as instances of constraint propagation or message passing algorithms.

- We develop learning algorithms to infer the structure of regulation from high-throughput binding observations.

- Our approach enables four new key predictions:
  - predictions to guide experiment design in an effective trial and error manner,
  - functional identification of complexes and joint binding events,
- a network representation of the predicted binding structure, and
- new measures of similarity to discover gene and protein function based on
  the predicted binding structure.

• we illustrate the uses of the economic approach in a small scale example and a
  large-scale example.

**Machine learning.** The economic models developed in this thesis illustrate a new
modelling paradigm. Specifically, we show by example how to map a problem onto
an economy and learn an economic structure from data. The economic structure
learned can then be used to make causally meaningful predictions in response to
perturbations. This use of game theory and abstract economies is new to machine
learning; the economic theory of revealed preferences is the closest theoretical analog
to our approach. We enumerate the distinctive features of the new framework

  • We associate observations with equilibrium and try to solve the inverse problem
    of determining the parameters of the utility functions that gave rise to such
    observations. Our learning approach assumes rationality of the agents. We say
    that we learn an economic structure owing to the formalism of graphical games
    and economies.

  • The new framework is scalable, and suitable for inference from limited observa-
    tions. While limited in depth, observations may come from high-throughput ex-
    periments, our framework assumes that observations may be partial and noisy.
    The type of structure learned is, however, restricted by the mapping of the
    problem onto an economy. Specifically, it is restricted by the choice of utility
    functions, which encode prior knowledge about the problem.

  • The framework is a new application of economic theory in an experimental
    setting. It is specifically suited to model problems where the global state of
    the system is not easily described with global considerations about energy or
    efficiency, but can be easily cast as emerging from the structure of local inter-
    actions.
1.4 Road Map

Chapter 2 provides background in computational biology and game theory. The biological overview is used to motivate the economic abstraction. We contrast the economic approach with more traditional approaches to modeling transcriptional regulation. The chapter ends with a brief overview of game theory and the theory of abstract economies.

Chapter 3 introduces the economic abstraction of genetic regulation. We develop three economic models with different assumptions about scale of binding and the competing and coordinating behavior of the agents.

Chapter 4 develops the properties of competitive equilibria. We prove the existence of competitive equilibria for each economy and under any setting of the parameters; and develop algorithms to compute such competitive equilibria.

Chapter 5 develops learning algorithms to infer the causal structure of the DNA binding economies from possibly noisy observations of equilibrium allocations.

Chapters 6 and 7 provide empirical analysis of our models. Results on small scale and genome-wide examples are discussed. We also present new kind of predictions that can be achieved with our economic models.

Chapter 8 outlines future extensions to this work, within computational biology and more broadly in machine learning.

Chapter 9 discusses our findings and conclusions.

Appendices A, B, and C present further detail on the proofs of existence of competitive equilibria, our learning algorithms and our empirical analysis.

Some content of this thesis appeared in Pérez-Breva et al. (2006) and Pérez-Breva et al. (2007).
Chapter 2

Background: Genetic Regulation,
DNA Binding, Economies

We review concepts in game theory, the theory of abstract economies and biology relevant to our economic abstraction of DNA binding. We also motivate the relevance of DNA binding in genetic regulation and as an illustration of the key features of the new modelling approach. Background on the application of computational learning to game theory is covered in Chapter 5.

The chapter is structured as follows, the first section introduces key biological notions about information flow, cell regulation and transcriptional regulation, and narrows the scope of our models down to DNA binding.

The second section overviews computational methods in genetic regulation. We categorize computational approaches to modeling genetic regulation as either probabilistic or mechanistic. This categorization applies to a broader class of problems. The method proposed in this thesis represents a new approach to modelling with respect to traditional probabilistic and mechanistic methods. It incorporates the convenient level of abstraction of probabilistic methods with the ability to make causal predictions of mechanistic methods. Nevertheless, it should not be understood as a replacement for either. Applied to DNA binding, the economic approach enables predictions about the average arrangement of proteins along DNA, which are inherently different from the descriptive predictions of probabilistic methods and the dynamic
simulations of mechanistic methods, and can be combined with both.

The third section introduces the framework of game theory. We first introduce game theory intuitively. Then, we formalize the concepts in game theory and its generalization to abstract economies that are required for our abstraction of DNA binding.

2.1 Biological Background: Genetic Regulation

The action of cell regulation is connected with the presence or absence of proteins in the cell medium. Regulation takes place via alterations in the flow of sequence information from DNA to protein. Sequence information refers here to the composition of the biological polymers (i.e., DNA, RNA, proteins).

The central dogma of molecular biology (Crick, 1956; Cric, 1958; Crick, 1970) summarizes the flow of sequence information between biological polymers with a categorization of all possible transfers of information. In its more succinct form the dogma is a negative statement: “Once sequential information has passed into protein it cannot get out again” (quoting Crick (1970)). Note the difference with the sequence hypothesis: “The transfer of nucleic acid (DNA, RNA) sequence information into proteins (sequence of amino-acids) occurs.” The diagrammatic representation of the dogma combines both the dogma and the sequence hypothesis. Figure 2-1 reproduces the diagram from (Crick, 1970).

The completeness of the central dogma has been challenged often (cf. Temin and Mizutani (1970); News and Views (1970) Henikoff (2002); Mattick (2003); W. (2003)). However, the only standing challenge to the central dogma, recognized already by Crick (1970), is the existence of prions (v. Keyes (1999)).

(Watson, 1965) introduced a simplified text-book illustration of the central dogma schematically presented in Figure 2-2 that has often been mistakenly cited as the central dogma.

From a computational perspective, the central dogma is a statement about which semantic transfers are permitted. The unit of meaning is an amino-acid (the basic
Figure 2-1: Schematic representation of sequence information flow. (left) Representation of all possible information transfers between bio-polymers. (right) Representation of sequence information transfers “Permitted” according to the central dogma of molecular biology and the sequence hypothesis. Type of arrows indicates category of transfers. Solid arrows are general transfers, dashed arrows are special transfers. Arrows were omitted for unknown transfers (diagrams are taken from Crick (1970))

**Sequence Information**

<table>
<thead>
<tr>
<th>DNA</th>
<th>Transcription</th>
<th>RNA</th>
<th>Splicing</th>
<th>Export</th>
<th>PROTEIN</th>
<th>Translation</th>
</tr>
</thead>
</table>

Figure 2-2: Illustration of a simplified schematic for the central dogma introduced by (Watson, 1965), intended as a textbook illustration and often mistakenly cited as the central dogma of molecular biology. The picture has been complemented with the name given to the biological mechanism at each step; the original schematic from (Watson, 1965) included also an arrow for replication omitted here because it is not part of regulation.

building block of proteins). So, the semantics of DNA correspond to sequences that encode proteins. Nothing is said about the mechanisms by which such transfers occur, nor about additional information pertaining to the syntax of bio-polymers. Syntax refers to, for instance, sequences of base pairs\(^1\) that indicate how DNA is read (e.g., operator sequences where factors bind DNA), or RNA segments that are spliced out (introns). (v. Darden (2006) for a study of the implications of the meaning of information in the central dogma, and Werner (2005) for a critique of the limitations of the central dogma in its application to computational research in molecular biology.)

Cellular regulation is in part concerned with the ability to affect (activate, accelerate, or inhibit) semantic transfers between bio-polymers. The simplified picture from

\(^1\)Base pairs are the base constituents of DNA, typically abbreviated with the letters GTCA.
Figure 2-2 is useful to identify the main transfer mechanisms that are affected by regulation and emphasizes the fundamental role of DNA regulation. In all generality the simplified picture from figure 2-2 is incorrect. It is, however, convenient to narrow down the scope of our models.

In Figure 2-2 DNA is the source of sequence information and proteins are the final product. Proteins have a structural or procedural function somewhere in the organism. At any point in this chain, the mechanisms of control of the cell can act to inhibit or magnify the information flow to change the amount of final protein product.

As a first approximation, the economic models proposed in this thesis focus on the first control “opportunity” in this chain: transcriptional regulation.

2.1.1 Transcriptional Regulation

Transcriptional regulation determines which sequences are read and consequently affects the processes that are active in the cell. Broadly speaking, transcription is the process by which a DNA reader (i.e., the protein RNA-polymerase) transcribes a sequence of DNA into RNA; the sequence starts at a promoter region and ends with a termination sequence. The initiation of transcription relies on the action of regulators, factors, and cofactors, that bind DNA and affect (facilitate or inhibit) binding of RNA-polymerase to the promoter region. Hence, regulators alter the way DNA is read. The name regulator is a placeholder for a wide range of chemical species: proteins, protein complexes, and several smaller molecules. While there are many ways in which regulators affect binding of RNA-polymerase, the function of regulators is closely related to that of catalysts: regulators accelerate or inhibit binding of RNA-polymerase to DNA.

The dynamic nature of transcriptional regulation is often encoded in a circuit representation. We show one example of a genetic circuit in Figure 2-3. The dynamic response of genetic circuits provides insights about the time-dependent component of regulation and its connection to control theory. Jacob and Monod (1961) and McAdams and Arkin (1998) present the cell as a complex and robust control system, with mechanisms at different scales being responsible for its regulation. We overview
previous work on simulation of genetic circuits in Section 2.2.1.

In this thesis we are interested in a complementary view of transcriptional regulation. We consider a population of cells in identical conditions. (We later reformulate the same idea using time averages.) At any given time, a snapshot of the regulators bound to DNA averaged across the population indicates which processes are being regulated at a transcription level. DNA binding is a precursor to transcription, and the average arrangement of regulators bound along DNA is an indication of the regulatory state of the cell. The models introduced in this thesis determine the arrangement of regulators, hence the average regulatory state, as a consequence of the availability of regulators.

Protein-DNA binding (a.k.a DNA binding)

Generally, regulators are proteins that bind specific DNA-sites activating or repressing transcription of one or a family of genes. The broad range of biological mechanisms employed by each regulator poses a challenge to introduce mechanistic models of DNA binding. Additionally, understanding the role of regulators often requires accounting for interactions with other DNA processes resulting in increasingly complex models. The following examples of characteristic regulatory mechanisms illustrate some of the challenges faced by detailed models of regulation.

- Regulators may unblock/block a promoter region and thus facilitate/impede RNA-polymerase binding, for example, by binding next to the promoter (this is the case for regulators cI2 and Cro in λ-phage, v. Ptashne (1987)).

- Conformational changes in either regulators or DNA may also be responsible for regulation by hiding (unfolding) previously available (unseen) sites; for instance, in E. coli, lactose produces a conformational change in the lactose repressor that frees LacI’s operator site and activates transcription (Monod and Jacob, 1961).

- In several processes, regulators may act alone or complexed with another regulator, (e.g., complex GAL4-GAL80 inhibits the galactose pathway and GAL4 alone activates it, v. Johnston (1987).)
• The action of one regulator may be affected by the availability of other regulators (e.g., growth in \( \lambda \)-infected E. Coli may alter the decision lysys/lysogeny, \( v \). Herman et al. (1993)).

• Regulators as DNA-encoded proteins are prone to regulation (when self-regulated, they define feedback loops).

To overcome these difficulties, models of DNA binding rest at a slightly higher level of abstraction that consider the mechanisms of transport, binding, and interaction between subsystems with different levels of detail. Even at higher levels of abstraction, however, our knowledge of the underlying physical mechanisms is partial. The following two mechanisms are crucial to every model of regulatory mechanisms, and are not yet fully accounted for

I how interaction between subsystems alters regulation, and

II how the limited numbers of regulators are allocated to each subsystem (i.e., transport mechanism).

The first of these two challenges was motivated already by Monod and Jacob (1961). While examples of unrelated regulatory subsystems that interfere with one another in predictable ways exist (e.g., Herman et al. (1993)), interaction between subsystems is often overlooked. Most approaches focus either on isolated subsystems or on discovery of qualitative relationships between regulators and DNA modules.

We believe that the notion of competitive equilibrium introduced in section 2.3 is adequate to frame this interaction as a byproduct of strategic allocation of resources. Hence, the economic models proposed in this thesis incorporate explicit mechanisms to model and learn the interaction between seemingly independent subsystems.

In our formalization, the strategic allocation of resources is strongly related to the interpretation of the second challenge above, transport mechanism, as a complement to DNA binding. The next sections overview the biological motivation to further decompose protein-DNA binding in two components: binding and transport.
2.1.2 DNA Binding as Transport and Binding

Berg et al. (1981), Stanford et al. (2000), and Jacob and Monod (1961) have studied the adequacy of different physical models to explain the transport mechanism. The general consensus is that an unspecified “sorting process” brings each protein near specific sites. The dynamics of this process are still not fully accounted for. There is experimental evidence to support diffusion-driven models. However, in some cases, empirical evidence suggests that transport occurs too quickly for 3D-diffusion alone. In Lac repression, facilitated diffusion seems to provide a better explanation for the diffusion process. In this case facilitated diffusion is a combination of 1-D diffusion along DNA and 3-D diffusion elsewhere. Intuitively, proteins “wander through the cell” until they “hit” DNA, then they “explore” DNA by hopping/sliding from site to site. At any point in the exploration of DNA, proteins may continue exploration (non-specific binding), spend more time at the sites where attraction is stronger (specific binding), or eject from DNA in which case 3-D diffusion exploration takes place again. Halford and Marko (2004) review the mechanics of the various explanations for the transport mechanism.

The empirical work of Sunney Xie’s group on the effect of transport, binding and stochasticity in genetic regulation using single-molecule detection techniques (Xie et al., 2006) supports facilitated diffusion driven models of transport. Elf et al. (2007) present empirical evidence in support of a non-specific binding phase (or “search” phase) that combines 1D diffusion along DNA segments and 3D “translocation” between segments. Yu et al. (2006) develops additional insights about the role of stochasticity in the dynamics of transcription once the “search” phase has concluded. This work also concludes that transcription has a stochastic component that causes proteins to be produced in bursts. The consequences of the stochasticity may be more apparent in early development, for instance, regulation timing in early development may help explain why identical twins have different fingerprints Xie (2007). (Cai et al., 2006; Friedman et al., 2006) complete this work with models of the observed stochasticity and their effect over cell populations and expression.
The search phase is said to conclude when factors reach the surroundings of their target site. Then it is customary to assume that the factor will remain near the target site, and that it will bind to it according to the dynamics of chemical equilibrium. The stochasticity noted by Yu et al. (2006) may be explained via thermal fluctuations and does not contradict the dynamic interpretation of chemical equilibrium.

The economic models in this thesis do not stand in contradiction with any of the existing hypotheses on the dynamics of the transport mechanism. We abstract out the mechanism dynamics and focus on transport only as a resource allocation process that brings proteins to the surroundings of the site. Our abstraction takes advantage of the freedom to define the “surroundings” of the site to model protein-DNA binding interactions at different scales, introducing three economies with different granularity.

Without invoking the transport mechanism, binding is generally modelled using chemical equilibrium. The computational literature that invokes the concept of chemical equilibrium introduces the assumption that the nucleus of the cell is a well-mixed system (cf. Shea and Ackers (1985); Arkin et al. (1998)). This assumption is required for chemical equilibrium to be applicable. However, it is contradicted by the very existence of a transport mechanism. The granularity in our economies is based on a precise definition of the concept of surroundings. Then, assumptions about homogeneity of the mixture depend on the scope defined for the transport mechanism.

2.2 Computational methods in Binding and Regulation

There are essentially two computational approaches to modeling binding and regulation: mechanistic and statistical approaches\(^2\).

\(^2\)In this section we use 'statistical approaches' and 'probabilistic approaches' interchangeably.
2.2.1 Mechanistic Approaches

Mechanistic approaches use physical-chemical relations to capture the mechanics and dynamics of the target system. Causal predictions about the system dynamics are made via simulation. Mechanistic models have popularized the analogy with electronic circuits, first introduced by Monod and Jacob (1961), (see Samoilov et al. (2002) for a rigorous signal processing study). Within the family of mechanistic models, stochastic simulation (Gillespie, 1976, 1977, 2001) has recently gained much popularity after McAdams and Arkin (1997) combined it successfully with genetic circuit representation to explain the apparent randomness observed in the dynamics of gene expression. Other examples of mechanistic models of genetic and molecular regulatory networks include (Arkin et al., 1998; Shea and Ackers, 1985). McAdams and Arkin (1998) motivate the recent interest in extending molecular regulation models with models of genetic regulation and review the state of the field. Mechanistic models capture successfully what is known about the physics of regulation; in particular, they encode concepts such as feedback, cascaded processes, and, through stochastic simulation they have been shown to mimic the apparent randomness of protein transcription.

The most notable drawback of mechanistic models lies in the complexity of the genetic circuits they simulate, and the scarcity of known genetic circuits. From a computer science perspective, numerical simulation of a large set of differential equations is a costly process that in the case of stochastic simulation has to be complemented with frequent estimation steps. As a result, mechanistic models are highly specific and difficult to scale. Davidson et al. (2002) describes research leading to a genetic circuit for endomesoderm differentiation in sea urchin embryos; it exemplifies the laborious procedure required to produce the regulatory genetic circuits. Figure 2-3 reproduces the resulting genetic circuit\(^3\). In general, mechanistic models cannot be easily adapted to discover networks of interaction, one exception to this rule is (Vance et al., 2002) that uses a perturbation analysis to estimate causal connectivities and part of the kinetics of networks of reactions. Finally, mechanistic models do not ad-

\(^3\)This image can be downloaded from http://sugp.caltech.edu/endomes/
Figure 2-3: Illustration of a genetic circuit. The circuit models the first 30 hours of development of the sea urchin, in which the endomesoderm differentiation of the embryo occurs. The work leading to this image is described by Davidson et al. (2002), and the image was obtained from http://sugp.caltech.edu/endomes/. This work illustrates the laborious procedure that building a circuit representation of a regulatory network requires and the complexity of the resulting circuit.

dress resource allocation or interaction between subsystems explicitly, since in most cases the physical knowledge about both is ad-hoc.

2.2.2 Statistical and Probabilistic Approaches

Statistical and probabilistic approaches have become increasingly important with the advent of micro-array technology. These methods ease the processing of vast amount of data coming from micro-arrays. In general they are used as data processing techniques to discover associative networks of regulation, binding sites, network motifs, gene modules, etc. from expression data (possibly genome-wide.)

The following examples illustrate statistical and probabilistic approaches in the context of genetic regulation. Lee et al. (2002) investigate regulator-gene interactions and introduce the idea of a combinatorial search algorithm to identify network motifs. Harbison et al. (2004) use comparative genomics to produce a genome-wide map of

\[^4\]A network motif is the simplest building unit of a regulatory network
\[^5\]A gene module is a group of genes that are regulated by the same regulator
\[^6\]comparative genomics helps identify conserved sequences across different species
regulators and sites in *Saccharomyces*; in that work, they suggest that environmental dependence of certain regulator binding patterns can provide clues about the underlying mechanisms. Segal et al. (2003) present a more elaborate probabilistic method involving clustering and a rule based system to infer regulators, regulated modules, and conditions of regulation from micro-array gene expression data. Bar-Joseph et al. (2003) propose a probabilistic search algorithm (called GRAM) to discover gene modules and regulatory networks from a combination of genome-wide location and expression datasets. de Jong (2002) surveys a number of statistical and computational data processing techniques typically used in the modeling of genetic regulatory networks.

The key strengths of probabilistic methods are their immediate applicability to genome-wide data, their ability to integrate data from different sources and their capability to discover associative networks with few assumptions. Obtaining causal predictions requires either impractical experimental designs with an accurate control of time and a prohibitive number of laboratory experiments, or building causality into the model; for instance Nachman et al. (2004) use generative probabilistic models (Bayesian networks) that encode some of the physics of the problem to infer quantitative models of regulatory networks.

Clearly, mechanistic and statistical models differ not only in their approach but also in their goals. The former simulate the evolution of the system to make predictions according to a physical model of the problem using, for example, genetic circuits; the latter identify all the elements involved in genetic regulation and the corresponding associative network. When used in combination mechanistic models help reduce the search space of statistical methods, and statistical models provide estimates of unknown parameters for mechanistic models.

### 2.2.3 This Work

We introduce a new modeling approach based in a game-theoretic and economic framework that can provide meaningful causal predictions about the arrangement of regulators along DNA. The new modeling approach does not require explicit repre-
sentation of the dynamics of the problem, nor does it depend on a pre-established representation of the regulatory network; however, it incorporates the physics of the problem and uses learning to infer the problem structure from gene expression data. This approach is not meant as a replacement of either statistical or mechanistic approaches but as a complement to both. For instance, it can be combined with dynamic simulations to study the evolution of regulatory states.

The economic approach captures both resource allocation and interaction between subsystems seamlessly. Through definition of the scope of the transport mechanism, it can be adapted to capture DNA binding interactions at different scales.

Our approach ultimately combines economic theory with learning to infer the structure of the economy, and can be used in conjunction with kinetic models to study the evolution of the binding arrangement of regulators. The new modeling approach is particularly apt to assist in experiment design. Combined with learning, it can analyze high throughput (and possibly noisy) observations of the state of equilibrium timely and make causal predictions about the effect of altered conditions on regulatory states genome-wide. Hence, it can be made part of a \textit{trial and error} process to fine-tune experimental hypotheses.

The next sections describe the key concepts in game theory and economic theory that are relevant to the abstraction of protein-DNA binding. Note that since abstract economies generalize game theory, the new modeling approach can also be understood with game-theoretic concepts.

\section{2.3 Game Theory and Economic Theory}

Game theory and economic theory provide the mathematical framework to addresses strategic problems with a focus on competition, coordination and resource allocation (von Neumann and Morgenstern, 1944; Nash, 1951; Fudenberg and Tirole, 1991; Arrow and Debreu, 1954). Applications of game theory range from economics (von Neumann and Morgenstern, 1944) and evolutionary biology (Lewontin, 1961; Maynard Smith, 1982; Fisher, 1930), to a broader class of problems related to local op-
timization and resource allocation in the social sciences and engineering. Computer science has widened the application of game theory with an emphasis in structure and strategic decision making (Kearns and Ortiz, 2003; Kearns et al., 2001), learning in games (Fudenberg and Levine, 1998), and the complexity of computing equilibria of games (v. Papadimitriou (1994); Goldberg and Papadimitriou (2005); Daskalakis et al. (2005); Daskalakis and Papadimitriou (2005); Chen and Deng (2005); Chen et al. (2006) for a general complexity analysis of computing equilibria, and Gilboa and Zemel (1989); Conitzer and Sandholm (2002) for the complexity analysis of specific types of equilibria.)

A set of players and corresponding set of actions (strategies), and a measure of payoff define a game. Game theory predicts the final outcome of the game to be an equilibrium under different assumptions of rationality of the players, and the information available to them.

In normal or strategic form, games are abstracted as ”single-shot” simultaneous-move game (much like rock-paper-scissors). The degree to which players are aware of other players’ strategies, and choose their actions individually or jointly with other players defines the frontier between cooperative and non-cooperative game theory. Here we focus on non-cooperative game theory.

For games in normal form, the goal of the analysis is to characterize the possible final equilibria of the game given all the information about the players preferences. We emphasize that the theory does not explain how is the game played, and, in most cases, it does not provide a way to compute the equilibrium either (the method of iterated dominance Nash (1951), however, helps constrain the space of equilibria.) In non-cooperative game theory a player’s payoff depends on the strategy of his rivals, but the player has no direct control on the choice of strategy of his rivals. Table 2.1 illustrates this distinction with a paradigmatic example.

The usual assumption is that players act rationally, which implies that each player aims to maximize his own payoff. Then, for many games some joint actions verify that every player has topped his payoff. Simply put, a player analyzing the game
Table 2.1: Prisoner’s dilemma in normal-form. Two thieves are being separately interrogated on charges of bank robbery. If neither suspect confesses they both get away and split the haul of the robbery (5 each.) If one confesses, he goes free and gets the whole loot (10 units) while his accomplice gets full sentence. If both confess both get reduced sentences but the loot is recovered. As illustrated in the table to the left, suspect 1 chooses to confess based on the payoff of his options conditioned on the options of suspect 2. While suspect 1 may be aware of suspect 2’s payoffs, he does not take them into consideration when making a decision. If rational, suspect 1 chooses based on the strategy that maximizes his payoff. Note, however, that in many games such a strategy may not exist, or it might exist but not be apparent to the player. This would be the case of multiple equilibria. The table to the right shows the perspective of the game theorist on the prisoner’s dilemma, which is a symmetric game. This game has only one Nash equilibrium in which both players choose to confess.

will conclude that no change in his strategy alone improves his payoff (i.e., either his payoff is maximal or to improve it, other players would have to modify their strategy as well.) We then say that the game has a Nash Equilibrium. For example, in table 2.1 the joint strategy (confess, confess) is the only Nash equilibrium.

Nash (1951) showed that if players are allowed to ”randomize” their strategies then at least one such equilibrium exists for all games. However, an equilibrium may not always exist if randomization is not allowed, and conversely, games may have not one, but several equilibria. Technically, every game has at least one equilibrium in mixed strategies, although not all of them have an equilibrium in pure strategies:

Nash equilibrium is the minimal necessary condition for achieving ”reasonable” predictions under the assumptions of rationality in the context of single-shot one move non-cooperative games.

Note that the concept of Nash equilibrium is a description “a posteriori”. That is players cannot target reaching a Nash equilibrium, from their perspective, reaching Nash equilibrium is an inverse problem. The concept of rationality makes the con-
nection between the “a priori” decision of the player and the “a posteriori” concept of Nash equilibrium. If the outcome of a game is not an equilibrium, then either the game had no equilibria in pure strategies, or somebody did not play rationally. The concept of rationality roots the discussion about the applicability of game theory in experimental settings and the meaning of the concept of Nash equilibrium when players learn to play a game. Fudenberg and Levine (1998) analyze player learning in the context of repeated and multiple-move games, and motivate the types of equilibria that arise as the players’ strategy set is updated with information from previous moves/games. In the context of learning in games, it is apparent that Nash equilibrium is not always the most reasonable outcome. We further overview learning in games in chapter 5.

In this thesis, learning is used in a different context; the goal is to determine the structure of the game, that is the parameters of the utility functions from noisy observations of Nash equilibrium, rather than the way players learn. Learning in this thesis relates to approximation theory and econometrics/statistics, not to an incrementally acquired knowledge about the game and the strategy of the opponents.

**Formal Game Theory basics**

This section introduces the basic formalism of games in strategic/normal form (i.e., games that can be represented as matrices like in Table 2.1 above,) and the definition of Nash Equilibria. For a thorough introduction to game theory and its different representations see (Fudenberg and Tirole, 1991).

A game in strategic or normal form is defined by

I a finite set of players \( i \in \mathcal{I} = 1, \ldots, I \)

II a set of pure strategies \( s^i \in \mathcal{S}_i \) for each player\(^7\)

III a payoff or utility function \( u_i(s) \) for each player (can be represented as a matrix, cf. Table 2.1)

\(^7\)We indicate the player with a superscript index
A *mixed* or *randomized strategy* for player $i$, $\sigma^i$, is a probability distribution over its pure strategies. For convenience it is customary to group the strategies chosen by all players in a tuple and call it *profile*. The tuple $s = (s^1, s^2, \ldots, s^I)$ is called *pure profile*; and similarly if some player uses a mixed strategy, it is denoted $\sigma$ and called *mixed strategy profile*. The notation $u_i(s^i, s^{-i})$ is oft used to emphasize that player $i$’s utility is conditioned on the strategies of its opponents, abridged as $s^{-i} = (s^j) \mid j \neq i, j \in I$.

A mixed strategy profile $\sigma^*$ is a Nash Equilibrium if for every player, the following condition holds

$$\sigma^{i*} \in \arg \max_{\sigma_i} u_i(\sigma^i, \sigma^{-i*}).$$

(2.1)

This definition matches the previous informal discussion about Nash equilibria. However, to make the connection explicit it is useful to write the condition for Nash Equilibria as a comparison

$$u_i(\sigma^{i*}, \sigma^{-i*}) \geq u_i(s^i, \sigma^{-i*}) \forall s^i \in \mathcal{S}_i,$$

and emphasize that although it is in principle possible for players to chose other strategies with equal payoff, hence the equality, none improves it. Note that if some player $j$ changes to another strategy, the resulting profile may not be a Nash Equilibrium, even if the payoff of player $j$ remains the same; the reason for that is that a change in player $j$’s strategy conditions the payoff of other players.

**Abstract Economy**

*Arrow and Debreu* (1954) introduced the concepts of abstract economy and competitive equilibrium. In what pertains to this thesis, abstract economies generalize games in two ways, informally:

1. The set of strategies is continuous rather than finite and discrete.

2. The domain of strategies available to a player can be constrained by the strategies of other players. (Recall that in games only the payoff depends on the strategy of other players.)
Abstract economies also differ from games semantically. Nevertheless, when the two conditions above are met, referring to the formalization as a game or as an abstract economy is a matter of convenience. In this work we will use the concepts of an abstract economy because they serve our purpose better; in particular, they emphasize naturally two of the concepts we are trying to model: resource allocation, global constrains on resources.

In an abstract economy players are referred to as individuals or agents. Two kinds of agents are defined: suppliers/distributors \((y^i \in Y_i)\) and consumers/individuals \((x^j \in X_j)\). Strategies and actions are now understood as a trade of commodities: some distributors supply (produce) commodities while maximizing their utility function \(u_i\), and individuals consume them in accordance to their utility function \(v_j\). The steady state of the economy is represented via a distribution or allocation of commodities. Finally, the concept of Nash equilibrium is generalized to that of a competitive equilibrium.

The idea of a competitive equilibrium is essentially the same than that of a Nash Equilibrium: an allocation of commodities such that agents have no incentive to unilaterally deviate. However the conditions for an economy to have a competitive equilibrium are slightly more elaborate. Here we present only the conditions relevant to this work:

I The utilities \(u_i(y^i)\), and \(v_j(x^j)\) are continuous quasi-concave functions on the corresponding strategy set \((X_j \text{ and } Y_i, \text{ respectively.})\)

II The set of production and consumption vectors \(i.e.,\) the strategy spaces \(X_j \text{ and } Y_i\) are closed convex subsets of \(\mathbb{R}^\ell\)

Note that abstract economies are defined naturally with a bipartite structure.

Abstract economies are generally defined to incorporate the notion of price for each commodity, which is set by a fictitious player. When price is incorporated there are some additional technical conditions to the competitive equilibrium. In this thesis we will not be using the notion of price explicitly, which is equivalent to say that we assume it fixed.
Part I

Mathematical Abstraction
This chapter presents the protein-DNA binding problem as an economy. This is a three step process.

I Identify the agents of the economy and means to rank their preferences.

II Associate the competitive equilibrium with observations/predictions.

III Define an algorithm to compute the competitive equilibrium.

In analyzing mathematically the resulting abstraction we are interested in the properties of the competitive equilibrium, its existence, uniqueness, and mapping the possible restrictions imposed on the strategy space of agents back into the original problem.

We begin by motivating the choice of economic theory to model protein-DNA binding, map the biological concepts onto an economic abstraction, and present three different economic models that synthesize the problem of protein-DNA binding at different scales. We examine the properties of the competitive equilibria for each of the resulting models and present algorithms to compute them (steps II and III above) in Chapter 4.
3.1 Motivation

DNA binding is a dynamic process. At any time, allocation along DNA arises from the interactions of individual proteins and sites with each other, as determined by the laws of physics, chemistry, and molecular biology in the different context of transport and local binding.

The purpose of our abstraction is to study the average allocation along DNA as a function of the availability of chemical species in the nucleus. We abstract the dynamics of binding out by making protein availability a function of time. We focus in the definition of a regulatory profile as the average allocation that would result, if protein resources remained fixed, from the structure of physical interactions that defines binding genome-wide. Figure 3-1, modified from Chapter 1, highlights the predictions we expect to achieve from the economic approach.

![Figure 3-1: Representation of the average regulatory profile. DNA binding regulators (proteins) modify the way DNA is read. Our models address the following question. Given the availability of regulators (caricatured as free-shape bodies above,) what is their expected allocation along DNA (caricatured as a cylinder above.) The expected allocation along DNA defines a regulatory profile. Like concentrations at chemical equilibrium the expected allocation can be interpreted as a dynamic steady state. Processes affecting protein availability (degradation and protein production) are abstracted out.](image)

Protein-DNA binding exhibits interactions at different scales. Interactions between DNA subsystems, mediated by protein availability, are only apparent at large scales; similarly interdependence between the bindings at close sites is only apparent at small scales. Biologically, the division between transport and binding sets the relative scale between physical interactions responsible for diffusion, and chemical interactions responsible for binding. For instance, assuming that the cell is an homogeneous mixture is equivalent to say that the effect of physical interactions in the diffusion process is negligible compared to chemical interactions. Conversely, as-
suming that only a small neighborhood around the sites is well mixed emphasizes physical interactions occurring between sites. The nature of the interactions our economic methods capture will depend on how we set the boundary between transport and binding mechanisms.

The problem of DNA binding exhibits several types of interactions at different scales: competition, compensation, constrained resource, and coordination.

- At each site, competing chemical reactions explain the binding proportions of each regulator.

- Overall availability of proteins constrains the proportion of bindings at each site.

- Binding events may be facilitated by a redistribution of protein allocation due to, for instance, excess binding by a third party at a different site. That is, binding exhibits compensating relations.

- Some regulators exhibit coordination ability when either binding as complexes or binding jointly at a promoter region.

The concept of competitive equilibrium is adequate to capture these types of interactions implicitly as a result of the choice of scale and the encoding of the utility functions.

Simulation methods that consider transcriptional regulation as a whole often model the dynamics of DNA binding using chemical kinetics and a stochastic abstraction. These simulations successfully capture the kinetics of DNA binding, and can be used to infer average allocation profiles by averaging multiple runs. There are situations, however, in which one might be interested in knowing just the average allocation profile, for instance, to study the genes being activated or repressed under a given set of conditions, and predict changes in the regulatory profile. In such cases there is value in a method that infers average allocation directly rather than as the result of multiple simulations. This situation that arises already in small scale illustrations is crucial to analyze medium and large-scale problems in which
multiple DNA subsystems are considered. Additionally, at their level of abstraction, simulation methods are hard to adapt for use with high-throughput data produced in microarray assays.

Economic approach and simulation methods are complementary. The relation between them is similar to the choice between kinetic studies and equilibrium dynamics in statistical mechanics. We may choose to study the chemical evolution of a system via simulation of chemical reaction kinetics or via equilibrium dynamics under assumptions of reversibility. While each approach is adequate to answer a specific set of questions, at equilibrium, both approaches are expected to predict the same average behavior. The economic approach to protein-DNA binding is the large scale analog to chemical equilibrium in statistical mechanics. The hypothesis underlying our approach is that given nuclear concentrations for the reacting species, it is possible to study average allocation of regulators genome-wide independently from the kinetics of regulator transport, binding and transcription. The goal is to reach conclusions about average allocation easier than through detailed kinetic simulation. An immediate consequence of these choices is that the level of abstraction of the economic approach is more suitable for inference from microarray assay data. So the economic approach bridges the gap from mechanistic methods to causal inference from microarray data (cf. Chapters 5 and 6).

### 3.2 DNA Binding Abstraction

The key concepts in DNA binding are now reintroduced as an abstract economy. As discussed above and in chapter 2 we define the global and local scopes of Protein-DNA binding via two subproblems, transport and binding. We introduce the main economic concepts in an economy where transport mechanism that brings near individual sites. We will later define two additional economies corresponding to coarser applications of the transport mechanism.
3.2.1 Agents

We can view this economy as having two types of agents, proteins and sites\(^1\). The protein-agent allocates proteins *near* sites along DNA in accordance with the transport mechanism. For instance, the RNA-Polymerase-agent allocates RNA-polymerase to sites. The transport mechanism defines the utility function of each protein-agent.

Sites along DNA, defined indirectly as sequences of DNA where proteins bind, are the *site-agents*. Sites choose which protein in their surroundings to bind every instant of time. Protein-site binding is defined by a chemical reaction, the utility of each site-agent ranks proteins according to the chemical activities of both product and reactants. Under assumptions of low concentration of reactants, the chemical equilibrium equation quantifies the reaction to good approximation. It follows that the utility of each site-agent has to encode the chemical equilibrium equation (*i.e.*, the law of mass action.)

3.2.2 Protein-Agent: Transport Mechanism

The transport mechanism captures how proteins are distributed around possible binding sites. While the dynamics of this mechanism is not well understood, a helpful image is that proteins are “DNA site explorers” that move near DNA and spend different amounts of time near each site (Berg et al., 1981). The exploration yields an average allocation over time near each DNA site. We choose to view this as a resource allocation problem and focus only on estimating the average allocation of proteins near DNA sites, not the way it is reached.

The transport mechanism invalidates the assumption of homogeneity of the nucleosol, which is often invoked to define the binding subproblem using chemical equilibrium concepts (*cf.* Shea and Ackers (1985); Arkin et al. (1998)). By introducing the transport mechanism explicitly we can relax that assumption. The transport mechanism allocates proteins to environments that are chemically well-mixed. This

---

\(^1\)In the language of Arrow and Debreu (1954) we should clearly identify the agents as distributors/suppliers and individuals and the commodities traded. In the DNA-binding economy the commodity traded is binding time, with sites being suppliers and proteins being consumers.
interpretation gives us control over the scale at which we examine binding. We will later take advantage of this degree of freedom, and propose three different versions of the DNA binding economy. As an example, consider an economy in which the transport mechanism brings proteins to the “neighborhood” of sites (defined functionally as DNA sequences where one protein binds). This definition for the transport mechanism implies that we assume that in a steady state, the local environment of the site is a closed and well-mixed system.

3.2.3 Site-Agent: the Binding Subproblem

The binding subproblem is the local counterpart to transport. Once proteins are transported near their respective target regions (sites, promoters, etc), a binding competition takes place. It is customary to assume that bindings are determined by statistical thermodynamic properties at equilibrium, (cf. Shea and Ackers (1985); Arkin et al. (1998)).

Controlling the limits of transport mechanism and binding subproblem is a convenient way to set the granularity of our models. It allows us to quantify protein allocations in steady state applying the law of mass action (i.e. chemical equilibrium equation) to all the species (proteins and sites) present in the region we have assumed to be chemically well-mixed.

3.2.4 Competitive Equilibrium and Steady State Behavior

Competitive equilibrium is reached when all agents in the economy have no incentive to allocate differently on average, given fixed availability of resources. That is, competitive equilibrium describes a steady state in which all agents allocate the resources available to maximize their utility function. A shift in the availability of resources will drive the economy to a different steady state described by a different allocation at competitive equilibrium.

Conceptually, in the DNA binding economy each steady state corresponds to a regulatory profile. It represents the allocation the system would attain given sufficient
time if the protein/site resources remained fixed. The competitive equilibrium is of biological interest because it defines a global state of the system from the "preferences" of each individual agent. The global state could be equivalently obtained by minimization of a global energy function. Using the energy function would translate into a \( nm \)-body (\( n \) sites, \( m \) proteins) problem requiring further assumptions about the nature of interactions between agents. The concept of competitive equilibrium, however, captures interactions succinctly; it does not require additional interaction terms, these are implicit in the definition of equilibrium and explicit in the utility functions: agents allocate optimally conditioned on the strategies (allocations) of other agents.

In some cases we may be able to assume that protein production and degradation occur at longer time-scales than the combination of transport and binding. Then the combination of transport and binding mechanisms yields a steady state that lasts while protein concentrations are fixed. In this interpretation, binding profiles can be interpreted as time-averages, and changes in the availability of resources drive the system through a sequence of steady states. This assumption is analogous to imposing reversibility in statistical mechanics (\textit{cf.} R. H. Perry (1997); Atkins (1986); DIU et al. (1989)). Many processes in biology and chemistry can be cast as reversible successions of steady states, for example, smooth addition of reactants to a system at chemical equilibrium drives the system into a succession of equilibrium states. By analogy, we see the arrangement of regulators at competitive equilibrium as defining a regulatory state.

Alternatively, binding profiles may be interpreted as averages over a population of cells. This interpretation does not assume anything about protein availability. As in statistical mechanics, the ergodicity principle is needed for population averages to approximate time averages.

### 3.3 Three DNA Binding Economies

We define three different binding economies that place different emphasis on the transport mechanism. From a computational standpoint we do so by imposing differ-
ent resource allocation models; from a game-theoretic standpoint our models explore
the tradeoff between purely competitive processes and processes that involve explicit
coordination.

3.4 A Site-Specific Economy

This version of the DNA binding economy is a purely competitive framework. How-
ever, as the empirical analysis will later demonstrate, the model can capture coordi-
nation as a byproduct of competition and resource availability.

3.4.1 Assumptions

The main assumptions in this economy are:

- the transport mechanism brings a protein next to the site where it eventually
  binds. An associated “affinity” parameter ($E_{ij}$) summarizes the ability of each
  site to attract proteins to its surroundings. The affinity parameter can be inter-
  preted as an activation energy, summarizing the contributions from factors such
  as the conformational structure and local environment of the site. Additionally
  the ability of the transport mechanism to bring proteins to the surroundings of
  a site is modulated by how occupied the site is.

- The immediate neighborhood of sites are closed and well mixed systems. That is
  chemical interactions at each site are considered independently from the chem-
  ical interactions occurring at other sites. We are free to assume so because
  the transport mechanism accounts for the nature of competition of resources
  between sites. The limiting reagent is determined by both, overall resource
  availability and site accessibility.

- Coordination, that is complex formation and joint binding events, is only im-
  plicitly encoded as a by-product of competition and resource allocation. Note
  that coordination could be included explicitly here as well. For simplicitly, and
to illustrate the trade-off between competition and resource allocation we choose not to introduce it here.

Figure 3-2 is a schematic representation of the competitive equilibrium of a site specific DNA binding economy.

Figure 3-2: Schematic representation of the site specific economy. Shaded areas represent the locally closed and well mixed systems that result from our choice of transport mechanism. DNA is represented as a line for simplicity. In practice affinities may encode the tertiary structure of DNA.

3.4.2 Utility Functions

Transport Mechanism: Protein Utility.

Protein-agents allocate proteins near sites based on their affinity to explore each site, modulated by how available the sites are. The higher the affinity the easier the approach. Affinities are specific to every protein-site pair and depend on the environmental conditions. This is consistent with viewing the transport mechanism as a diffusion process. We control the trade-off between non-specific binding and diffusion adding a term that rewards close-to-uniform distributions.

The utility function that describes this reasoning for an arbitrary protein $i$ is

$$u_i(p_i, s) \equiv \sum_j p_{i,j} E_{ij} \left( 1 - \sum_{j'} s_{j',i'} \right) + \beta \sum_j p_{i,j} \log \left( \frac{1}{p_{i,j}} \right), \quad (3.1)$$

The optimal strategy for a protein-agent is to allocate first to sites for which it
has higher affinity, and as these are filled, start allocating to sites with lower affinity.

\( \beta \) dimensions the affinities, and controls the specificity of the allocation.

The strategy set for any protein \( i \) defines protein allocation as a distribution with 
\[ 1 - \sum_j p_{i,j} \] unallocated (“free”) protein.

\[ \mathcal{P}_i \equiv \{ p_i : \sum_j p_{i,j} = 1, p_{i,j} \geq 0, \text{ for all } j \} \]

Note that the utility function for a given protein-player depends on the strategy of the site-players and is independent of the strategy of other protein-players.

**Chemical Equilibrium: Site Utility**

We have assumed that *in steady state*, the local environment of the site is a *closed* and *well-mixed* system. This assumption is consistent with the notion that inhomogeneities in the cell are due to transport and handled by the protein-utility. As discussed earlier (cf. Section 3.4.1), this assumption is necessary to apply equilibrium thermodynamic properties, and is less restrictive than the usual hypotheses (cf., Shea and Ackers (1985); Arkin et al. (1998)). Accordingly, we quantify site allocations\(^2\), \( s_{j,i} \)'s, using the binding reaction between each site and the proteins in its surroundings,

\[ i_{\text{PROTEIN}} + j_{\text{SITE}} \rightleftharpoons ij_{\text{BINDING}}. \]  

\[ (3.2) \]

At chemical equilibrium the concentrations of reactants and products are related by a binding equilibrium constant \( K_{ij} \) according to the law of mass action (*i.e.*, chemical equilibrium equation)

\[ \frac{[ij_{\text{BINDING}}]}{[i_{\text{PROTEIN}}][j_{\text{SITE}}]} = K_{ij}. \]  

\[ (3.3) \]

Using the notation from Table 3.2, we identify site allocation \( s_{j,i} \) with \([ij_{\text{BINDING}}]\), and enforce balance of mass through the availability of unbound protein \([i_{\text{PROTEIN}}]\) as \( (p_{i,j} - s_{j,i}) \), and the availability of site \([j_{\text{SITE}}]\) as \( (1 - \sum_{\nu} s_{j,\nu}) \). The chemical

\( \text{We are using a special convention to order subindex. In } s_{j,i} \text{, the first subindex indicates that the vector } s_j \text{ is a variable of site-agent } j \text{, the second indicates that vector } s_j \text{ runs over proteins } i. \) Similarly for protein-agent \( i \), with allocation \( p_{i,j} \), \( p_i \) is a vector-valued variable indexed by sites \( j \).
equilibrium equation then reads

$$\frac{s_{j,i}}{(p_{i,j}f_i - s_{j,i})(1 - \sum_{i'} s_{j,i'})} = K_{ij}. \quad (3.4)$$

In section 6.1.2 we quantify the units in this equation and relate $K_{ij}$ to Gibbs’ free energy.

We construct the sites’-utility function reordering the terms in the chemical equilibrium equation (3.4), and summing over all proteins. The site-agent allocates bindings “slots” to remain at chemical equilibrium with all the chemical species in its surroundings. Put in a different way, the site-agent allocates bindings to maximize binding to each protein as permitted by the chemical equilibrium equation,

$$v_j(s_j, p_j) \equiv \sum_i s_{j,i} - K_{ij}(p_{i,j}f_i - s_{j,i}) \left(1 - \sum_{i'} s_{j,i'}\right). \quad (3.5)$$

To completely determine the optimal strategy we will constrain the strategy space so that if a binding equation does not reach chemical equilibrium there is an excess of reactants rather than an excess of products.

The corresponding strategy set ($S^j(p_j)$) is defined as the set of allocations $s_j$ (i.e., $s_j \in S^j(p_j)$) that meet the following three constrains:

I. $s_{j,i} \leq K_{ij}(p_{i,j}f_i - s_{j,i})(1 - \sum_{i'} s_{j,i'})$

II. $s_{j,i} \leq p_{i,j}f_i$

III. $\sum_{i'} s_{j,i'} \leq 1.$

The first constraint states that if a binding equation does not reach chemical equilibrium there will be an excess of reactants rather than an excess of products. This constraint emphasizes that site availability is the limiting resource. Constraints II and III are mathematical statements that sites cannot allocate more protein than is available in their surroundings, and, sites have limited availability (it is useful to think of a site as having a limited number of time-slots, and site allocation $s_{j,i}$ as a rate of occupancy.)
Figure 3-3: Bipartite structure of the site specific economy. Nodes represent agents and edges represent interactions between agents \((K_{ij}, E_{ij})\). Strategies for each site only depend on the strategies of proteins. Shaded circles emphasize dependencies between site and protein agent strategies, partial shading indicates indirect dependencies. In the diagram, the strategy of site \(j_1\) depends on the strategies of protein-agents \(i_1\) and \(i_2\) but not protein-agent \(i_3\). Site \(j_1\) only depends on the strategy of sites \(j_3\) and \(j_1\) through the strategy of protein-agents \(i_1\) and \(i_2\).

By construction, the sites’-utility (3.5) guarantees that competitive equilibrium of the DNA-binding economy is characterized by sites being at chemical equilibrium. To see that this is the case note that (3.5) has a maximum at \(s_{j,i} = 0\) for all \(i\) and that then each term in the sum is simply a reorganization of each equilibrium equation. We also emphasize that the strategy set of each site depends on the strategies of the proteins in its surroundings and not on the other sites.

As a result of our choices for site and protein utilities, the DNA-binding Economy has the structure of a bipartite graph (shown in Figure 3.4.2) protein strategy depends on other proteins only indirectly through the sites they approach, and the same happens for sites. Furthermore, dependence on the strategy of other agents is solely mediated by how filled each site is; that is, in equations (3.1) and (3.5) only the term \(\sum_{s'} s_{j,s'}\) includes the strategy of other agents explicitly.

This structure is particularly useful to prove the existence of a competitive equilibrium and how to compute it. It can be used to summarize the maximization of the utility functions as a fixed point mapping.
3.4.3 Fixed Point

Let \( \alpha_j = \sum_s s_{j,s} \) denote site \( j \) occupancy, an indication of how “filled” the sites are. \( \alpha_j \)'s are real numbers in the interval \([0, 1]\). If we fix \( \alpha = (\alpha^1, \ldots, \alpha^m) \), i.e., the occupancies for all the \( m \) sites, we can readily obtain the maximizing strategies for proteins expressed as a function of site occupancies: \( p_{i,j}(\alpha) \propto \exp(E_{ij}(1 - \alpha_j)/\beta) \), where the maximizing strategies are functions of \( \alpha \).

Then, we can express the joint strategy at competitive equilibrium as a function of how filled the sites are, and reduce the problem of finding equilibria to finding fixed points of a monotone function.

Maximizing protein utility.

The protein-utility function is convex, it can be maximized using the method of Lagrange multipliers with the constraint \( \sum_j p_{i,j} = 1 \).

Specifically,

\[
L(p_{i,j}, \lambda; \alpha, \beta) = \sum_j p_{i,j} E_{ij} (1 - \alpha^j) - \beta \sum_j p_{i,j} \log p_{i,j} - \lambda \left( 1 - \sum_j p_{i,j} \right). \tag{3.6}
\]

To get an expression for \( p_{i,k} \) we take the derivative of the lagrangian with respect to \( p_{i,k} \) and set it to zero:

\[
\frac{\partial L(p_{i,j}, \lambda; \alpha, \beta)}{\partial p_{i,k}} = E_{ik} (1 - \alpha^k) - \beta (\log p_{i,k} + 1) + \lambda = 0, \tag{3.7}
\]

by rearranging terms,

\[
p_{i,k} = e^{\frac{\lambda - \beta}{\beta}} \times e^{\frac{1}{\beta} E_{ik}(1 - \alpha^k)}, \tag{3.8}
\]

and eliminating the dependence on \( \lambda \) using the constraint, we obtain

\[
p_{i,k} = \frac{e^{\frac{1}{\beta} E_{ik}(1 - \alpha^k)}}{\sum_j e^{\frac{1}{\beta} E_{ij}(1 - \alpha^j)}}, \tag{3.9}
\]
Maximizing Site Utility.

By definition (at competitive equilibrium,) each site-player achieves a local chemical equilibrium specified in (3.4). By replacing $\alpha_j = \sum_{i'} s_{j,i'}$, and solving for $s_{j,i}$ in the law of mass action (3.4), we get

$$s_{j,i}(\alpha) = \frac{K_{ij}(1 - \alpha_j)}{1 + K_{ij}(1 - \alpha_j)} p_{i,j}(\alpha) f_i. \quad (3.10)$$

So, the fraction of time the site is bound by a specific protein is proportional to the amount of that protein in the neighborhood of the site, modulated by the equilibrium constant. Note that $s_{j,i}(\alpha)$ depends not only on how filled site $j$ is, but also on how occupied the other sites are through $p_{i,j}(\alpha)$.

Equations (3.10) and (3.9) redefine the DNA binding Economy in terms of $\alpha$. Given an optimal $\alpha$, these relations give us the optimal arrangement of proteins along DNA.

The equilibrium condition can be now expressed solely in terms of $\alpha$ and reduces to a simple consistency constraint: overall occupancy should equal the fraction of time any protein is bound:

$$\alpha_j = \sum_i s_{j,i}(\alpha) = \sum_i \frac{K_{ij}(1 - \alpha_j)}{1 + K_{ij}(1 - \alpha_j)} p_{i,j}(\alpha) f_i = G^i(\alpha). \quad (3.11)$$

We have therefore reduced the problem of finding equilibria of the game to finding fixed points of the mapping $G^i(\alpha) = \sum_i s_{j,i}(\alpha)$.

### 3.5 A Promoter-Specific Economy

In this version of the DNA binding economy we refine our assumptions about independence between sites. The assumption that all sites can be dealt with independently was convenient as a first approximation to the formalization of transport and binding. Nonetheless, the assumption may be overly simplistic to properly describe various regions of the genome. We assume here that the transport mechanism dis-
tributes proteins to regions of DNA rather than individual sites. Biologically, regions of possibly overlapping DNA sites can be associated with promoters. The protein allocated to a promoter region is shared by all the sites within the promoter region. The sharing creates specific challenges to the algorithms to finding the competitive equilibrium.

### 3.5.1 Assumptions

The main assumptions in this version of the economy are:

- As before, the transport mechanism is associated with an “affinity” parameter \( E_{ij} \) that can be interpreted as an activation energy. It summarizes the contributions from factors such as the conformational structure and local environment of a promoter region. The ability of the transport mechanism to bring proteins to a promoter region is now modulated by the level of occupancy of the sites within the promoter region.

- Promoters, rather than sites, are the closed and well mixed systems.

- Neither sites nor promoters overlap.

- Coordination, complex formation and joint binding events, are implicitly encoded as a by-product of competition and resource allocation.

Figure 3-4 is a schematic representation of the competitive equilibrium of a promoter specific DNA binding economy. Compare with Figure 3-2.

### 3.5.2 Formal Description

Let \( \mathcal{R} \) represent possible promoter regions \( (r \in \mathcal{R}) \) each of which may be bound by multiple proteins.

We index sites as in the site specific economy and \( s_{j,i} \) represents the occupancy of some site \( j \) by protein \( i \).
Let $p_i = \{p_{i,r}\}_{r \in \mathcal{R}}$ represent an allocation of protein $i \in \mathcal{P}$ into promoter regions in a manner that is not specific to the possible sites within each promoter.

Let $N(r)$ $r \in \mathcal{R}$ be the set of possible binding sites within promoter region $r$. $N(r) \cap N(r') = \emptyset$ whenever $r \neq r'$ (i.e., we assume promoters do not share sites).

We define also auxiliary variables for binding and occupancy rates.

The binding rate (or occupancy rate) of site $j$ $\alpha_j = \sum_{i \in \mathcal{P}} s_{j,i}$

The occupancy rate of promoter $r$ $a_r = \sum_{j \in N(r)} \alpha_j$

### 3.5.3 Utility Functions

We study economies with utility functions of the following general form

**Transport Mechanism**

The utility for protein $i$ is given by

$$u_i(p_i) = \sum_{r \in \mathcal{R}} p_{i,r} E_{ir}(a_r) + \beta H(p_i)$$

We assume only that $E_{ir}(a_r)$ is a decreasing and differentiable function of $a_r$. The protein utility is based on the assumption that attraction to the promoter decreases as the number of proteins already bound at the promoter grows. The maximizing strategy for protein $i$ given $a_r = \sum_{j \in N(r)} \alpha_j$ for all $r$, is $p_{i,r}(a) \propto \exp(E_{ir}(a_r)/\beta)$.
Binding Mechanism

Sites \( j \in N(r) \) within a promoter region \( r \) are at chemical equilibrium, hence follow the law of mass action

\[
\frac{s_{j,i}}{(f_i p_{i,r}(a) - \sum_{k \in N(r)} s_{k,i})(1 - \alpha_j)} = K_{ij}
\]

The site occupancies \( \alpha_j \) are now tied within the promoter as well as influencing the overall allocation of proteins across different promoters through \( a = \{a_r\}_{r \in R} \).

3.5.4 Fixed Point

We can derive expressions for the occupancy rates of sites, neighborhood and promoter regions from the law of mass action. We start rewriting the law of mass action

\[
\frac{s_{j,i}}{(f_i p_{i,r}(a) - \sum_{k \in N(r)} s_{k,i})(1 - \alpha_j)} = K_{ij}
\]

Rearranging terms

\[
s_{j,i} = (1 - a_{r,j})K_{ij}p_{i,r}(a)f_i - \left( \sum_{k \in N(r)} s_{k,i} \right) K_{ij} (1 - \alpha_j)
\]

summing over \( j \in N(r) \)

\[
\sum_{k \in N(r)} s_{k,i} = p_{i,r}(a)\sum_{j \in N(r)} K_{ij} (1 - \alpha_j)
\]

\[
- \left( \sum_{k \in N(r)} s_{k,i} \right) \sum_{j \in N(r)} K_{ij} (1 - \alpha_j)
\]

and summing over proteins \( i \in P \) we reach an expression for the binding rate of each promoter region

\[
a_r = \sum_{i \in P} \frac{\sum_{j \in N(r)} K_{ij} (1 - \alpha_j)}{1 + \sum_{j \in N(r)} K_{ij} (1 - \alpha_j)} p_{i,r}(a)f_i
\]
plugging these results back into the law of mass action we reach an expression for the allocation to site $s_{j,i}$

$$s_{j,i} = \frac{K_{ij}(1 - a_{r,j})}{(1 + \sum_{j \in N(r)} K_{ij}(1 - a_{r,j}))} p_{i,r}(a_r) f_i \quad \text{for } j \in N(r) \quad (3.13)$$

which, summing over proteins yields the binding rate of site $j$

$$\alpha_j = \sum_{i \in P} \frac{K_{ij}(1 - \alpha_j)}{(1 + \sum_{j \in N(r)} K_{ij}(1 - \alpha_j))} p_{i,r}(a_r) f_i. \quad (3.14)$$

Equation (3.14) is a fixed point equation on $\alpha_j$. As we show in Chapter 4 solving this fixed point equation is slightly more difficult than for the site-specific DNA binding economy because this equation is no longer an increasing function of $a^{-r}$.

We summarize the fixed point below,

$$\alpha_j = \sum_{i \in P} \frac{K_{ij}(1 - \alpha_j)}{1 + \sum_{j \in N(r)} K_{ij}(1 - \alpha_j)} p_{i,r}(a_r) f_i \equiv G_r^j(\alpha_j, \alpha^{-j}, a^{-r}) \quad (3.15)$$

### 3.5.5 Considerations about the Promoter Specific Economy

The promoter specific version of the DNA binding economy introduces a new structure of dependencies. Each site’s occupancy depends directly on the occupancy of other sites within the same promoter region with which it competes and, through transport mechanism, on the occupancy of other promoter regions. The additional structure captures the notion that binding at each site should only be affected by binding at farther sites through transport mechanism. Technically, the promoter structure makes the problem more feasible to estimate as it reduces the set of parameters required to describe transport.

In its more general form the promoter specific economy as introduced here opens up the possibility to extend the algorithms with sequence specific protein motifs and
different levels of overlap. Indeed, allowing overlapping between sites to distinguish between totally and partially bound protein can be seen as a way to introduce coordination between sites.

As measurement techniques with increased resolution improve (cf. Johnson et al. (2007)) it will be possible to consider refinements to the promoter specific economy that use protein motifs and varying levels of site overlap. Some of these refinements are discussed further in Chapter 8.

### 3.6 A Nucleus-Specific Economy

The site and nucleus specific DNA binding economies consider purely competitive settings to introduce the transport mechanism with different levels of granularity. Regulators, however, often exhibit coordinate behavior; for instance, when binding as complexes. This version of the economy considers the case of coordinate binding. To simplify the introduction of complexes, in this version of the game we assume that effects of the transport mechanism are negligible and consider the entire nucleus of the cell to be a closed and well-mixed system.

#### 3.6.1 Assumptions

The main assumptions in this version of the economy are:

- no transport mechanism is considered, all proteins present in the nucleus are available to all the sites.

- The entire nucleus is considered to be a closed and well mixed system.

- Coordination, that is complex formation and joint binding events, is considered explicitly. Proteins may bind DNA both individually and/or via complex formation.

Figure 3-5 is a schematic representation of the competitive equilibrium of a nucleus specific DNA binding economy. Compare with Figures 3-2 and 3-4.
3.6.2 Formal Description

Let $C_j$ be a subset of complexes $c$ ( $c \in C_j$) that can bind to site $j$. Each complex $c$ is defined as a set of proteins $i \in \mathcal{P} = \{1,...,n\}$, so that $c \subseteq \mathcal{P}$. For notational simplicity we use the word complex when referring to a single protein as well. So when protein $i$ can bind a site individually we consider the “singleton” complex $c = \{i\}$. Note that with these definitions, $C_j$ is a collection of sets, and in the worst case corresponds to the powerset of all proteins. In most cases, though, it will be possible to significantly reduce the cardinality of $C_j$ based on observations or literature (hence $C_j \subset PowerSet(\mathcal{P})$).

Let $P_i$ be the set of all complex-site pairs $(c, j)$ that involve protein $i$, that is $P_i = \{(c, j) \mid c \in C_j, i \in c\}$. Finally, let $s_{j,c}$ denote the rate of binding of complex $c$ at site $j$.

In this version of the economy, DNA-binding is solely driven by chemical equilibrium. We can write the chemical reaction of a complex at a site as

$$c(1) + c(2) + \cdots + c(m) + j_{SITE} \rightleftharpoons c_j^{BINDING}$$

where $c(1)$ selects the first protein in the complex $c$. At equilibrium we obtain the concentrations of proteins and binding in accordance with the following expression of the law of mass action

$$\frac{[bound\ c_j]}{[free\ protein\ c(1)] \cdots [free\ protein\ c(m)] [free\ site\ j]} = K_{c_j} \quad (3.16)$$
Where $K_{cj}$ is a constant and the notation $[·]$ stands for concentration (interpreted as a proportion of a population, a time average, or a measure of units per volume).

### 3.6.3 Utility functions

#### Transport Mechanism

Since we have assumed that the whole cell is a well mixed-closed system this economy does not consider a transport mechanism. Abusing notation we can write a chemical reaction for the whole of DNA as:

$$
\sum_i \nu_i \text{PROTEIN} + \sum_j \text{j_{DNA-sites}} \rightleftharpoons \sum_c c_j \text{BOUND}, \tag{3.17}
$$

where $\nu_i$ is the stoichiometric coefficient of protein $i$ in the DNA reaction; it equals the number of appearances of protein $i$ in any complex, $\nu_i = \sum_c [i \in c]$.

#### Binding Mechanism

Using $s_{j,c}$ to denote the rate of binding of complex $c$ at site $j$, and $f_i$ to denote the amount of protein $i$ available, we may rewrite the expression at equilibrium as

$$
s_{j,c} \prod_i (f_i - \sum_j \sum_{c' \in C_j} s_{c'j})(1 - \sum_{c' \in C_j} s_{c'j}) = K_{cj} \tag{3.18}
$$

which can be expressed as a utility function for each site $j$,

$$
v_j(s^i) = \sum_{c \in C_j} s_{j,c} - K_{cj} \prod_i (f_i - \sum_j \sum_{c' \in C_j} s_{c'j})(1 - \sum_{c' \in C_j} s_{c'j}) \quad \forall \text{ site } j \tag{3.19}
$$

with strategy set

I. $s_{j,c} \leq K_{cj} \prod_i (f_i - \sum_j \sum_{c' \in C_j} s_{c'j})(1 - \sum_{c' \in C_j} s_{c'j})$

II. $\sum_{j \in C_j ; i \in e} s_{j,c} \leq f_i$

III. $\sum_{c \in C_j} s_{j,c} \leq 1$
3.6.4 Fixed Point

We can obtain a fixed point relation that maximizes the utility function for sites. We introduce two new variables that simplify the derivation

\[ \beta_j = 1 - \sum_{c \in C_j} s_{j,c} \quad \text{site } j \text{ availability} \quad (3.20) \]

\[ R_i = f_i - \sum_j \sum_{c \in C_j : i \in c} s_{j,c} \quad \text{Free protein } i. \quad (3.21) \]

The equilibrium equation then becomes

\[ \frac{s_{j,c}}{\prod_{t \in c} (R_t)} \beta_j = K_{cj} \quad (3.22) \]

which yields the following expressions for \( \beta_j, s_{j,c}, \) and \( R_i \)

\[ s_{j,c} = K_{cj} \beta_j \left[ \prod_{c \in e} (R_c) \right] \]

\[ \beta_j = 1 - \sum_{c \in C_j} s_{j,c} = 1 - \sum_{c \in C_j} K_{cj} \beta_j \left[ \prod_{i \in c} (R_i) \right] \]

\[ R_i = f_i - \sum_{j, c \in C_j : i \in c} s_{j,c} = f_i - \sum_{j, c \in C_j : i \in c} K_{cj} \beta_j \left[ \prod_{t \in c} (R_t) \right] \]

\[ = f_i - \left( \sum_{j, c \in C_j : i \in c} K_{cj} \beta_j \left[ \prod_{t \in c \neq i} (R_t) \right] \right) R_i \quad (3.23) \]

Rearranging terms above we reach an expression for a fixed point in \( R_i \)

\[ \beta_j = \frac{1}{1 + \sum_{c \in C_j} K_{cj} \left[ \prod_{t \in c} (R_t) \right]} \quad = B_j(R) \quad (3.24) \]

\[ R_i = f_i \left( \sum_{j, c \in C_j : i \in c} 1 + K_{cj} \beta_j \left[ \prod_{t \in c \neq i} (R_t) \right] \right)^{-1} \quad = G_i(\beta, R) \quad (3.25) \]
3.6.5 Considerations about the Nucleus Specific Binding economy

In the absence of assumptions about the transport mechanism we have taken the scope to be the whole of the nucleus. However we could also have taken it to be one promoter region. Then we could use this economy as a replacement for the binding mechanism in the promoter specific version of the DNA binding economy. Under this reinterpretation equation (3.17) is corrected as

\[ \sum_i \nu_i i_{\text{PROMOTER-PROTEIN}} + \sum_{j \in r} j_{\text{PROMOTER-SITES}} \rightleftharpoons \sum_{c,j} c_j \text{BINDING:} j_{\in r}, \]  \tag{3.26}

Alternatively, we can reinterpret complexes as joint binding events, and sites as promoter regions. At any point in time a collection of proteins is bound on distinct sites within the promoter. Then this DNA binding economy fully replaces the promoter specific version of the economy. Equilibrium constants \((K_{c,j})\) in this case summarize affinities and site specific equilibrium constants, and the Nucleus specific economy stands for a coarse description of DNA-binding:

\[ \sum_i \nu_i i_{\text{PROTEIN}} + \sum_r r_{\text{PROMOTER}} \rightleftharpoons \sum_{c,r} c_{r} \text{BINDING:}, \]  \tag{3.27}

3.7 Fixed Point Equations or Economy?

In each abstraction we found a fixed point equation that describes the joint strategy at competitive equilibrium. One could argue that approaching the problem as an economy was unnecessary. A similar statement can generally be made about any game-theoretic model as the notion of equilibrium in an economy or a game is tightly connected with that of fixed point.

The value of using game-theoretic and economic concepts resides in the level of abstraction, which eases the interpretation of biological concepts in computational terms. The fixed points we reached for each of the economies hide the underlying biological mechanisms that inspired our approach. For instance, the granularity in-
roduced in the economic approach via the different interpretations of the transport mechanism is more explicit in the economic viewpoint.

We could also have reached similar expressions for the equilibrium strategies by minimizing a global energy function. Indeed we shall do so when introducing learning in the nucleus specific version of the game. The economic formalism emphasizes the structure of the problem by making explicit player preferences. The concept of competitive equilibrium summarizes the interdependence between the optimizations local to each agent. This abstraction suits the notion that proteins are affected by their local environment, rather than controlled by an unobserved mechanism to remain at a global energy minimum.

Finally, the economic formalism provides a conceptual connection with statistical mechanics that is convenient to reason about chemical systems. Like chemical equilibrium, the concept of competitive equilibrium is a dynamic concept; it summarizes an average behavior that results from reaching a state in which the exchange of commodities has reached a steady state. Chemical equilibrium dynamics abstract kinetics and focus in average allocation of chemical species when reaction happens at the same rate in both directions. Similarly competitive equilibrium summarizes the average distribution of commodities when all the agents in the economy have no incentive to change their average allocation of resources. In both cases, the concept of equilibrium is intimately tied to the notion of fixed point. And the choice for one interpretation over the other is intimately connected with the purpose of the analysis. In chemistry, the fixed point interpretation is convenient to study the evolution of the equilibrium as the concentration of some species changes. Also here, the concept of fixed point for each economy proves valuable to quantify the effect of each change in the concentration of chemical species.

We find the fixed point interpretation convenient also when learning these models from data. There $f$’s as well as $K$’s and $E$’s change.

3.7.1 Economic Interpretation

The DNA binding economies can be cast in the language of abstract economies.
Sites are production units. In the time-average interpretation of the steady state, each site produces binding “time-slots” as a commodity. Binding supply is indicated by the variable $\alpha_j$ in the site and nucleus specific economies, and by the variable $a^r$ in the promoter-specific economy.

Protein-agents are consumption units. Each protein agent consumes binding-time slots in exchange for allocating sufficient proteins to the surroundings of the site, the promoter region, or the nucleus. The initial amount of protein available can be interpreted as the income of each protein-agent. The demand for each binding time is given by the functions $G_j$ in the site specific economy, $G_i^j$ in the promoter specific economy, and $(f_i - G_i)$ in the nucleus specific economy. The demand function results from combining the concept of chemical equilibrium and the balance of mass.

Competitive equilibrium results when each agent has maximized his own utility function. The fixed point relation is then simply the result of balancing supply $\alpha_j$ and demand $G$ at that maximum, given a fixed (announced) equilibrium price. In the DNA binding economies we have not considered the notion of price explicitly. We may assume that it is fixed to one or given as an implicit function of the chemical equilibrium constants.

### 3.7.2 Advantage over Traditional Reaction-Equation models

The introduction of the economic abstraction simplifies the study of the set of joint chemical reactions that define protein-DNA binding genome-wide in two ways: 1) through the definition of the transport mechanism and 2) the competitive equilibrium.

The transport mechanism, defines the inhomogeneities in the chemical mix of the nucleosol. These inhomogeneities may be due to the high specificity of binding, dynamic changes in site availability, and other bindings and conformational changes in DNA. Our abstraction of the transport mechanism allows us to encode these explicitly in the notion of “affinity”. Furthermore, modeling the transport mechanism as a resource allocation problem allows us to model the interdependence between DNA subsystems (i.e., groups of sites that participate in the same regulatory process) explicitly as a consequence of shifts in local availability of sites and proteins. This
interdependence has been observed (v., Herman et al. (1993)) and can only be explained using the concept of chemical equilibrium in combination with the notions of limiting reagent and reallocation of excess reagent. Our abstraction of the transport mechanism provides us with a transparent way to connect these notions under global resource constraints resource (sites and proteins).

In the case of the nucleus specific economy we neglect the effects of the transport mechanism. The economic abstraction, nevertheless, offers also an advantage over minimizing a global function based on Gibbs Energies. While minimizing a global Gibbs energy function can be readily accomplished for a reduced set of joint reactions, defining a global function that takes into account pairwise interactions for a larger set of reactions is challenging. The concept of competitive equilibrium simplifies that task in the case of the nucleus specific economy, and makes it possible in the case of the site and promoter specific economies, for which defining a global Gibbs energy function would require considering a large number of interaction terms. The competitive equilibrium defines a global optimum propagating constraints imposed by the preferences local to each agent under global constraints.

As we show in Chapters 4 and 5, the concept of competitive equilibrium operates at a convenient level of abstraction to compute steady state allocations, and to define the learning task to estimate chemical parameters and coordinate control from high-throughput data. Accomplishing both tasks using traditional methods would require introducing several assumptions about coordinate control and pairwise interactions in the expression of the global Gibbs energy function. In most cases, however, finding an expression that describes the behavior observed globally is a challenge. We believe that it is easier to encode such assumptions in the utility functions and rely on the concept of competitive equilibrium to propagate the effect of the assumptions globally.

Taken together, transport mechanism and the notion of competitive equilibrium provide the economic abstraction with greater flexibility to model protein-DNA binding at different scales. For instance, using model selection, to determine the scale at which to model each region of DNA (choosing over site, promoter and nucleus specific economies locally), or viewing the problem of protein-DNA binding as as involving
aspects of all three economies.

3.8 Summary

This chapter introduced three computational abstractions of DNA-binding. Each of the economies required additional notation to capture the different levels of detail they model. In subsequent chapters (except in the experimental section) the discussion about DNA binding focuses in computational rather than biological concepts. To ease the exposition we summarize here the general mapping from biology to economic theory and all the notation introduced.

3.8.1 Conceptual Mapping

Table 3.1 summarizes the conceptual mapping from biology to an abstract economy.

<table>
<thead>
<tr>
<th>Biological Concept</th>
<th>DNA binding economies</th>
<th>Abstract Economy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>Protein-agent</td>
<td>Consumption Unit</td>
</tr>
<tr>
<td>“transport mechanism”</td>
<td>“allocates protein to sites/promoters”</td>
<td></td>
</tr>
<tr>
<td>Sites</td>
<td>Site-agent</td>
<td>Production Unit</td>
</tr>
<tr>
<td>“chemical equilibrium”</td>
<td>“allocates proteins to binding-time-slots”</td>
<td></td>
</tr>
<tr>
<td>Steady state</td>
<td>Competitive equilibrium</td>
<td>supply = demand</td>
</tr>
</tbody>
</table>

Table 3.1: Conceptual mapping from biology to economic theory. The mapping of protein-DNA binding onto an abstract economy requires the definition of the utilities of each type of agent and a proof that an equilibrium exists and can be computed.

3.8.2 Conventions

Throughout this Chapter we have adhered to the convention that $i$ indexes proteins, $j$ sites, $r$ promoter regions, and $c$ complexes. The following table summarizes the notation introduced for each economy

73
Site specific economy

\[ s_{i,j} \] binding rate of protein \( i \) at site \( j \).

\[ p_{i,j} \] Fraction of protein \( i \) in the neighborhood of site \( j \).

\[ \alpha_j \] total binding rate at site \( j \).

\[ u_i(p_i) \] utility function of protein agent.

\[ v_j(s_j) \] utility function of site agent.

\[ f_i \] Total protein \( i \) available in the cell/nucleus.

\[ E_{ij} \] Affinity of protein \( i \) for site \( j \) (Activation Energy).

\[ K_{ij} \] Chemical equilibrium constant for the binding of protein \( i \) and site \( j \)

\[ i_{\text{PROTEIN}} + j_{\text{SITE}} \rightleftharpoons i_j \text{BINDING} \]

\[ (\bar{p}, \bar{s}) \] allocations at equilibrium

Promoter specific economy

\[ \mathcal{R} \] set of promoter regions.

\[ s_{j,i} \] binding rate of protein \( i \) at site \( j \).

\[ p_{i,r} \] Fraction of protein \( i \) allocated to promoter \( r \).

\[ \alpha_j \] total binding rate at site \( j \).

\[ \alpha_{r,j} \] occupancy rate of site \( j \in r \).

\[ \alpha_r \] occupancy rate of promoter \( r \).

\[ u_i(p_i) \] utility function of protein agent.

\[ v_j(s_j) \] utility function of site agent.

\[ f_i \] Total protein \( i \) available in the cell/nucleus

\[ E_{ir} \] Affinity of protein \( i \) for site \( j \) (Activation Energy)

\[ K_{ir} \] Chemical equilibrium constant for protein \( i \) at site \( j \)

\[ i_{\text{PROTEIN}} + j_{\text{SITE}, j \in r} \rightleftharpoons i_j \text{BINDING}, j \in r \]

\[ (\bar{p}, \bar{s}) \] allocations at equilibrium

Nucleus specific economy

\[ C_j \] set of complexes \( c \) that bind at site \( j \).

\[ c \] complex, defined as a set of proteins \( \{i\} \).

\[ P_i \] set of complex-site pairs involving protein \( i \)

\[ P_i = \{(c, j) | c \in C_j, i \in c\} \]

\[ s_{j,c} \] binding rate of complex \( c \) at site \( j \).

\[ \alpha_j \] total binding rate at site \( j \).

\[ \beta_j \] availability of site \( j \).

\[ R_i \] protein \( i \) that remains unbound.

\[ v_j(s_j) \] utility function of site agent.

\[ f_i \] Total protein \( i \) available in the cell/nucleus

\[ K_{cj} \] Chemical equilibrium constant for complex \( c \) at site \( j \)

\[ c(1) + \cdots + c(m) + j_{\text{SITE}} \rightleftharpoons c_j \text{BINDING}, c \in C_j \]

\[ (\bar{s}) \] allocations at equilibrium

Table 3.2: Notation
Chapter 4

Predictions of the Economies

We establish existence of competitive equilibria in pure strategies, under any set of conditions, for each economy, and provide algorithms to compute such equilibria.

Existence of a competitive equilibrium in pure strategies guarantees predictions about single binding allocation arrangements, rather than a linear combination of them (mixed strategy). Intuitively, a mixed strategy would imply a superposition of chemical equilibrium states for a fixed amount of chemical species that is not possible in practice. Proving existence of pure competitive equilibria is generally straightforward for abstract economies that meet certain convexity criteria Fudenberg and Tirole (1991); Arrow and Debreu (1954). Proving existence of a competitive equilibrium for the DNA binding economies, however, requires additional work, because the strategy set for the site-agents is non-convex.

Existence of an equilibrium under any conditions for each DNA binding economy guarantees predictions about the final allocation of proteins for any given set of parameters of the economy.

We develop fast algorithms specific to each DNA binding economy. Our algorithms use the monotonicity properties of the fixed point equations to converge iteratively to the competitive equilibrium. Each iteration of our algorithms is polynomial in the number of proteins and sites. The algorithms will be scalable to genome-wide scope. (note that game theory and economics is a formal framework to reason about the existence of equilibria, not how they are computed.)
4.1 Site Specific Economy

4.1.1 Equilibrium Existence

Finding a competitive equilibrium for the site-specific DNA-Binding economy is equivalent to solving the following coupled maximization:

\[
\max_{p_i} u_i(p_i) \quad \forall \text{ protein } i
\]

subject to

\[
\begin{align*}
\sum_j p_{i,j} &= 1 \\
u_i(p_i) &= \sum_j p_{i,j} E_{ij}(1 - \sum_i s_{j,i})
\end{align*}
\]

\[
\max_{s_j} v_j(s_j) \quad \forall \text{ site } j
\]

subject to

\[
\begin{align*}
s_{j,i} &\leq (p_{i,j} f_i - s_{j,i}) (1 - \sum_i s_{j,i}) K_{ij} \\
s_i^j &\leq p_{i,j} f_i \\
\sum_i s_i^j &\leq 1 \\
v_j(s_j) &= \sum_i s_{j,i} - K_{ij} (p_{i,j} f_i - s_{j,i}) (1 - \sum_i s_{j,i})
\end{align*}
\]

The first condition in the strategy set of sites is non-convex, and thus, we cannot directly apply previous work of Fudenberg and Tirole (1991); Arrow and Debreu (1954) to assert the existence of a competitive equilibrium. However, it is possible to show that an equilibrium exists for any setting of the parameters of the economy.

**Theorem 1.** Every site based DNA binding economy has an equilibrium.

The theorem can be proved using Brouwer’s fixed point theorem and constructively with an algorithm that relies on the monotonicity properties of the fixed point (cf. section 3.11)

\[
\alpha_j = \sum_i s_{j,i}(\alpha) = \sum_i \frac{K_{ij}(1 - \alpha_j)}{1 + K_{ij}(1 - \alpha_j)} p_{i,j}(\alpha) f_i = G^j(\alpha).
\]

Brouwer’s fixed point proof to theorem 1 is given in Appendix A.1. The constructive proof accompanies the description of the algorithm in section 4.1.3.
4.1.2 Specification and Use of the Economy

The site specific DNA binding economy is defined by the set of parameters \( \{E_{ij}\}, \{K_{ij}\}\), along with the utility functions \( \{u_i\} \) and \( \{v_j\} \) and the allocation constraints \( \{P^i\} \) and \( \{S^j\} \).

Given each \( \{E_{ij}\}, \{K_{ij}\}\) and \( \{f_i\} \), a competitive equilibrium exists defined by an average allocation of proteins \( (p, s) \) near or bound to DNA sites. We interpret one such allocation as a regulatory state. With an algorithm to compute a competitive equilibrium, we may use the economy in a variety of settings:

- with fixed \( \{E_{ij}\}, \{K_{ij}\}\), to track the evolution of a regulatory state (i.e., a competitive equilibrium) for varying availability of resources (i.e., changes in \( \{f_i\} \)).
- with fixed \( \{E_{ij}\}, \{f_i\}\), to reproduce site knock-out experiments altering the value of \( \{K_{ij}\} \).
- with fixed \( \{K_{ij}\}, \{f_i\}\), to explore the effect of conformational and local interactions altering the value of \( \{E_{ij}\} \).

4.1.3 Equilibrium Computation

Let \( \alpha(t) \) denote the site occupancies at the \( t^{th} \) iteration of the algorithm. \( \alpha_j(t) \) specifies the \( j^{th} \) component of this vector, while \( \alpha^{-j}(t) \) contains all but the \( j^{th} \) component. The algorithm to compute a competitive equilibrium for the site specific economy proceeds as follows:

**Algorithm 1:** Site specific equilibrium computation.

**Input:** \( K_{ij}, E_{ij}, \text{and } f_{ti} \)

**Output:** Regulatory state \( (p, s) \)

1. Initialize \( \alpha_j(0) \leftarrow 0 \), for all \( j = 1, \ldots, m \)

2. Iterate over \( t \)

3. for Each \( j \) do

4. Update \( \alpha_j(t+1) \) so that \( \alpha_j(t+1) = G_j(\alpha_j(t+1), \alpha^{-j}(t)) \) from (3.11)

5. Stop when \( \alpha_j(t+1) \approx \alpha_j(t) \) for all \( j = 1, \ldots, m \).
Figure 4-1: Illustration of how the algorithm exploits the structure of the economy. Note the bipartite structure; e.g., strategies for each site only depend on the strategies of proteins. Shaded circles emphasize dependencies between site and protein agents (cf. equations 3.9 and 3.10). The information required to update some $\alpha_j$ concerns proteins that transport to that site directly, and other sites where these proteins transport to indirectly, all of which appear shaded.

Note that the inner loop of the algorithm, i.e., finding $\alpha_j(t + 1)$ on the basis of $\alpha^{-j}(t)$ reduces to a simple binary search. Figure 4.1.3 illustrates the update procedure of the algorithm and its relation to the structure of the economy.

The algorithm generates a monotonically increasing sequence of $\alpha$'s that converge to a fixed point (equilibrium) solution. We provide here a formal convergence analysis of the algorithm that may also be seen as a constructive proof of the existence of a competitive equilibrium. The proof relies on the monotonicity properties of the function $G$ to show first that the inner loop has a unique fixed point, and then show convergence of the resulting sequence of $\alpha_j(t)$. The following lemmas introduce the key monotonicity properties.

**Lemma 2.** Let $\alpha^{-j}$ denote all components $\alpha_j$ except $\alpha_j$, and $\alpha_j^{(1)} < \alpha_j^{(2)} \forall j$. Then, for each $j$, $G^j(\alpha_j^{(1)}, \alpha^{-j}) > G^j(\alpha_j^{(2)}, \alpha^{-j})$. That is $G^j(\alpha_j, \alpha^{-j})$ is a strictly decreasing function of $\alpha_j$ for any fixed $\alpha^{-j}$.

The proof is immediate since the derivative of $G^j(\alpha_j, \alpha^{-j})$ with respect to $\alpha_j$ is strictly negative. This lemma is sufficient to show that the inner loop converges to a unique fixed point. $G$ is a strictly decreasing function of $\alpha_j$ with $G(1, \alpha^{-j}) = 0$, $G(0, \alpha^{-j}) \geq 1$, and $\alpha_j$ is defined in the range $[0, 1]$. It follows that $G$ intersects once and only once the line ($\alpha_j = \alpha_j$), and the recurrence relation (3.11) has a unique fixed point for a fixed $\alpha^{-j}$. 

78
Lemma 3. Let $\alpha_1$ and $\alpha_2$ be two possible assignments to $\alpha$. If for all $k \neq j$, $\alpha_k^{(1)} \leq \alpha_k^{(2)}$, then $G^j(\alpha_j, \alpha_1^{-j}) \leq G^j(\alpha_j, \alpha_2^{-j})$ for all $\alpha_j$.

The proof is essentially based on the fact that $\alpha_1^{-j}$ and $\alpha_2^{-j}$ appear only in the normalization terms for the protein allocations and for these terms

$$\sum_{k \neq j} \exp(E_{ik}(1 - \alpha_k^1)/\beta) \geq \sum_{k \neq j} \exp(E_{ik}(1 - \alpha_k^2)/\beta)$$

as $\alpha_k^{(1)} \leq \alpha_k^{(2)}$ for all $k \neq j$.

On the basis of these lemmas, we can show that the algorithm indeed generates a monotonically increasing sequence of $\alpha$'s.

Theorem 4. $\alpha_j(t+1) \geq \alpha_j(t)$ for all $j$ and $t$.

Proof. By induction. Since $\alpha_j(0) = 0$ and the range of $G^j(\alpha_j, \alpha^{-j}(0))$ lies in $[0, 1]$, clearly $\alpha_j(1) \geq \alpha_j(0)$ for all $j$. Assume then that $\alpha_j(t) \geq \alpha_j(t-1)$ for all $j$. We extend the induction step by contradiction. Suppose $\alpha_j(t+1) < \alpha_j(t)$ for some $j$. Then

$$\alpha_j(t+1) < \alpha_j(t) = G^j(\alpha_j(t), \alpha^{-j}(t-1)) \leq G^j(\alpha_j(t), \alpha^{-j}(t)) < G^j(\alpha_j(t+1), \alpha^{-j}(t)) = \alpha_j(t+1)$$

which is a contradiction. The second line follows from the induction hypothesis and lemma 3, and the third line derives from lemma 2 and the assumption $\alpha_j(t+1) < \alpha_j(t)$. Figure 4-2 illustrates this result graphically. □

Since $\alpha_j(t)$ lies in the interval $[0, 1]$, and because of the continuity of $G^j(\alpha_j, \alpha^{-j})$ in the two arguments, the algorithm is guaranteed to converge to a fixed point solution.

Theorem 5. The algorithm converges to a fixed point $\alpha^*$ such that $\alpha_j^* = G^j(\alpha_j^*, \alpha^{-j})$ for all $j$. 79
Figure 4-2: Graphical illustration of the induction step in the proof of Theorem 4.

**Proof.** The result is a direct consequence of the Monotone Convergence Theorem for sequences and the continuity of $G^j$'s. \qed

## 4.2 Promoter Specific Economy

### 4.2.1 Equilibrium Existence

Finding a competitive equilibrium for the promoter specific DNA-Binding economy is equivalent to solving the following coupled maximization:

$$
\begin{align*}
\max_{p_i} u_i(p_i) & \quad \forall \text{ protein } i \\
\text{subject to} & \\
\sum_r p_{i,r} &= 1 \\
\sum_r p_{i,r} E_i r (a^r) + \beta H(p_i) \\
\max_{s_j} v_j(s_j) & \quad \forall \text{ site } j \\
\text{subject to} & \\
s_{j,i} & \leq \left( p_{i,r} (a^r) f_i - \sum_{k \in N(r)} s_{k,i} \right) \left( 1 - \alpha_j \right) K_{ij} \\
s_j^2 & \leq p_{i,r} f_i \\
\sum_i s_j^2 & \leq 1 \\
v_j(s_j) &= \sum_i s_{j,i} - K_{ij} \left( p_{i,r} (a^r) f_i - \sum_{k \in N(r)} s_{k,i} \right) \left( 1 - \alpha_j \right)
\end{align*}
$$

The condition that there might be an excess of reactants but not products, that is, the first condition in the strategy set of sites is non-convex. We cannot use previous
work Fudenberg and Tirole (1991); Arrow and Debreu (1954) to assert the existence of a competitive equilibrium. However, as in the site specific economy, we show that an equilibrium exists for any setting of the parameters of the economy using the properties of the recurrence relation 3.15,

$$\alpha_j = \sum_{i \in P} \frac{K_{ij}(1 - \alpha_j)}{1 + \sum_{j \in N(r)} K_{ij}(1 - \alpha_j)} p_{i,r}(a) f_i$$

$$\triangleq G^j_i(\alpha, a^{-r})$$

**Theorem 6.** Every promoter based DNA binding economy has an equilibrium.

The proof is conceptually similar to the constructive proof of existence of an equilibrium for the site specific economy. It relies on monotonicity properties of $G^j_i(\alpha, a^{-r})$. The following lemma and theorem provide the basis for proving the existence of an equilibrium and solving the fixed point equations.

The following lemma introduces two important monotonicity properties pertaining the allocation of proteins to promoters.

**Lemma 7.** Let $a^{-r} = \{a^l\}_{l \neq r}$. Then

1. $p_{i,r}(a)$ is a monotonically decreasing function of $a^r$ for any $a^{-r}$.

2. $p_{i,r}(a^r, a^{-r}) \leq p_{i,r}(a^r, a^{-r})$ for all $a^r$ whenever $a_1^l \leq a_2^l$ for all $l \neq r$.

**Proof.** write

$$p_{i,r}(a^r, a^{-r}) = \frac{\exp \left( E_{ir}(a^r) \right)}{\exp \left( E_{ir}(a^r) \right) + \sum_{l \neq r} \exp \left( E_{il}(a^l) \right)}.$$

where we have defined $E_{ir}(a^r) \triangleq E_{ir}(a^r)/\beta$, and note that $x/(x+y)$ is an increasing function of $x$ and by hypothesis $E_{ir}$ is a decreasing function of $a^r$ (part 1) and

$$\sum_{l \neq r} \exp \left( E_{il}(a^l) \right) \geq \sum_{l \neq r} \exp \left( E_{il}(a^l) \right)$$
for part 2, since again $E_{ir}$ is a decreasing function of $a^l$. □

On the basis of this lemma, the following theorem shows that the best response of each promoter to the allocation of other promoters is an increasing function of the allocation of other promoters. That is, the solution of the fixed point $\alpha_j = G^j_r(\alpha, a^{-r})$ yields a sequence of allocations $a^r$ at promoter $r$ that increases monotonically with $a^{-r}$.

**Theorem 8.** Let $\{\hat{\alpha}^{(1)}_j\}$ be the fixed point solution $\alpha_j^{(1)} = G^j_r(\alpha_1, a_1^{-r})$ and $\{\hat{\alpha}^{(2)}_j\}$ the solution to $\alpha_j^{(2)} = G^j_r(\alpha_2, a_2^{-r})$. If $a^l_1 \leq a^l_2$ for all $l \neq r$ then $\hat{a}^r_1 \leq \hat{a}^r_2$.

The proof of theorem 8 can be found in Appendix A.2.

The existence of a competitive equilibrium (i.e., theorem 6) follows from lemma 7 and theorem 8. Like for the site-specific economy, equilibrium existence is a consequence of the monotonicity properties of $G^j_r$, its continuity, and the fact that $\alpha_j \in [0, 1]$.

### 4.2.2 Specification and Use of the economy

The specification of promoter specific DNA binding economy is analogous to the specification of the site specific DNA binding economy. The promoter specific economy is defined by the set of parameters $\{E_{ir}\}$, $\{K_{ij}\}$ and $\{f_i\}$, along with the utility functions $\{u_i\}$ and $\{v_j\}$ and the allocation constraints $\{P^i\}$ and $\{S^j\}$.

Given each $\{E_{ir}\}$, $\{K_{ij}\}$ and $\{f_i\}$, there is a competitive equilibrium $(p_{ir}, s_{ij})$ specified by an average allocation of proteins to promoters and an average allocation of bindings within each promoter. We interpret each such allocation as a regulatory state. Compared to the site-specific economy, however, the predicted regulatory state results from a coarser representation of the transport mechanism. The coarser structure emphasizes interdependence of DNA regions (promoters) that group together sites. At this scale, transport mechanism explains displacements in the chemical equilibria of a number of (possibly distant) promoters as a consequence of the reallocation of protein.
With an algorithm to compute a competitive equilibrium, we may use the economy in a variety of settings:

• with fixed \( \{E_{ir}\}, \{K_{ij}\} \), to track the evolution of a regulatory state for varying availability of resources (i.e. changes in \( \{f_i\} \)).

• with fixed \( \{E_{ir}\}, \{f_i\} \), to reproduce site knock-out experiments, altering the value of \( \{K_{ij}\} \).

• with fixed \( \{K_{ij}\}, \{f_i\} \), to explore the effect of conformational changes in DNA and the effect of these on the reallocation of proteins between promoters, altering the value of \( \{E_{ir}\} \).

The promoter structure mirrors the structure of sequence information in DNA. For most species, however, promoter regions are only partially identified, therefore, limiting our ability to test this economy. We believe, however, that the abstraction is useful in many contexts, including detecting commonalities in the regulatory states of tissues. This line of work will be discussed further in Chapter 8.

4.2.3 Equilibrium Computation

We redefine the notation slightly to illustrate an iterative algorithm for finding the solution \( \alpha_j = G^j_i(\alpha, a^{-r}) \). Let

\[
G^j_i(\alpha_j, \alpha^{-j}, \alpha^{-j}, a^{-r}) = \sum_{i \in P} \frac{K_{ij}(1 - \alpha_j)}{1 + K_{ij}(1 - \alpha_j) + \sum_{k \neq j} K_{ik}(1 - \alpha_k)} f_ip_{ir}(\alpha_j, \alpha^{-j}, a^{-r}). \tag{4.1}
\]

The first argument refers to \( \alpha_j \) anywhere on the right hand side, the second argument refers to \( \alpha^{-j} \) in the denominator of the first expression in the sum, and the third
argument refers to $\alpha^{-j}$ in $p_{i,r}(\cdot)$. The algorithm is as follows:

**Algorithm 2: Promoter Specific equilibrium computation.**

**Input:** $K_{ij}, E_{ir}$, and $f_i$

**Output:** Regulatory state $(p, s)$

1. Initialize $\overline{\alpha}_j(0) \leftarrow 1$ for all $j = 1, \ldots, m$
2. Initialize $\underline{\alpha}_j(0) \leftarrow 0$ for all $j = 1, \ldots, m$
3. Iterate over $t$
4. for Each promoter $r$, update $\overline{\alpha}_j(t)$ and $\underline{\alpha}_j(t)$ do
5. for Each $j$ in promoter $r$ do
6. find $\overline{\alpha}_j$ s.t.
7. $\overline{\alpha}_j(t) = G^j_{j,r}(\overline{\alpha}_j(t), \overline{\alpha}^{-j}(t-1), \alpha^{-j}(t-1), a^{-r}(t-1))$
8. for Each $j$ in promoter $r$ do
9. find $\underline{\alpha}_j$ s.t.
10. $\underline{\alpha}_j(t) = G^j_{j,r}(\underline{\alpha}_j(t), \overline{\alpha}^{-j}(t-1), \alpha^{-j}(t), a^{-r})$
11. Stop when $\overline{\alpha}_j(t) - \underline{\alpha}_j(t) \leq \epsilon$ for all or when $\Delta_t \overline{\alpha} < \epsilon$ and $\Delta_t \underline{\alpha} < \epsilon$

Note that each step in the inner loop of the algorithm, i.e., finding upper and lower bounds for each promoter, reduces to a simple binary search. The algorithm successively narrows down the gap between upper and lower bounds. Specifically, $\overline{\alpha}_j(t+1) \leq \overline{\alpha}_j(t)$ and $\underline{\alpha}_j(t+1) \geq \underline{\alpha}_j(t)$. The fact that these indeed remain upper and lower bounds follows directly from lemma 7 and theorem 8; $G^j_{j,r}(\cdot, \alpha^{-j}, \overline{\alpha}_j, a^{-r})$, viewed as a function of the first argument, increases uniformly as we increase the components of the second argument. Similarly, it uniformly decreases as a function of the third argument.

### 4.3 Nucleus Based Economy

#### 4.3.1 Equilibrium Existence

Finding a competitive equilibrium for the nucleus-based DNA-Binding economy is equivalent to solving the following maximization:
\[
\max_{s_j} \; v_j(s_j) \; \forall \; \text{site } j
\]

subject to
\[
\begin{align*}
    s_{j,c} & \leq K_{cj} \prod_{i \in e} (f_i - \sum_{j} \sum_{c' \in C_j} s_{j,c'}) (1 - \sum_{c' \in C_j} s_{j,c'}) \\
    \sum_{c \in C_j} s_{j,c} & \leq f_i \\
    \sum_{c \in C_j} s_{j,c} & \leq 1 \\
    v_j(s_j) & = \sum_{c \in C_j} s_{j,c} \\
    -K_{cj} \prod_{i \in e} (f_i - \sum_{j} \sum_{c' \in C_j} s_{j,c'}) (1 - \sum_{c' \in C_j} s_{j,c'}) & \\
\end{align*}
\]

Like in the previous two economies, the first condition of the strategy set is not convex. We have to show that an equilibrium exists. Intuitively, we may expect an equilibrium to exist by analogy with chemistry. In the absence of a transport mechanism, this maximization corresponds to finding the chemical equilibria of a set of joint reactions. Generally the concentrations at chemical equilibrium of a set of joint reactions are obtained by minimizing the global Gibbs energy of the system of chemical reactions (Atkins, 1986; R. H. Perry, 1997). The economic approach differs conceptually in that it places the emphasis locally on each complex-site pair.

**Theorem 9.** *Every nucleus based DNA binding economy has a unique equilibrium.*

The theorem can be proven in two ways. Mapping the recurrence relation defined in equations (3.24, 3.25),

\[
\beta_j = \frac{1}{1 + \sum_{c \in C_j} K_{cj} \prod_{i \in e} (R_i)} = B_j(R) \\
R_i = f_i \left( 1 + \sum_{j,c \in C_j; i \in e} K_{cj} \beta_j \left[ \prod_{l \in C \neq i} (R_l) \right] \right)^{-1} = G_i(\beta, R),
\]

to the minima of a global objective function. Or constructively with an algorithm that uses the monotonicity properties of the fixed point. The constructive proof however does not assert uniqueness.
Here we show that an equilibrium exists and is unique with a mapping to an objective function. In Section 4.3.3 we introduce an algorithm to compute the competitive equilibrium that uses the monotonicity properties of the fixed point, and emphasizes the similarity with the previous two economies.

**Objective Function and Uniqueness of Equilibrium**

We define a new objective function that will allow us to show that any fixed point \( R_i = G_i(B(R), R) \) is unique over the domain \( R_i \in (0, f_i] \)

\[
F(R) = - \sum_i f_i \log R_i + \sum_i R_i + \sum_j \log \left( 1 + \sum_{c \in C_j} K_{cj} \prod_{\ell \in c} R\ell \right) \quad (4.2)
\]

This objective function recovers the fixed point at its minima,

\[
\frac{dF(R)}{dR_i} = - \frac{f_i}{R_i} + 1 \sum_j \left( \frac{1}{1 + \sum_{c \in C_j} K_{cj} \prod_{\ell \in c} R\ell} \right) \sum_{c \in C_j : i \in c} K_{cj} \prod_{\ell \in c \setminus i} R\ell
\]

\[
= - \frac{f_i}{R_i} + 1 \sum_j \sum_{c \in C_j : i \in c} B_j(R)K_{cj} \prod_{\ell \in c \setminus i} R\ell
\]

\[
= - \frac{f_i}{R_i} + \frac{f_i}{G_i(B(R), R)} = 0. \quad (4.3)
\]

This shows that the extrema of the objective function \( F \) are in correspondence with the fixed points that define the competitive equilibria of our economy. We now show that the objective function has a unique minimum. Let \( R_i = \exp(\xi_i) \) so that,

\[
F(\xi) = - \sum_i f_i \xi_i + \sum_i e^{\xi_i} + \sum_j \log \left( 1 + \sum_{c \in C_j} K_{cj} e^{\sum_{\ell \in c} \xi\ell} \right). \quad (4.4)
\]

\( F(\xi) \) is a sum of convex functions of \( \xi \), hence convex. Additionally, for strictly positive \( f_i \), \( F(\xi) \) tends to \( +\infty \) when \( \xi_i \) approaches either \( \pm \infty \), and the function has a unique
minimum. Application of the chain rule for differentiation,

\[
\frac{dF(\xi)}{d\xi_i} = \frac{dF(\xi(R))}{dR_i} \frac{dR_i(\xi_i)}{d\xi_i}
= \frac{dF(R)}{dR_i} R_i = -f_i + \frac{f_i R_i}{G_i(B(R), R)} = 0,
\]

shows that the minimum of \( F(\xi) \) is zero if and only if \( R_i = G_i(B(R), R) \). That is, the unique minimum is defined by the fixed point relation. Hence the fixed point relation is unique.

4.3.2 Specification and Use of the economy

The *Nucleus-based DNA binding economy* is defined by the set of parameters \( \{K_{ij}\} \) and \( \{f_i\} \), along with the utility function \( \{v_j\} \) and the allocation constraints \( \{S^j\} \).

Given values for each \( \{K_{ij}\} \) and \( \{f_i\} \), there is a unique competitive equilibrium specified by an average allocation of proteins and complexes along DNA (s). Unlike the site and promoter based economies, unbound protein is assumed to be shared across sites. We interpret each such allocation as a regulatory state. The predicted regulatory state of the nucleus specific economy summarizes the competing and coordinating relations between proteins.

With an algorithm to compute a competitive equilibrium, we may use the economy in a variety of settings:

- with fixed \( \{K_{ij}\} \), to track the evolution of the regulatory state and changes in complex formation for varying availability of proteins (i.e. changes in \( \{f_i\} \).)

- with fixed \( \{f_i\} \), to reproduce site knock-out experiments by altering \( \{K_{ij}\} \) values.

Chapter 5 shows how to learn the structure of the economy from protein binding observations. We can then use the economy to identify likely complex candidates (we discuss all the potential interpretations of complexes there.)
4.3.3 Equilibrium Computation

We introduce an iterative algorithm to compute the equilibrium of the nucleus specific economy, the algorithm uses monotonicity properties of the fixed point relation

\[ \beta_j = B_j(R) \quad (4.6) \]
\[ R_i = G_i(\beta, R), \quad (4.7) \]

and provides a constructive proof of existence of an equilibrium of the nucleus specific economy. The equilibrium is also unique. This is not guaranteed by the algorithm but by the analysis from Section 4.3.1.

The algorithm finds the fixed point from above and below \((R^*_i \in [R^{(0)}_i, \overline{R}^{(0)}_i])\) iteratively. Every iteration is guaranteed to produce tighter bounds on the fixed point \([R^{(t)}_i, \overline{R}^{(t)}_i] \subset [R^{(t-1)}_i, \overline{R}^{(t-1)}_i]\).

**Algorithm 3**: Nucleus specific equilibrium computation.

**Input**: \(K_{ij}\) and \(f_i\)

**Output**: Regulatory state \((s)\)

1. Initialize \(\overline{R}^{(0)}_i := f_i\)
2. Initialize \(R^{(0)}_i := 0\)
3. Iterate over \(t\)
   
   4. for Each site \(j\) update bounds on each \(\beta^{(t)}_j\): do
      
      5. \(\beta^{(t)}_j := B_j(\overline{R}^{(t)})\)
      6. \(\overline{\beta}^{(t)}_j := B_j(R^{(t)})\)
   
   7. for Each protein \(i\) update \(t+1\) bounds on each \(R^{(t+1)}_i\) do
      
      8. \(\overline{R}^{(t+1)}_i := G_i(\beta^{(t)}, R^{(t)})\)
      9. \(\overline{R}^{(t+1)}_i := G_i(\overline{\beta}^{(t)}, \overline{R}^{(t)})\)
   
   10. Stop when \(\overline{R}^{(t)}_i - \overline{R}^{(t)}_i < \epsilon\)

We show that each iteration results in tighter bounds on the fixed point, that is
\[ [R_i^{(t)}, \overline{R}_i^{(t)}] \subset [R_i^{(t-1)}, \overline{R}_i^{(t-1)}]. \] Consider the mappings:

\[
\begin{align*}
\overline{R}_i^{(t+1)} &= G_i(B(\overline{R}_i^{(t)}), R_i^{(t)}) \\
\underline{R}_i^{(t+1)} &= G_i(B(R_i^{(t)}), \overline{R}_i^{(t)})
\end{align*}
\] (4.8)

We will first show monotonicity of each bound by induction. By simple inspection of (3.24, 3.25) we see that induction holds for the first iteration:

\[
\begin{align*}
\overline{R}_i^{(1)} &= G_i(B(\overline{R}_i^{(0)}), R_i^{(0)}) \leq f_i = \overline{R}_i^{(0)} \\
\underline{R}_i^{(1)} &= G_i(B(R_i^{(0)}), \overline{R}_i^{(0)}) \geq 0 = R_i^{(0)}
\end{align*}
\] (4.9)

For the induction step we assume that

\[
\begin{align*}
\overline{R}_i^{(t)} &\leq \overline{R}_i^{(t-1)} \\
\underline{R}_i^{(t)} &\geq \underline{R}_i^{(t-1)}
\end{align*}
\] (4.10)

and examine the value of \( \overline{R}_i^{(t+1)} \)

\[
\begin{align*}
\overline{R}_i^{(t+1)} &= G_i(B(\overline{R}_i^{(t)}), R_i^{(t)}) \\
&\leq G_i(B(\overline{R}_i^{(t-1)}), R_i^{(t)}) \\
&\leq G_i(B(\overline{R}_i^{(t-1)}), \underline{R}_i^{(t-1)}) \\
&= \overline{R}_i^{(t)}
\end{align*}
\] (4.11)

The second and third line follows from the fact that \( G \) decreases with \( B \) as a function of both arguments and the assumptions. Symmetric reasoning shows monotonicity for the lower bound. Monotonicity and \( R_i \in [0, f_i] \) guarantees convergence to a fixed point.
4.4 Strategic complementarity

Here we make the connection of the DNA binding economies with strategic complementarity and the theory of super-modular games (Milgrom and Roberts, 1990; Fudenberg and Tirole, 1991; Bulow et al., 1985) and illustrate additional properties that result from super-modularity.

The term strategic complementarity describes a situation in which an agent’s best response to an increase in the strategy of its opponents’ is an increase of its own strategy. The fixed point relations introduced for each DNA binding economy can be interpreted as best response functions of one player to its opponents’ strategy. For instance, the fixed point relation for the site specific economy $\alpha_j = G(\alpha_j, \alpha^{-j})$; there, theorem 4-2, illustrated in Figure 4 showed that the best response of each player is a monotonically increasing function of the strategies of other players.

To verify strategic complementarity, strategies have to be comparable. In all generality this requires defining strategies on a lattice and a partial order. In the DNA binding economies, strategies are naturally ordered because they are indexed by a real number ($\alpha_j$ or $R_i$), and bounded, hence they trivially define a complete lattice.

The class of super-modular games is the class of games in which each players’ strategy set is ordered and the game exhibits strategic complementarity. A super-modular game with a compact strategy set is said to be smooth super-modular if it has the following properties,

I The strategy set $S_i$ of each player is compact.

II $u_i(s_i, s_{-i})$ (the utility function of player $i$) is upper semi-continuous in $s_i$ (cf. Andreu Mas-Colell (1995)) and continuous in $s_{-i}$

III $u_i$ has increasing differences

The DNA binding economies are smooth supermodular games. We use the site specific economy to illustrate that the monotonicity properties we have derived for

$$1u(s_i(t+1), s_{-i}(t+1)) - u(s_i(t), s_{-i}(t)) \geq u(s_i(t), s_{-i}(t+1)) - u(s_i(t), s_{-i}(t))$$

$$s_i(t), s_{-i}(t+1) \geq s_{-i}(t).$$
all the DNA binding economies are connected with the notion of super-modularity. We consider the fixed point relation (3.11). While we have not introduced a utility function in terms of $\alpha$, we could easily construct one subtracting both sides of (3.11). Verification of the first two properties for the site specific economy is trivial; $\alpha \in [0,1]$ is compact and $G(\alpha_j, \alpha^{-j})$ is a continuous function in both its arguments. The increasing differences property can be derived from lemmas 2 and 3, and theorem 4.

It follows that the site specific economy is a super-modular game and the algorithm to compute an equilibrium is an instance of iterated dominance, particularly, the upper bound calculation. Computationally, iterative dominance is reminiscent of constraint propagation. At each iteration, we update a bound that constrains the strategy space of that player (node) for future iterations based on the updated bounds of the other players (Ortiz). If we seed the algorithm with $\alpha_j \leftarrow 0$ and $\alpha_j \leftarrow 1$, super-modularity guarantees that the resulting updates will successively bind all possible competitive equilibria. If the bounds meet, there is a unique equilibrium and the competitive equilibrium is said to be dominance solvable.
Chapter 5

Learning the Economic Structure

This chapter introduces algorithms to learn the parameters of DNA binding economies from binding observations. The parameters are $K$, $E$ that define the interactions between agents, as well as $f$. We interpret binding observations as partial noisy observations of equilibria. Binding observations indicate the binding arrangement of regulators along DNA, that is, the joint strategic allocation that results when each agent has maximized its utility function conditioned on the strategies of its opponents. In this case the utility functions $\{u_i\}, \{v_i\}$ are unknown. The learning task is to find the utility functions that best fit the noisy observations about the joint strategic allocation. Figure 5-1, modified from Chapter 1, sketches our approach to learning.

![DNA Proteins](image)

Figure 5-1: In learning we try to solve the inverse problem of determining the economic structure that gave rise to the observed equilibrium.

The parameters $\{K\}, \{E\}$ define the dependence of the utility function of each agent on the strategies of other agents. They can be seen to define a relational graph between agents as in the formalism of graphical games and economies (e.g. Kakade et al. (2004); Kearns et al. (2001)). Owing to that formalism we say that our learning
algorithms learn the structure of the economy. (This is consistent with the meaning of structure in the work of Kearns and Ortiz (2003)). Learning economic structures has broad applicability. Specifically it is well suited to problems that are best explained in terms of the preferences of individuals, and where only limited slices of observations about the aggregate behavior are available. A key advantage is that there is no need to introduce a function to describe the global behavior. Instead, the form of the utility functions and the choice of an equilibrium concept explicitly define the assumptions about the emerging aggregate behavior.

The chapter is divided into two parts. First, we give an overview of previous work at the interface of learning and games. Generally speaking, previous work has been from the point of view of the agents; it focuses on the process by which agents learn how to play or how to set the rules of the game given partial observations of the actions of their opponents. This approach has generated a rich debate around the adequacy of Nash Equilibrium to describe the final outcome of a game and the concept of rationality. To the best of our knowledge, our approach to learning economic theories is new to machine learning. It’s closest analog is in the use of approximation within the theory of revealed preferences (Varian, 2006), a subfield of economics and econometrics.

In the second part we introduce algorithms to learn the structure of the site and nucleus specific economies from high-throughput experimental data, specifically microarray experiments. We emphasize that our approach deviates from previous work at the interface of machine learning and games in that we identify observations with the equilibrium, and try to infer the underlying structure of the economy. Therefore, our approach is not concerned with rationality of the agents and the adequacy of the concept of equilibrium. These are taken as given and justified in terms of the physics of the problem.
5.1 Previous Work on the Theory of Learning Applied to Game Theory and Economics

Four main research directions lie at the intersection of learning and game theory / economic theory. *Learning in Games* (including evolutionary game theory), *Mechanism Design*, *Reinforcement Learning/Multi-Agent Systems*, and *Revealed preferences in econometrics*. The first three are generally concerned with the procedure by which agents/players acquire knowledge about the game and adapt. We review the three approaches together and refer to them jointly as *player learning*.

For further reference, (Fudenberg and Levine, 1998) exhaustively overview learning in games from a game theoretic vantage point, (Nicolò Cesa-Bianchi, 2006) is a review from the computational learning theory perspective. (Andreu Mas-Colell, 1995) is a microeconomics reference work and devotes various sections to learning in economics. (Varian, 2006) overviews the theory of revealed preference.

5.1.1 Player Learning

Learning in Games

Learning in games (Fudenberg and Levine, 1998), is generally concerned with the dynamics of play. The field is closely connected to rationality (Simon, 1957), the applicability of Nash equilibrium to real life situations (Binmore, 2007), and the scope of applicability of Bayesian analysis in game theory (Binmore, 1993).

Learning in games generally assumes that the utility function and the strategy set of a player’s opponents are unknown. Players only have access to partial information of the actions of their opponents to make a decision on their next move. In such a setting, there is no obvious way for a player to use the concept of Nash equilibrium. Nash Equilibrium is a description of the game *a posteriori* that requires full information about the game. Learning in games focuses on the dynamics of play to explain how and when a game can attain equilibrium, and under which learning conditions it is reasonable to expect such equilibrium to be a Nash equilibrium and what it means.
to play rationally.

The two foundational methods of learning in games are fictitious play and evolutionary game theory. Fictitious play (Robinson, 1951; Brown, 1951) is a general method to update one player’s belief on the choice of strategies of its opponents after successful rounds of playing. The opponents strategy set is assumed fixed and the utility function of each player is conditional on the updated beliefs. The method to assess the best response can be interpreted as Bayesian inference (Fudenberg and Levine, 1998), and the idea of updating beliefs is closely connected with reinforcement learning. The study of fictitious play addresses the problem of whether successful rounds of the game will converge and if so to which form of equilibrium. It provides a general framework to introduce learning how to play with varying choices of update rule and assumptions about the role of mixed strategies (cf. stochastic fictitious play).

Another foundational method in learning in games is evolutionary game theory (Maynard Smith, 1982; Lewontin, 1961). In contrast with fictitious play evolutionary game theory studies the evolution of strategies in populations rather than single players, and consequently puts less emphasis on attaining an equilibrium. It was originally developed in the context of evolution in biology; since then it has been reabsorbed as part of learning in games where it is often referred to as replicator dynamics (Fudenberg and Levine, 1998).

Mechanism Design

A mechanism is defined as a set of rules that governs how a game is played. The goal of mechanism design is to select an optimal set of rules given partial observations of the values that players assign to an strategy. The precise meaning of optimality of a mechanism depends on the context (v. Parkes (2004) for a review, and Balcan and Blum (2005) for a description from the machine learning standpoint).

In the context of auctions, the auctioneer is the principal agent. The auctioneer seeks to maximize its own utility function choosing an optimal set of rules to govern the auction (mechanism), given incomplete knowledge about the utility functions of the rest of the agents. More general formulations of the problem (cf. Andreu
Mas-Colell (1995) consider a set of agents that collectively create a mechanism that implements a social choice function based on incomplete information about the players preferences.

Multi-Agent Learning, Reinforcement Learning, and Stochastic games

Littman (1994) extended the Markov decision process framework used in reinforcement learning to consider multiple agents individually as in Markov games (van der Wal, 1981) rather than lumped in a static environment. The field of multi-agent learning has since grown at the interface of game theory (Markov games) and computer science (reinforcement learning). In a recent special issue of Artificial Intelligence (cf. Vohra and Wellman (2007)), Shoham et al. (2007) overview the state of the art in multi-agent learning from the economic and computer science vantage points. In the same special issue, (Fudenberg and Levine, 2007; Tuyls and Parsons, 2007; Stone, 2007) complement the overview with the specifics from economic theory, evolutionary game theory and computer science. Shoham et al. (2007) define the general setting of multi-agent learning as a stochastic game. A stochastic game is a multi-stage game in which a stochastic transition function determines the probability of the next game to be played based on the outcome of the last game played.

Shoham et al. (2007) restrict their attention to known fully observable games, and motivate two learning approaches; model-based and model-free learning. In model-based learning each agent learns the strategies of its opponents to devise a best response. In model-free learning the agent learns a strategy that does well against the opponents, without explicitly learning the opponents strategy. When games are unknown Q-learning (Watkins and Dayan, 1992; Watkins, 1989) provides a general framework to learn policies based on the introduction of a probabilistic reward and a discount rule. The original Q-learning was developed in the context of Markov decision processes in reinforcement learning. In its extension to the setting of multi-agent games, each agent ignores the other agents and assumes a “stationary environment”. (Littman, 1994) suggested ways to adapt Q-learning to a multi-agent setting in the case of zero-sum two-players games.
5.1.2 Revealed Preferences

Revealed preference is the subfield of economic theory concerned with the inverse problem of assessing consumer preferences from their actions (v. Varian (2006) for an overview).

One example is the integrability problem (v. Andreu Mas-Colell (1995, page 75 of)): find utility functions that rationalize a given demand function. Where rationalize is understood as in the definition of Nash equilibrium; the choice that maximizes the utility function. The integrability problem is mostly concerned with existence of utility functions and their determination when the demand function is given. There is a body of research that discusses the theoretical implications of revealed preference and its ability as a tool to assess rationality.

(Afriat, 1967) gives an explicit method to approximate the inverse mapping and construct a utility function that is consistent with observations of demand. (Dievert, 1973; Varian, 1982) present Afriat’s approach to revealed preference as a constrained optimization problem. (Beigman and Vohra, 2006) study the learnability of classes of demand functions.

Our approach to learn the structure of the economy is related to the problem of revealed preferences in that we associate the observations with strategies at equilibrium. Revealed preferences, however, focuses in asserting consistency of a set of utility functions or recovering a set of utility functions that supports observations. Our approach deviates from Afriat’s in that we assume that observations are partial and noisy, and focus in finding the utility function from the space defined by the physics of the problem that best explains the observations.
5.2 Site-Specific DNA Binding Economy

5.2.1 Set Up

We assume that we are given noisy observations about binding in regions of DNA. These could be observations of binding ratios generated via Chip-Chip Microarray assays (Ren et al., 2000). Specifically, we assume that data is available as a profile of binding ratios $\hat{o}_{tij}$ per experiment $t$ for each protein-site pair $(i, j)$. We index experiments by the triple $(t, i, j) \in T$. The word experiment refers to a set of experimental conditions. Depending on the experiment design each experiment may be the result of multiple Chip-Chip Microarray assays, for instance, when one Microarray assay is required to analyze the binding profile of each regulator under the same experimental conditions of species concentration, and temperature.

Our goal is to learn the parameters of the economy, $K_{ij}, E_{ij}, f_{ti}$ minimizing the following cost function:

$$F(\hat{o}_{tij}, s_{tij}, K_{ij}, E_{ij}, f_{ti}) = \sum_{(t,i,j) \in T} L(\hat{o}_{tij}, s_{tij}) + \lambda \sum_{ij} K_{ij} + \lambda \sum_{ij} E_{ij},$$

(5.1)

where $L(\hat{o}_{tij}, s_{tij})$ is a loss function that will be discussed further in the experimental section. Subject to two types of constrains. The following three equality constrains are provided by the algorithm to find a competitive equilibrium

$$s_{tij} = \frac{K_{ij}\beta_{tj} f_{ti} P_{tij}}{1 + K_{ij} \beta_{tj} f_{ti} P_{tij}}, \quad P_{tij} = \frac{e^{E_{ij} \beta_{tj}}}{\sum_l e^{E_{il} \beta_{tl}}}, \quad \beta_{tj} = 1 - \sum_i s_{tij},$$

(5.2)

the remaining three inequality constraints,

$$K_{ij} \geq 0, \quad E_{ij} \geq 0, \quad f_{ti} \geq 0,$$

(5.3)

arise from the definition of the variables and are also needed to introduce 1-norm regularization terms ($\lambda \sum_{ij} \cdots$) in the cost function.

Dependence on the experiment is restricted to nuclear concentration $f_{ti}$ and final
allocation \(s_{tij}\). Equilibrium constants \(K_{ij}\) and affinities \(E_{ij}\) are assumed to remain constant across experimental conditions.

The optimization problem defined above may have multiple local minima. The existence of multiple local minima is in part a consequence of the existence of internal degrees of freedom between the parameters that define the economy. For instance, in the absence of additional constraints, it is generally possible to explain increased binding at a site both increasing transport affinity and increasing the chemical equilibrium constant. Similar considerations explain the internal degree of freedom between nuclear concentration and equilibrium constants or affinities. One of the motivations to introduce the promoter-specific economy is to incorporate structure present in DNA that limits these degrees of freedom.

In practice, existence of local minima implies that the optimization problem will converge to different values depending on the initialization. Nonetheless, the optimization may still be able to capture the behavior of the system, and should display the same general trend that experimental values do.

We can solve the optimization problem above via gradient descent. To simplify derivation it is useful to think of the procedure to solve this optimization as tracing a path in the \((s, \beta, K, E, f)\) space. That is, at any given point \((s^0, \beta^0, K^0, E^0, f^0)\) the gradient \((\dot{s}, \dot{\beta}, \dot{K}, \dot{E}, \dot{f})\) defines the direction of maximum growth in the cost function as a function of the parametrization of the optimization path. Dotted variables represent derivatives with respect to the path parameter. Since we are not interested in this parameter we may proceed without naming it.\(^1\)

In appendix B.1, we show how to reduce 5.1 to an unconstrained minimization problem. The path derivative of the cost function that results,

\[
\dot{F} = \nabla F(k, g, e)^T \cdot \begin{bmatrix} \dot{k} \\ \dot{g} \\ \dot{e} \end{bmatrix},
\]

\[\text{Eq. (5.4)}\]

\(^1\)To clarify notation, however, note that dotted variables are a shorthand for \(\frac{d}{d\theta}\), where \(\theta\) would be the parameter that parametrizes the advance in the optimization path, the so called path parameter.
can be expressed in terms of only three variables \((k, g, e)\) defined as,

\[
k_{ij} \triangleq \log K_{ij} \quad e_{ij} \triangleq \log E_{ij} \quad g_{ti} \triangleq \log f_{ti}. \tag{5.5}
\]

With corresponding path derivatives

\[
\dot{k}_{ij} = \frac{\dot{K}_{ij}}{K_{ij}} \quad \dot{e}_{ij} = \frac{\dot{E}_{ij}}{E_{ij}} \quad \dot{g}_{ti} = \frac{\dot{f}_{ti}}{f_{ti}}. \tag{5.6}
\]

The change to log-scale variables is sufficient to ensure non-negativity of the variables provided that the initial point of the optimization path \((s^o, \beta^o, K^o, E^o, f^o)\) satisfies the inequality constraints. The expression for the variation in \(F\) when tracing a path through \((k, g, e)\) space along the direction \((\dot{k}_{ij}, \dot{g}_{ti}, \dot{e}_{ij})\) is given here for completeness,

\[
\dot{F} = \sum_{ij} \left( M_{ij}^k + \sum_{ij'} M_{ij'}^\gamma [D_t^{-1}]_{jj'} N_{ij}^k \right) \dot{k}_{ij} + \sum_{ti} \left( M_{ij}^g + \sum_{ij'} M_{ij'}^\gamma [D_t^{-1}]_{jj'} s_{ti} \right) \dot{g}_{ti} + \sum_{ij} \left( M_{ij}^e + \sum_{ij'} M_{ij'}^\gamma [D_t^{-1}]_{jj'} s_{ti} \left[ \delta_{(i,j)} - P_{ij} \right] E_{ij} \beta_{ij} \right) \dot{e}_{ij}. \tag{5.7}
\]

The variables \(M_{ij}^k, M_{ij}^\gamma, D_t^{-1}, N_{ij}^k, M_{ij}^g, M_{ij}^e, s_{ti}, \text{ and } P_{ij}\) are placeholders for expressions of \(e, k, g\) defined in equation (B.16) from appendix B.

### 5.2.2 Algorithm

We use a gradient descent method. For a given descent direction \((\dot{k}, \dot{e}, \dot{f})\) we obtain the change in the value of the parameter \(K\) with a multiplicative update that results from applying an additive update on \(k\) as follows

\[
k_{ij} = k_{ij} + \gamma \dot{k}_{ij} \quad \log K_{ij} = \log K_{ij} + \gamma \dot{k}_{ij} \quad K_{ij} = K_{ij} e^{+\gamma \dot{k}_{ij}}. \tag{5.8}
\]
Note that $\dot{k}_{ij}$ is already a descent direction, hence the positive sign. For instance, if we followed a path along the gradient then $\dot{k}_{ij} = -\nabla_{k_{ij}} F(k, e, g)$. The procedure to obtain the updates for $R_i$ and $f_{ti}$ is analogous.

The algorithm is a standard gradient descent algorithm.

**Algorithm 4**: Standard gradient descent algorithm adapted to learning the parameters of the site-specific economy. (v. Bertsekas (2004) for details on the Armijo Rule and different alternatives to find a descent direction)

**Input**: Observations $\hat{s}_{tij}$ preprocessed from $\hat{o}_{tij}$, a choice of Loss function, line minimization parameters.

**Output**: $K_{ij}$, $E_{ij}$, and $f_{ti}$ that minimize regularized cost function and predicted allocation at competitive equilibrium $s_{tij}$

1. **Initialize** $K_{ij}$, $E_{ij}$, and $f_{ti}$ to some feasible value
2. **Iterate** Main Loop
   3. Find descent direction $(\dot{k}_{ij}, \dot{e}_{tij}, \dot{g}_{tij})$
   4. Use line minimization with (5.8) to update $K_{ij}$, $E_{ij}$, and $f_{ti}$, and Algorithm 1 to then compute $s_{tij}$
5. **Stop** when cost function converges.

Note that the choice of the initial point will affect the results. Also, depending on the data source, some extra work may be required to compare binding ratios $\hat{o}_{tij}$ with the output of the fixed point Algorithm $s_{tij}$. These and other practical considerations are discussed in the context of our experimental results in Chapter 6.
5.3 Nucleus-Specific DNA Binding Economy

5.3.1 Set Up

We start by recalling definitions and clarifying initial assumptions:

Let $C_j$ be the set of complexes $c$ ( $c \in C_j$) that can bind to site $j$. Each complex $c$ is defined as a set of proteins $i \in P = \{1,...n\}$ that bind jointly. For notational simplicity we use the word complex when referring to a single protein as well. When protein $i$ binds a site alone we consider the “singleton” complex $c = \{i\}$. Note that with these definitions, $C_j$ is a collection of sets, and in the worst case it is the powerset of all proteins. In most cases, though, it will be possible to significantly reduce the cardinality of $C_j$ based on observations or literature (hence $C_j \subset \text{PowerSet}(P)$)

Let $P_i$ be the set of complex-site pairs $(c, j)$ that involve protein $i$, $P_i = \{(c, j) | c \in C_j, i \in c\}$

Let $s_{t,c,j}$ denote the rate of binding of complex $c$ at site $j$ in experiment $t$. These are the predictions of our algorithm at equilibrium.

Ideally we would like to compare the predictions of our algorithm to binding observations $\hat{s}_{tij}$ of complex $c$ at site $j$ in experiment $t$. However, high-throughput data generally available from Chip-Chip microarray intensities (cf. (Ren et al., 2000)) does not distinguish between protein $i$ binding site $j$ alone or jointly with another protein. Our algorithm will determine for each site whether proteins bind alone or as a complex. We then define $\hat{s}_{tij}$ as an observation of the allocation of protein $i$ at site $j$; it includes all complexes in which protein $i$ participates, including “singleton complex” $\{i\}$. $s^*_{tij}$ is the average occupancy of site $j$ by complex $c$ inferred directly from the observations. Finally, the set $\hat{S}$ defines all the complex allocations that are consistent with the observed binding values of protein $i$. The vector of $s^*_{tij}$ enumerating all the complexes across all sites and experiments is one element of the set $\hat{S}$.

Other choices of input data (cf. (Johnson et al., 2007; Qi et al., 2006)) are possible. In practice, the choice of input data will only affect the preprocessing steps to translate raw observations $\hat{o}_{tij}$ into binding observations $\hat{s}_{tij}$.

Table 5.1 refreshes the notation introduced in chapter 3 for the nucleus specific
economy and summarizes the new notation introduced for the learning algorithm.

Notation for the nucleus specific economy

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_j$</td>
<td>set of complexes $c$ that bind at site $j$.</td>
</tr>
<tr>
<td>$c$</td>
<td>complex, defined as a set of proteins {i}.</td>
</tr>
<tr>
<td>$P_i$</td>
<td>set of complex-site pairs involving protein $i$</td>
</tr>
<tr>
<td>$P_i = {(c, j)</td>
<td>c \in C_j, i \in c}$.</td>
</tr>
<tr>
<td>$s_{t, c, j}$</td>
<td>binding rate of complex $c$ at site $j$ in experiment $t$.</td>
</tr>
<tr>
<td>$\alpha_j$</td>
<td>total binding rate at site $j$.</td>
</tr>
<tr>
<td>$\beta_j$</td>
<td>availability of site $j$.</td>
</tr>
<tr>
<td>$R_i$</td>
<td>protein $i$ that remains unbound.</td>
</tr>
<tr>
<td>$v_j(s_j)$</td>
<td>utility function of site agent.</td>
</tr>
<tr>
<td>$f_i$</td>
<td>Total protein $i$ available in the cell/nucleus</td>
</tr>
<tr>
<td>$K_{cj}$</td>
<td>Chemical equilibrium constant for complex $c$ at site $j$</td>
</tr>
<tr>
<td>$c(1) + \cdots + c(m) + j_{SITE} \rightleftharpoons c_{BINDING, c \in C_j}$</td>
<td></td>
</tr>
</tbody>
</table>

New notation introduced for learning

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{s}_{tcj}$</td>
<td>observed rate of binding of complex $c$ at site $j$ in experiment $t$ (generally unavailable).</td>
</tr>
<tr>
<td>$\hat{s}_{tij}$</td>
<td>observed binding rate of protein $i$ at site $j$ in experiment $t$; it includes all complexes in which protein $i$ participates, including “singleton complex” {i}.</td>
</tr>
<tr>
<td>$s^*_{tcj}$</td>
<td>binding rate of complex $c$ at site $j$ in experiment $t$ inferred inferred directly from the observations.</td>
</tr>
<tr>
<td>$\hat{S}$</td>
<td>set of all complex allocations consistent with the observed protein bindings($\hat{s}$). Note that $(s^*) \in \hat{S}$.</td>
</tr>
</tbody>
</table>

Table 5.1: Notation

Our learning algorithm uses the fact that the nucleus-specific economy has a unique equilibrium that maps to the minimum of the objective function (4.2) (cf. Section 4.3). We introduce the objective function (4.2) as a “biological” regularizer, and solve the learning problem and the economy jointly.

The learning cost function to minimize includes a loss term $F_L$, a regularization penalty to control complexity in the absence of observations $F_{L1}$, a penalty term $F_{fR}$ to couple $f_i$ and $R_i$ at equilibrium, and a term $F_{BIO}$ that penalizes parameter choices that deviate significantly from competitive equilibrium (that is the biological regularizer). We list all these terms below including a reference to the experiment $t$
where appropriate.

\[
F_{BIO}(R_{ti}, K_{cj}, f_{ti}) = - \sum_{ti} f_{ti} \log R_{ti} + \sum_{ti} R_{ti} + \sum_{ti} \log \left( 1 + \sum_{c \in C_j} K_{cj} \prod_{t \in c} R_{tt} \right)
\]

\[
F_{fR}(f_{ti}) = + \sum_{ti} f_{ti} \log f_{ti} - (1 - \log \varepsilon) \sum_{ti} f_{ti}
\]

\[
F_{L1}(K_{cj}) = \sum_{j,c \in C_j} K_{cj}
\]

\[
F_{L}(R_{ti}, K_{cj}, f_{ti}) = \sum_{tcj: (t,c,j) \in T} \text{Loss} \left( s_{tcj}, s_{tcj}(R_{ti:i \in c}, K_{cj}) \right)
\]

\[
s.t. \quad s_{tcj}(R_{ti:i \in c}, K_{cj}) = \frac{K_{cj} \prod_{i \in c} R_{ti}}{1 + \sum_{c' \in C_j} K_{c'j} \prod_{i \in c'} R_{ti}}
\]

\[
K_{cj} > 0, f_{ti} > 0, R_{ti} > 0
\]

(5.9)

where in coupling \( f_i \) with the availability of protein \( i \) at equilibrium \( R_i^* \) we have introduced a parameter \( \varepsilon \) that sets the proportion of protein that remains in excess at equilibrium according to the following equation

\[
f_i - R_i^* = (1 - \varepsilon) f_i.
\]

(5.10)

We discuss our choice of loss function \( \text{Loss} \left( s_{tcj}, s_{tcj}(R_{ti:i \in c}, K_{cj}) \right) \) in the algorithm section.

Note that the cost function terms above are not convex in the variables of the game, but are convex in the following log-scale variables,

\[
k_{\alpha\beta} = \log K_{\alpha\beta} \quad dk_{\alpha\beta} = 1/K_{\alpha\beta} dK_{\alpha\beta}
\]

\[
\xi_{ti} = \log R_{ti} \quad d\xi_{ti} = 1/R_{ti} dR_{ti}.
\]

(5.11)

We know from Section 4.3.1 that the biological regularizer is convex in \( \xi \), similar arguments show that it is convex in \( k \). The rest of the terms are either linear or exponential in these variables, hence also convex. The combined cost function \( F(\xi_{ti}, k_{cj}, f_{ti}) \),
shown below, is jointly convex on $k$ and $\xi$.

$$F(\xi_{ti}, k_{cj}, f_{ti}) = F_L + \lambda_1 (F_{BIO}(\xi, k) + F_fR(f, \xi)) + \lambda_2 F_L(k)$$

s.t.

$$f_{ti} > 0$$

$$s_{tej}(\xi_{ti}; i \in c, k_{cj}) = \frac{e^{k_{cj} e^{\sum \xi_{ti}}} e^{\sum \xi_{ti}}}{1 + \sum e^{k_{cj} e^{\sum \xi_{ti}}}}.$$ (5.12)

Note that in addition to ensuring convexity of the cost function, the switch to log-scale incorporates the non-negativity constraints on $K$ and $R$ into the cost function.

5.3.2 Algorithm

Preprocessing

Our preprocessing is based on our interpretation of the comments and warnings from the Chip-Chip experimental protocol (Lee et al., 2006). We assume that Chip-Chip assays for one protein are generally comparable, and that the sensitivity of the immunoprecipitation step, which is protein-specific, is the largest contributor to the variations in signal-to-noise ratio across chips.

The simplest normalization of the observed binding ratios $\hat{o}_{ti \jmath}$ assumes that Microarray assays are comparable and that measurements with excess noise have already been processed. We normalize all entries corresponding to protein $i$ by the second largest intensity observed for that protein across sites and experiments (indicated as $\text{max}_{2n, t, j}$ below) and let both first and second largest entries be equal to one.

$$\hat{s}_{ti \jmath} = \gamma \times \text{min} \left( \frac{\hat{o}_{ti \jmath}}{\text{max}_{2n, t, j} \hat{o}_{ti \jmath}}, 1 \right)$$ (5.13)

The parameter $\gamma$ controls the bias to form complexes and has to be set experimentally. We interpret the resulting $\hat{s}_{ti \jmath}$ as the combined allocation to site $j$ of all complexes in

106
which protein $i$ participates, including single protein $i$,

$$\hat{s}_{tij} = \sum_{c \in C_j; i \in c} s^*_{tcj}. \quad (5.14)$$

In practice, however, depending on the source of data, the assumptions above may need to be complemented with additional considerations about noise levels and measurement confidence. In appendix C we review two modifications to this preprocessing method that are specific to the kind of Microarray assays used in our empirical analysis.

**Learning Algorithm**

We propose an algorithm to minimize (5.12) in three steps.

**Algorithm 5**: Algorithm to learn the nucleus specific economy and its allocation at equilibrium jointly.

**Input**: Observations $\hat{s}_{tij}$ preprocessed from $\hat{o}_{tij}$ and a choice of Loss function. Line minimization parameters. Constraints $\hat{s}_{tij} = \sum_{c \in C_j; i \in c} s^*_{tcj}$.

**Output**: $K_{cj}, R_{cj}, s_{tcj}$ and $f_{ti}$ that minimize (5.12). Note that this minimization is a compromise between solving the fixed point system, fitting the observations and regularizing the parameters of the economy.

Consequently the allocations $s_{tcj}$ predicted in this manner may differ from the prediction of the fixed point equations using the same $K_{cj}, f_{ti}$ due to noise.

1. **Iterate** Outer Loop
2. Find $\{s^*\} \in \hat{S}$ given $s_{tcj}$ and a Loss function $Loss(\hat{S}, s)$.
3. **Iterate** Inner Loop
4. Find $\{R, K\}$ that minimize $F(R, K)$, (see (5.16) below).
5. Find $\{f\}$ using the relation $f_i - R^*_i = (1 - \varepsilon)f_i$.
6. **Stop** when $F_L$ remains constant within the inner loop
7. **Stop** when $F_L$ remains constant in the outer loop

107
Find the set \( \{ s_{tcj}^* \} \in \hat{S} \), given predictions \( s_{tcj} \)

We are interested in choosing the set \( s_{tcj}^* \) that best fits the given solution of the game \( (s_{tcj}) \). The Kullback-Leibler divergence is adequate to compare two distributions. We first define the loss function to compare predictions with observations

\[
Loss(s) = \min_{s^* \in \hat{S}} D(s_{tcj}^* \| s_{tcj}).
\]

And define an extended loss function to find \( s_{tcj}^* \) as

\[
Loss(s, s^*) = D(s_{tcj}^* \| s_{tcj}),
\]

where \( Loss(s) = \min_{s^* \in \hat{S}} Loss(s, s^*) \).

To turn \( s_{tcj}^* \) into a distribution over all complexes \( c \in C_j \) we extend the set \( C_j \) with the “empty protein” (\( \emptyset \)) that represents the time the site is unoccupied. We can then find \( s_{tcj}^* \) via the following minimization:

\[
\min_{s^* \in \hat{S}} D(s_{tcj}^* \| s_{tcj}) = \min_{s^* \in \hat{S}} - \sum_{c \in C_j \cup \emptyset} s_{tcj}^* \log \frac{s_{tcj}}{s_{tcj}^*} \quad s.t. \quad \hat{s}_{tji} = \sum_{c \in C_j : i \in c} s_{tcj}^*, \quad \forall i
\]

\[
\sum_{c \in C_j \cup \emptyset} s_{tcj}^* = 1 \quad (5.15)
\]

This problem can be seen as a maximum entropy problem with equality constraints. It can be solved with an infeasible start Newton’s method for equality constrained optimization (cf. Boyd and Vandenberghe (2006))
**Find** \( \{R, K\} \) **that minimize the** \( F(R, K) \)

For fixed \( s^*_{cj}, f_{ti} \), the learning cost function simplifies to,

\[
F(R_{ti}, K_{cj}) = F_L + \lambda_1 F_{BIO} + \lambda_2 F_{L1}
\]

s.t.

\[
K_{cj} > 0 \\
f_{ti} > 0 \\
R_{ti} > 0
\]

\[
s_{tcj}(R_{ti;i\in c}, K_{cj}) = \frac{K_{cj} \prod_{i\in c} R_{ti}}{1 + \sum_{c'\in C_j} K_{c'} \prod_{i\in c'} R_{ti}} \tag{5.16}
\]

This is an optimization problem jointly convex in the log-scale variables \( \xi, k \). We can solve it using an unconstrained newton’s method (cf. (Boyd and Vandenberghe, 2006)). In practice we just need to compute the Hessian. We can compute the hessian using a path parameter or directly. The details of this computation are in appendix B.2

**Find** \( \{f\} \)

The total availability of each protein, \( \{f\} \), is related to the scale (units) of the problem. In chemistry, the total availability of protein \( f \) is an integration constant in the integration of the rate of reaction, and is usually set by balance of mass. Estimating \( f \) from microarray data poses a challenge because in the absence of scale information, the values of \( \{f\} \) will be largely dependant on the choice of normalization.

In our approach \( f_i \) defines the tradeoff between unbound protein \( R \) and the equilibrium constants \( K \) of each of the reactions in which the protein participates. \( F_{fR} \) as defined above simply sets that trade-off \( K - R \) so that almost all protein \((1 - \epsilon)\) is bound. To prevent \( f \) from growing to infinity, however, we also need to couple \( f \) with the actual binding rates. This is equivalent to determining the limiting reagent.
Choosing the site to be the limiting reagent implies

\[ f_{ti} = \frac{\sum_{j} \sum_{c \in C : i \in c} s_{tcj}}{(1 - \varepsilon)} \]

When using Microarray data without information about scale, we fix

\[ f_{ti} = \frac{\sum_{j} \hat{s}_{tij}}{(1 - \varepsilon)}, \]

because there is no relative scale information to be learned (if there was any, it was removed in the normalization during preprocessing.)

When information about scale is available, we may add the constrain above as a bias. For instance, as a new KL divergence term that compares the distribution of \( f_{ti} \) to the distribution defined by the limiting reagent \( (s_{tcj}) \).

\[ F_{fR} = \sum_{ti} f_{ti} \log \frac{\sum_{j} \hat{s}_{tij}}{(1 - \varepsilon)f_{ti}} \]
Part II

Experimental results
Two key results are presented. The first is a demonstration that the DNA binding economies reproduce a genetic switch in accordance with the qualitative and quantitative literature on the subject. The genetic switch is the viral infection of E. Coli by the \( \lambda \)-phage virus \textit{Ptashne} (1987). The second main result shows that the economic approach can complete the map of coordinate binding of regulators in yeast (\textit{Harbison et al.}, 2004).

The following two chapters are structured as follows. First we analyze the DNA binding equilibria when the parameters of the economy are given. This part of the analysis is limited to \( \lambda \)-phage. The number of examples of genetic regulation with known tabulated constants is limited. The second part of the analysis focuses on the case where the parameters are learned from Microarray data; beginning with \( \lambda \)-phage to enable comparison with the previous section, and followed by a large-scale example in Chapter 7.
Chapter 6

Small Scale Experiments

6.1 Forward Predictions with Parameters from Experiments

6.1.1 Overview of λ-phage Infection

λ-phage is a well-studied example of viral infection in Escherichia Coli (Herskowitz and Hagen, 1980; Shea and Ackers, 1985; Arkin et al., 1998). Bacteria infected by the λ virus display no symptoms when the virus is dormant; they are said to be in lysogeny state. When the virus is activated, aggressive replication of the viral DNA leads to the bacterium bursting; bacteria in this state are said to be in lysis state. The state of the λ-phage infection is governed by a genetic two-state control switch that specifies whether to transition from lysogeny to lysis. The components of the λ—switch are 1) two adjacent genes cI and Cro that encode cI2 and Cro proteins, respectively; 2) the promoter regions P_{RM} and P_{R} of these genes, and 3) an operator (O_{R}) with three binding sites O_{R}1, O_{R}2, and O_{R}3 that partially overlaps the promoter regions. Figure 6-1 illustrates schematically the geometry of the switch. We focus on the lysogeny phase, during which cI2 dominates over Cro. There are two relevant protein-agents, RNA-polymerase and cI2 and three sites, O_{R}1, O_{R}2, and O_{R}3. Presence of cI2 in either O_{R}1 or O_{R}3 effectively blocks access of RNA-polymerase to promoter regions
Figure 6-1: Representation of cI and Cro genes, promoters and operator sites as they appear in the λ chromosome. Arrows indicate direction of transcription. OR1 overlaps with the Cro promoter PR, and OR3 overlaps with PRM.

PR, or PRM respectively; hence we can safely restrict ourselves to operator sites as the site-agents.

Full specification of an economy for lambda-phage requires transport affinities and chemical equilibrium constants. Chemical equilibrium constants are available, we discuss their computation in Section 6.1.2. Affinities, however, involve a certain ambiguity. Affinities may refer to an activation energy of binding, an energy of binding, or a combination of both. Here we use the word affinity as an indication of activation energy.

We first outline key considerations about the affinities in the lysogeny cycle. RNA-polymerase can bind either promoter but does not bind OR2. The affinity of cI2 protein to bind the operator sites can be summarized as OR1 > OR2 ≃ OR3. The cycle goes through three phases largely dependent on the concentration of cI2:

1. cI2 binds to OR1 first and blocks the Cro promoter PR

2. Slightly higher concentrations of cI2 lead to binding at OR2 which in turn facilitates RNA-polymerase to initiate transcription at PRM

3. At sufficiently high levels cI2 also binds to OR3 and inhibits its own transcription

The first two phases appear to contradict known affinities: cI2 binds almost immediately at both OR1 and OR2 (in this order), and the presence of cI2 in OR2 results in increased transcription of cI2. These effects are attributed to protein-protein interactions between cI2 dimers at OR1 and OR2, and between cI2 and RNA-polymerase. Such protein-protein interactions could be encoded in our game theoretic model via
additional structural constraints on the utilities and strategies of the site players. We specifically avoid such encoding, however, and instead attempt to explain the observed effects simply through competition and resource constraints. For example, in our model, the order in which $cI_2$ binds to the three operator sites is a byproduct of $cI_2$ being transported differently to these sites, and the competition between $cI_2$ and RNA-polymerase to bind $O_{R3}$ and $O_{R2}$. Note that the spatial proximity of the sites places physical constraints on binding and such constraints may be implicit in the affinities.

Table 6.1 summarizes the qualitative affinities with and without protein-protein interactions.

The complete $\lambda$-phage switch has been extensively studied with simulation methods (Shea and Ackers, 1985; Arkin et al., 1998). Our economic abstraction of the $\lambda$-phage switch complements these approaches. Our goal is to abstract out the dynamics of the binding arrangements in the same way that the chemical equilibrium constant provides an abstraction to compute the concentration of chemical species at equilibrium. Computing similar averages using a full kinetic analysis involves averaging the responses of a genetic circuit to a stochastic input. Figure 6-2 illustrates intuitively the difference in complexity between the economic models of lambda-phage and the corresponding genetic circuit.

The $\lambda$-phage economy can be complemented with kinetic information mapping changes in protein availability to time to study the dynamics of equilibrium states. Without time information it is still possible to examine the dynamics of equilibrium states as a function of protein availability. Our results show that DNA binding economies can successfully track the order of events in $\lambda$-phage lysogeny as a function
Figure 6-2: The figure to the left is a graphical representation of the site-specific economy for λ-page. The figure to the right is a genetic circuit representation of the full λ switch taken from (Arkin et al., 1998). For simplicity our economies only consider the lysogeny state. Reproducing the entire switch would have required time dependent concentrations of protein species. While possible, we have restricted our attention to the lysogeny cycle to introduce an uncluttered perspective of the site specific economy.

of the concentration of cI2.

### 6.1.2 Parameters for the λ-phage Site Specific Economy

The site specific economy requires three sets of parameters: chemical equilibrium constants, affinities, and initial availabilities of each different protein species. We set the chemical equilibrium constants in accordance with the Gibbs’ Free energies $\Delta G$ tabulated by Shea and Ackers (1985),

<table>
<thead>
<tr>
<th>$\Delta G$ (kCal)</th>
<th>$O_{R3}$</th>
<th>$O_{R2}$</th>
<th>$O_{R1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cI2</td>
<td>−10.1</td>
<td>−10.1</td>
<td>−11.7</td>
</tr>
<tr>
<td>RNA-p</td>
<td>−11.5</td>
<td>0</td>
<td>−12.5</td>
</tr>
</tbody>
</table>

To incorporate these energies into our simulation, we first have to quantify the units of the equilibrium equation (3.4),

$$K_{ij} = \frac{s_i^j}{(p_j^i f_i^j - s_i^j)(1 - \sum_{\nu} s_{j,\nu}^i)}.$$
To ensure the consistency of units, we define $f^i$ as the total number of proteins $i$ available, and arrange the units of $K_{ij}$ accordingly:

$$f^i \equiv \tilde{f}^i V_T N_A,$$

$$K_{ij} \equiv \tilde{K}_{ij} \frac{1}{N_A V_S},$$

(6.1) (6.2)

where $V_T$ and $V_S$ are the volumes of cell and site neighborhood, respectively, $N_A$ is the Avogadro number, $\tilde{f}^i$ is the concentration of protein $i$ in the cell, and $\tilde{K}_{ij}$ is the equilibrium constant in units of $\ell/\text{mol}$. The equilibrium equation can then be rewritten as

$$s_j^i \left( p_j^i \tilde{f}^i V_T N_A - s_j^i \right) \left( 1 - \sum_{i'} s_{j,i'} \right) = \tilde{K}_{ij} \frac{1}{N_A V_S}$$

and rearranging the terms, the relationship with the Gibbs' free energies unfolds

$$s_j^i \left( p_j^i \frac{\tilde{f}^i V_T N_A}{V_S N_A} - s_j^i \right) \left( 1 - \sum_{i'} s_{j,i'} \right) = \tilde{K}_{ij} \frac{1}{V_S N_A}$$

$$= e^{-\Delta G / RT},$$

(6.3)

where $R$ is the universal gas constant and $T$ is temperature. For a typical *Escherichia coli* (2$\mu$m length), we obtain the following chemical equilibrium constants from the tabulated free energies

<table>
<thead>
<tr>
<th>$K_{ij}$</th>
<th>$O_R 3$</th>
<th>$O_R 2$</th>
<th>$O_R 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cI$_2$</td>
<td>.0020</td>
<td>.0020</td>
<td>.0296</td>
</tr>
<tr>
<td>RNA-p</td>
<td>.0212</td>
<td>0</td>
<td>.1134</td>
</tr>
</tbody>
</table>

Note that when such parameters are learned from data any dependence on the volumes will be implicit.

Similarly, we set the transport affinities in accordance with the qualitative description in (Ptashne, 1987; Ptashne and Gann, 2002), summarized in Table 6.1:
We set $\tilde{f}_{\text{RNA}−p} = 30\text{nM}$ (cf. Shea and Ackers (1985)) which for a typical $E.\ coli$ is equivalent to setting $f_{\text{RNA}−p} \simeq 340$ copies. And then varied $f_{cI_2}$ from 1 to 10,000 copies to study the dynamical behavior of the lysogeny cycle. The results are reported as a function of the ratio $f_{cI_2}/f_{\text{RNA}−p}$. We set $\beta = 10^{-5}$.

### 6.1.3 Reproducing λ-phage Lysogeny

The predictions from the site-specific DNA binding economy exactly mirror the known behavior. We emphasize that our model does not encode protein-protein interactions, yet is able to account for the experimental observations.

We present the results as a function of varying concentrations of $f_{cI_2}$. Table 6.2 shows the distribution of protein near and bound at each site at specific concentrations of $cI_2$. The amount of RNA-p is assumed to remain constant. Table 6.2 shows that the simulation mirrors the behavior of the lysogeny cycle discussed earlier.

1 When no $cI_2$ is present, RNA-polymerase is only slightly more likely to bind to $O_{R1}$ than $O_{R3}$. The natural tendency towards lysogeny observed experimentally

---

<table>
<thead>
<tr>
<th>$p_{ij}$</th>
<th>$f_{cI_2}/f_{\text{RNA}−p}$</th>
<th>$O_{R3}$</th>
<th>$O_{R2}$</th>
<th>$O_{R1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$cI_2$</td>
<td>1/100</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RNA-p</td>
<td></td>
<td>0.46</td>
<td>0</td>
<td>0.54</td>
</tr>
<tr>
<td>$cI_2$</td>
<td>2</td>
<td>0</td>
<td>.88</td>
<td>0.12</td>
</tr>
<tr>
<td>RNA-p</td>
<td></td>
<td>0.49</td>
<td>0</td>
<td>0.51</td>
</tr>
<tr>
<td>$cI_2$</td>
<td>10</td>
<td>0</td>
<td>.55</td>
<td>0.45</td>
</tr>
<tr>
<td>RNA-p</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$cI_2$</td>
<td>100</td>
<td>0.32</td>
<td>.39</td>
<td>0.28</td>
</tr>
<tr>
<td>RNA-p</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s_{ij}$</th>
<th>$f_{cI_2}/f_{\text{RNA}−p}$</th>
<th>$O_{R3}$</th>
<th>$O_{R2}$</th>
<th>$O_{R1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/100</td>
<td></td>
<td>0</td>
<td>.001</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.77</td>
<td>0</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>0.56</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.78</td>
<td>0</td>
<td>0.85</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0</td>
<td>0.83</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.88</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>.78</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.19</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.2: (left) Distribution of proteins near the sites. The numbers represent the fraction of proteins allocated to the sites. (right) predicted occupancy of site $j$ by protein $i$, can be interpreted as a probability of binding at each site.
can only be explained by factors external to this analysis such as transcription time, or presence of additional proteins.

2 As more cI₂ becomes available, it accumulates near O₁, O₂, increasing the probability of finding cI₂ at either sites to nearly one. Further increase of cI₂ leads to almost exclusive binding of RNA-polymerase to O₃.

3 Finally, at high concentrations cI₂ blocks its own promoter and reduces RNA-p binding at O₃. Figure 6-3(a) shows how increased cI₂ concentration sharply reduces RNA-polymerase binding at O₃.

Figure 6-3: Predicted protein binding to sites O₃, O₂, and O₁ for increasing amounts of cI₂. The curves are created running Algorithm 1 with different values of f\(_{cI₂}\). The rightmost figure illustrates a comparison with (Arkin et al., 1998). The shaded area indicates the range of concentrations of cI₂ at which stochastic simulation predicts a decline in transcription from O₁. Our model predicts that cI₂ begins to occupy O₁ at the same concentration.

Figure 6-3 illustrates how the binding at different sites changes as a function of increasing f\(_{cI₂}\). The simulation mirrors the behavior of the lysogeny cycle discussed earlier. Although our model does not capture dynamics, and the figure does not involve time, it is nevertheless useful for assessing quantitative changes and the order of events as a function of increasing f\(_{cI₂}\). Note, for example, that the levels at which cI₂ occupies O₁ and O₂ rise much faster than at O₃. While the result is expected, the behavior is attributed to protein-protein interactions which are not encoded in our model. Similarly, RNA-polymerase occupation at O₃ bumps up as the probability that O₂ is bound by cI₂ increases.
Finally, Figure 6-3(c) shows a comparison with stochastic simulation (v. (Arkin et al., 1998)). Our model predicts that cI$_2$ begins binding O$_R1$ at the same level as (Arkin et al., 1998) predicts a decline in the transcription of Cro. While consistent, we emphasize that the methods differ in their goals; stochastic simulation focuses on the dynamics of transcription while we study the strategic allocation of proteins as a function of their concentration.

Figure 6-4 illustrates the distinct nature of the predictions that can be achieved with this method. The figure shows the $\lambda$-phage economy as a block element. Dynamic behavior has been abstracted out of the computation of the strategic allocation of proteins. However it is still possible to consider the dynamic response of this block as part of a larger system by studying the evolution of concentrations $f_i$ in time.

**Causal Predictions: Simultaneous Occupancy of O$_R1$ and O$_R2$.**

O$_R1$ knockout experiments have shown that protein-protein interactions between cI$_2$ dimers are largely responsible for simultaneous occupancy of sites O$_R2$ and O$_R1$. While agreeing with that observation, Figure 6-3 suggests that the cooperative binding can also be obtained as a by-product of competition involving RNA-polymerase, cI$_2$, O$_R1$, and O$_R2$. To assess the validity of this hypothesis we simulated O$_R1$ knockout experiments by substantially reducing the equilibrium constants at this site. This simulation illustrates the ability of the economic models of DNA binding to make causal predictions.
With \( O_R1 \) knocked out, our model predicts that \( cI_2 \) will bind \( O_R3 \) and \( O_R2 \) similarly, with minor initial differences due to competition between \( cI_2 \) and RNA-polymerase at \( O_R3 \). Figure 6-5 reproduces the qualitative behavior observed in knock-out experiments. RNA-polymerase binds \( O_R3 \) at first but \( cI_2 \) takes over at the same rate as it binds to \( O_R2 \). Only if concentration of \( cI_2 \) became sufficiently high do we find \( cI_2 \) at the mutated \( O_R1 \) as well. Note, however, that \( cI_2 \) inhibits transcription at \( O_R3 \) prior to occupying \( O_R1 \). Thus the binding at the mutated \( O_R1 \) could not be observed without interventions.

### 6.1.4 Empirical Complexity Analysis of the Site Specific Economy.

Analyzing the complexity supermodular games and economies is generally hard. As indicated in Chapter 4, the inner loops of our algorithms depend polynomially on the number of sites and proteins. However, ultimate convergence depends on the relative values of affinities and equilibrium constants.

For instance, consider Algorithm 1 for the site specific economy. Convergence for the binary search is sublinear and site occupancy (\( \alpha \)) updates are polynomial on
the number of sites and proteins. The outer loop, however, depends on the rate at which site occupancies increase, which is a function of how dissimilar are the economic parameters at each site, $\beta$, and the tolerance of the algorithm. To quantify the rate of convergence we analyze empirically the number of outer iterations the algorithm required to converge for different number of proteins and sites, and randomly set economic parameters.

Figure 6-6 shows the number of iterations required to converge to a solution for different number of proteins and sites. For each economy (defined by the number of proteins and sites) we ran 15 different trials with random values for the economic parameters. Figure 6-6 confirms the intuition that the number of outer iterations does not depend on the number of proteins and sites.

Figure 6-6: Number of outer iterations required to converge for different instances of the site specific economy. We ran the forward prediction algorithm for different configurations of the site specific economy. Each configuration corresponds to a fixed number of proteins and sites with random values for affinities, chemical equilibrium constants, and initial protein availability. To ensure that each configuration was biologically meaningful, we scaled the economic parameters to be similar in average to the values reported for $\lambda$-phage. We ran 15 random trials for each configuration. The figures above show the mean and standard deviation of the number of iterations the algorithm required to converge with tolerance $10^{-4}$ and $\beta = 10^{-3}$. Variability for different random trials is a consequence of relative changes in affinities and equilibrium constants. (a) Shows the dependence of the number of iterations in the number of proteins when the site number is fixed to 10. There is a slightly upward trend in the mean number of iterations (from 2 iterations to 8 iterations when going from 10 to 200 proteins), which seems to indicates a logarithmic dependence on the number of proteins. The increase in standard deviation, however, would indicate that this effect is more likely due to an increase in the number of similar parameters per site due to randomization. (b) Shows dependence in the number of sites when protein number is kept at 10. Both (a) and (b) indicate that the dependence of the number of iterations in the number of proteins or sites is minimal. (c) shows the number of iterations as a function of total number of proteins and sites. Figure (c) indicates that the number of iterations of the outer loop does not depend appreciably in the number of proteins and sites.

122
Figure 6-6 illustrates a key feature of the DNA binding economies, their scalability. Only the update rule depends on the biological scale of the problem, and dependence is polynomial. Ultimate convergence is a function of tolerance and the values of the economic parameters.

6.1.5 Toy example: Promoter Specific Economy

The coarser structure of the promoter specific economy is well suited for modelling binding arrangements of seemingly independent subsystems. For instance, promoters within independent subsystems that compete for the same proteins. Herman et al. (1993) discovered one such example in Escherichia Coli infected by λ-phage. Even in the context of this well-known sub-system, however, few quantitative experimental results are available about binding.

We illustrate the difference between site specific and promoter specific economies with a toy example constructed from λ-phage infection. This situation is a convenient artificial construct. Ptashne (1987) indicates that, in natural conditions, only one λ-phage DNA is inserted in the host DNA. Figure 6-7 compares the binding profiles at two identical λ-promoters with equal affinity. As expected both binding profiles are identical.

The most notable difference between the site specific economy (Figure 6-3) and Figure 6-7 is the smoother evolution of the binding profiles. In the promoter specific economy, affinities pertain to promoters, and no longer capture site-specific activation energy differences. Binding at sites within the promoter is exclusively governed by chemical equilibrium. In contrast in the site specific economy transport affinities modulate protein availability at the site level. The site-specific setting is designed to capture finer allocation to sites, while the promoter-specific setting focuses on the interplay between subsystems.

Consider a slight modification to the previous setting in which one of the lambda-promoters has a (10-fold) reduced affinity for $cI_2$ and the other has a (100-fold) reduced affinity for $RNA - p$. Equilibrium constants remain unaltered. The resulting binding profiles are shown in Figure 6-8. The most notable difference is the concentra-
Figure 6-7: Predicted protein binding to sites $O_R3$, $O_R2$, and $O_R1$ in each of the two identical $\lambda$-phage promoters, for increasing amounts of $cI_2$. As expected both $\lambda$-phage promoters behave identically when the affinities are equal. Note the smoother evolution of binding compared to the site specific economy (Figure 6-3). Affinities in the promoter specific economy only affect promoters and thus binding at sites is exclusively governed by chemical equilibrium. In the site specific economy affinities had an additional effect on each site individually.

The concentration of $cI_2$ at which binding reversals occur. In a hypothetical experiment we would observe that while one $\lambda$ switch has reached self-repression the other continues to produce $cI_2$ at a fairly high rate. The different behavior is due to resource allocation.

Figure 6-8 illustrates the ability of the economic approach to capture resource allocation explicitly. General probabilistic and simulation approaches do not explicitly consider limiting resources in the activation/repression of different DNA subsystems. In general the hypothesis that the cell is a well-mixed system leaves the resolution to the chemical equilibrium equation. The promoter and site specific economies capture the role of limiting resources at different scales by means of the convenient abstraction of affinities.
Figure 6-8: Predicted protein binding to sites $O_R3$, $O_R2$, and $O_R1$ in two $\lambda$-phage promoters, for increasing amounts of $cI_2$. In this example the affinity of the first promoter for $cI_2$ has been decreased 10-fold and the affinity of the second promoter for $RNA-p$ has been decreased 100-fold. The same general behavior is observed independently for each $\lambda$-phage promoter. However, binding reversal events occur at different concentrations of $f_{cI2}$.

Uniqueness of The Solution

Figure 6-9 concludes the empirical analysis of the promoter specific economy with an empirical analysis of the bounding algorithm. While section 4.2 does not guarantee uniqueness of the competitive equilibrium of the promoter based economy, Figure 6-9 indicates that for the cases considered in this section the equilibria corresponding to varying concentrations were unique.

6.2 Learning the Parameters

6.2.1 Learning a Site Specific Economy for $\lambda$-phage Lysogeny

We show that the learning algorithm can recover the economic structure that accurately predicts the whole lysogeny cycle in a realistic experimental setting. We consider
Figure 6-9: Upper and lower bounds on protein binding to sites $O_R^3$, $O_R^2$, and $O_R^1$ in each of the two $\lambda$-phage promoters, for increasing amounts of $cI_2$. Predicted using algorithm 2. We note also that for each promoter and setting of $f_i$, $E_{ir}$, and $K_{ij}$ bounds converged to within $10^{-14}$ in under 10 iterations.
an experiment design setup likely to occur in early laboratory experiments; we explore a range of values of $f_{cI_2}$ and $f_{RNA-p}$ that do not reveal any of the binding reversals observed above. In particular we are interested in seeing that the structure uncovered by our learning algorithms can assist the choice of the next experiment, even when the initial design omits information about key events in the subsystem.

**Simulated Measurements**

We simulated 5 measurements with constant $f_{RNA-p} = 340$ and increasing $f_{cI_2} \in [1, 45]$ and another set of five experiments with constant $f_{cI_2} = 340$ and $f_{RNA-p} \in [1, 45]$. The first set of experiments covers the beginning of the curves in Figure 6-3, the second set has sufficient $cI_2$ to bind at the first two operator sites but not enough RNA-p to show any appreciable effect of competition. 10 repetitions of the above experiments were simulated by addition of gaussian noise $\mathcal{N}(0, 0.1)$. Note that 0.1 noise is substantial for low-binding events. Figure 6-10 compares one of the noisy simulated experiments with the binding ratios used as source.

We discuss how are microarray measurements obtained experimentally and the processing required to compare them to binding predictions in appendix C. Additionally, our algorithm to learn the site-specific economy requires an initial guess for the parameters of the economy $(K^0, E^0, f^0)$. Ideally the algorithm should be seeded with an initial guess that contains overall scale information that is missing from the microarray measurements. Appendix C discusses one way to initialize the parameters when such information is unavailable.

**Results**

Table 6.3 shows the parameters learnt for the site-specific λ-phage economy using the initialization discussed in appendix C. The actual values differ from the values obtained in section 6.1.2. However, affinities and equilibrium constants follow roughly the same partial order. The most noticeable difference is the high variability of protein availability.

The values 6.3 are a perfect illustration of the difficulty to set the scale using
Figure 6-10: The top row is the predicted protein binding at sites $O_{R3}$, $O_{R2}$, and $O_{R1}$ for two sets of 5 experiments each. In the first set $f_{RNA-p} = 340$ and $f_{cI^2} \in [1, 45]$, in the second set, $f_{cI^2} = 340$ and $f_{RNA-p} \in [1, 45]$. Note that the data set does not contain experiments during the binding reversal. The bottom row is the result of adding gaussian noise to the predictions to simulate a microarray experiment. Scaling effects due to signal amplification are not shown here.

Figure 6-11 shows that these parameters recover the behavior of $\lambda$-phage (compare to Figure 6-3) to great accuracy. This result illustrates the ability to uncover the structure of the economy from a set of noisy observations of equilibrium states that was severely limited by experiment design choices.

In a real experimental setting Figure 6-11 would be an indication that our experiments missed an important range of concentrations. The range of concentrations where binding reversals occur. In this particular example the site-specific economy would help us determine the best next set of experiments to produce. We note that the range of concentrations at which Figure 6-11 predicts a binding reversal matches very closely the range from Figure 6-3. To complete the results, Figure 6-12 shows the pattern of allocation to sites achieved via the transport mechanism. Note that
Table 6.3: Parameter values learnt using the initialization method described in appendix C. Values and standard deviations result from averaging the parameters learnt from ten different instances of simulated microarray measurements.

The transport mechanism continues to adapt allocation to protein near sites even after the site has reached a constant binding profile.

The last result in this section illustrates the advantage of using an informed initial guess for \((K^0, E^0, f^0)\) with an extreme example. Table 6.4 shows the parameters learnt by our algorithms in a setting that mirrors that of Table 6.3. However, in this case, the values computed in Section 6.1.2 with 10% additive noise are used as initial guess. Our algorithms succeed in removing the noise from our initial guess.

Table 6.4: Parameter values learnt when the learning algorithm is fed with the physical constants given in section 6.1.2 with 10% additive noise. Values and standard deviations result from averaging the parameters learnt from ten different instances of simulated microarray measurements.

This table may also be interpreted as an illustration of a sensitivity analysis of our learning algorithms. A complete sensitivity analysis however would have to consider a wider range of noise values individually for measurements and each of the initial guesses, as well as a study of the range of concentrations included in the initial set of
Figure 6-11: Predicted protein binding to sites $O_{R3}$, $O_{R2}$, and $O_{R1}$ for increasing amounts of $cI_2$, obtained using the parameters from Table 6.3. In these simulations $f_{RNA-p} = 261$, the mean value observed in the constant $RNA-p$ experiments. The figures reproduce the expected behavior for $\lambda$-phage to great accuracy, particularly the binding reversal events that were not present in the training data. Note that the “bump” in RNA-p binding in $O_{R3}$ is smoother due to the high binding prediction. Binding of RNA-P at $O_{R2}$ is the most noticeable difference with our previous behavior and it is a side-effect of the addition of noise. If this were a true experimental setting it could have been due to either noise or non-specific binding.

Figure 6-12: Predicted transport of proteins to sites $O_{R3}$, $O_{R2}$, and $O_{R1}$ for increasing amounts of $cI_2$. 
Chapter 7

Large Scale Experiments

Systems for which the biological parameters such as equilibrium constants and affinities are known are rare and generally limited to small scale subsystems studies in isolation. The study of large-scale systems is mostly limited to functional studies with probabilistic methods that use Microarray assays as input. Our results show that the learning algorithms we introduced in Chapter 5 can be used to scale the economic models to large-scale examples.

Validating large-scale results is, however, challenging because our knowledge of the underlying processes at that scale is often incomplete. We address this challenge by introducing three new types of predictions that summarize the economic structure in different ways to enable comparison with existing work.

This chapter presents results on yeast regulation and new kinds of predictions derived from the economic structure: a new functional similarity measure, and new predictions about coordinate behavior. Our results show that our models can be used to complete the map of regulators in Yeast.
7.1 Scaling up: Learning a Nucleus Specific Economy for Yeast

We demonstrate the scalability of the economic approach using the nucleus specific economy. The nucleus specific economy can predict both competitive and coordinating behavior.

We model our economy after Chromatin Immuno-Precipitation (ChIP) Microarray data from (Harbison et al., 2004). Data includes binding regulation profiles for up to 203 regulators. All regulators were profiled in a rich media environment (YPD) and at least 84 among those were profiled in at least one of 12 different growth conditions.

Our algorithms explain variations between growth conditions as a function of regulator concentration. However, the equilibrium constants learnt are valid across growth conditions.

We first illustrate the main results from the nucleus specific economy in a small sample of proteins and sites, and then show results for a large sample of the network. We show that our algorithms learn a sparse representation of the regulatory network, and introduce some examples of the additional conclusions that can be inferred from the structure we learn.

7.2 Results

7.2.1 Reduced Set of Genes and Proteins.

We first illustrate the predictions of our model in a subsample of genes(42), proteins(15), and conditions(10). The sample was selected to include: regulators profiled in at least 4 conditions, conditions for which at least 3 regulators were profiled, and genes with at least a 1% of binding events across conditions and regulators; where a binding event is defined as in Lee et al. (2006). Preprocessing of the Microarray data was done according to the method explained in appendix C. Missing data was filled with YPD values.
Figure 7-1 shows the chemical equilibrium constants (i.e., the economic structure) learnt by our algorithms interpreting microarray assays as noisy observations of competitive equilibrium. Note the sparse structure.

Figure 7-1: Genes appear in abscissa, proteins and complexes in ordinates. The map displays the logarithm of the equilibrium constant for each complex-site pair, which can be interpreted as an energy measure. The color-scale indicates magnitude in log-scale above background noise. Darker values indicate higher-valued equilibrium constants. The dendrograms at the left and the top show hierarchical clusters of complexes and genes respectively. Clusters where computed using the average cosine distance between energy profiles. Where the energy profile of, for instance, a gene \( j \), is the vector of equilibrium constants of all complexes for that site in logarithmic scale (\( \{\log K_{cj}\}, \forall c \)). Note the sparsity of the structure learnt; white values indicate null equilibrium constants.

Functional Similarity by Reactive Profile

Figure 7-1 also illustrates the use of the structure of the economy to uncover regulatory patterns. The dendrogram at the top shows hierarchical clusters of genes, obtained by comparing their reactive profiles. We define the reactive profile of a gene by the vector of logarithms of equilibrium constants of the binding reaction of that gene with each complexes. Intuitively, genes that display preferences for binding the same
regulators (up or downstream) with similar strength can be expected to be involved in
the same regulatory processes. Note that unlike sequence comparisons, our measure
relies on the binding reactions at each gene. Therefore it does not require making
assumptions about the relevance of particular base-pair deletions, translocations, and
insertions.

We compared biological gene function as annotated in “YPD” and “genRE”
databases Hodges et al. (1999); Mewes et al. (2004); Gldener et al. (2006) in the
top two clusters. Already for this small sample, gene clusters map the biological
function adequately. The left cluster, groups genes primarily known for their role in
protein synthesis, whereas the cluster to the right groups a more varied set of genes,
many of which (50%) seem to have a regulatory or catalytic role. In particular,
genues known to be active under stress by heat, with functions in many processes
(regulatory/catalytic, cell rescue, metabolism, cell transport, cell fate) were grouped
together. Note that our economic algorithms predicted an energy profile for gene
YDL027C, whose function is currently unknown, similar to that of the heat stress
group. Table 7.1 lists the functions associated with each gene in the aforementioned
databases.

Finally, we can define a reactive profile for each complex similarly. Each cluster of
genues can then be associated with a family of proteins and complexes. By inspection
Figure 7-1 reveals that the left cluster which we associated with protein synthesis
is primarily defined by a set of regulatory proteins, FHL1, MSN4, MOT3, SKN7,
SFP1. These regulators, which bind the remaining sites more sparsely, are known
to be related to RNA synthesis. Our analysis would indicate that the set of genes
involved in protein synthesis is bound nearly exclusively by regulators involved in
RNA synthesis, and complexes thereof.

7.2.2 Extended Set of Genes and Proteins.

We considered an extended set of proteins (27) and sites(414) over 8 different con-
ditions (Alpha, BUT14, H2O2Hi, H2O2Lo, HEAT, RAPA, SM, and YPD). Only
proteins profiled in at least 3 conditions and conditions with at least 3 regulators
<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>YER103W</td>
<td>Protein synthesis and translation</td>
</tr>
<tr>
<td>YDL027C</td>
<td>?</td>
</tr>
<tr>
<td>YBR140C</td>
<td>6,2,11,7</td>
</tr>
<tr>
<td>YJR046W</td>
<td>10</td>
</tr>
<tr>
<td>YJR045C</td>
<td>4,7,11,5 (heat)</td>
</tr>
<tr>
<td>YDR258C</td>
<td>4,7,11,5 (heat)</td>
</tr>
<tr>
<td>YMR186W</td>
<td>8,5,11 (heat)</td>
</tr>
<tr>
<td>YDR152W</td>
<td>?</td>
</tr>
<tr>
<td>YDR151C</td>
<td>5,12</td>
</tr>
<tr>
<td>YLR042C</td>
<td>?</td>
</tr>
<tr>
<td>YJR123W</td>
<td>1</td>
</tr>
<tr>
<td>YGR211W</td>
<td>11,10</td>
</tr>
<tr>
<td>YGR210C</td>
<td>?</td>
</tr>
<tr>
<td>YNL007C</td>
<td>8,10,1,11</td>
</tr>
</tbody>
</table>

1 Protein synthesis and translation  5 Cell Fate  9 Biogenesis
2 Cell Sensing  6 Signal Transduction  10 Cell Cycle
3 Pheromone Response  7 Metabolism  11 Binding Function or Cofactor Requirement
4 Cell Transport  8 Cell Rescue  12 Transcription

Table 7.1: Genes and function assigned to them in MIPS and YPD databases. Genes appear in the same order as in figure 7-1. The line indicates the division between the right cluster (top rows) and the left cluster (bottom rows).

were considered. Additionally, to ensure that the reference condition \( (YPD) \) was fully populated, sites with inconclusive bindings\(^1\) at \( (YPD) \) were rejected. In these experiments \( YPD \) was used as a reference for the total availability of each protein in conditions were it was unobserved. Missing observations, however, were left as unobserved and the corresponding binding inferred from the total availability of that protein in \( YPD \).

Figure 7-2 \(^2\) shows the structure learnt for the nucleus specific economy over 414 sites and 27 proteins leading to a total of 130 complex candidates (only the 313 sites and 117 complexes with relevant binding are shown). Note first that 130 complexes is a considerably smaller set of complexes than the powerset of 27 proteins and the set of all possible pairwise complexes \( (27 \cdot 26 = 702) \). The structure learned by the nucleus specific economy is remarkably sparse, leading to only 10 to 20 complex or

\(^1\)we considered a slightly relaxed definition of binding with respect to Lee et al. (2006), intensities greater that 1.5-fold were considered conclusive bindings with a p-value \( p < 0.001 \).

\(^2\)Due to their size, figures showing results on the extended set of proteins appear at the end of the chapter rather than embedded in the text.
protein bindings per site.

Figure 7-2, like Figure 7-1 earlier shows clear regulatory patterns. It is possible to identify gene-modules co-regulated by the same group of complexes, complexes that have a broad regulatory spectrum, and highly specific complexes.

**Network Representation**

The map of equilibrium constants from figure 7-2 can also be represented as a network. Figure 7-3 shows as an example the network representation of the upper right cluster of genes and regulators that groups genes known to be involved in protein synthesis.

Note that 7-3 is an intermediate between a Bayesian network and a genetic circuit representation. Like Bayesian networks, this network does not encode information about the kinetics of binding. Nevertheless, the relations shown in the network are causal (edges encode chemical equilibrium constants rather than similarities), and thus, like genetic circuits, the network can be used to make causal prediction about regulation dynamics (here we distinguish as it is common in chemistry, between reaction kinetics and dynamics of equilibria.)

**Coordinate Behavior**

The nucleus specific economy models coordinate as well as competitive behavior. We have referred to coordinate behavior loosely as complex formation. Note, however, that the prediction of complex formation from our algorithms has multiple interpretations. A combination of two or more proteins is considered a complex when their joint presence at a site can be explained consistently throughout conditions by means of the following reversible chemical reaction

\[
i_{\text{PROTEIN} 1} + i_{\text{PROTEIN} 2} + j_{\text{SITE}} \rightleftharpoons i_{\text{PROTEIN} 1}i_{\text{PROTEIN} 1}j.
\]

That is, our algorithms deduces complex formation when the presence of two proteins is best explained by a single chemical reaction involving both proteins rather than by two independent chemical reactions. This however does not clarify the mechanism by
which both proteins bind at the site. Several interpretations about the mechanism are possible. We emphasize, however, that complex formation is not simply the result from a measure of binding correlation. Table 7.2.2 indicates various possible interpretations of the meaning of complexes

<table>
<thead>
<tr>
<th>Description</th>
<th>Site specific protein complexes</th>
<th>Protein complexes</th>
<th>Coordinate binding events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexed proteins that only exist when bound at the site</td>
<td>( i_{\text{PROTEIN} \ 1} + i_{\text{PROTEIN} \ 2} + j_{\text{SITE}} \rightleftharpoons i_{\text{COMPLEX} \ 1-2j} )</td>
<td>Complexed proteins that exist both when bound at the site and when not bound</td>
<td>Binding of two proteins at a site that cannot be explained in terms of protein-protein interactions.</td>
</tr>
<tr>
<td>( i_{\text{PROTEIN} \ 1} + i_{\text{PROTEIN} \ 2} \rightleftharpoons i_{\text{COMPLEX} \ 1-2} + j_{\text{SITE}} \rightleftharpoons i_{\text{COMPLEX} \ 1-2j} )</td>
<td></td>
<td></td>
<td>( i_{\text{PROTEIN} \ 1} + i_{\text{PROTEIN} \ 2} + j_{\text{SITE}} \rightleftharpoons i_{\text{PROTEIN} \ 1i_{\text{PROTEIN} \ 1j}} )</td>
</tr>
</tbody>
</table>

Figure 7-4 illustrates predictions about coordinate behavior that can be obtained from our algorithms; complexes detected in a sample of conditions, and variations in their binding levels across conditions.

We can use the ability to detect complexes of the economic approach as an additional measure to validate the predictions of the economic approach as well as to hypothesize the formation of complexes between proteins that have not been observed yet experimentally. We conclude the illustration of the predictions of our economies with an analysis of site-specific bindings of a complex and its composing proteins.

One way to do so is by examining the complexes from figure 7-4 that display stronger binding across genes. We focus on three of the complexes predicted to change in Figure 7-4, \textit{STE12:TEC1}, \textit{YAP6:CIN5}, \textit{SKN7:CIN5}. Note that our algorithm predicted these as complex candidates. In this particular case, the first two are reported (\textit{cf. YPD-proteome database}) to either form a complex or be part of a coordinated behavior, only \textit{SKN7:CIN5} remains unreported as a complex and thus constitutes a hypothesis of our algorithms.

**STE12:TEC1.** Figures 7-5 and 7-6 indicate that our economies capture adequately the role of STE12 and TEC1 in YPD and their expected variation across conditions.
However they appear to be unable to resolve the differences between the conditions of filamentation (*But14*) and mating (*Alpha*). We attribute this lack of resolution to limitations in our sample selection; a relevant condition (*BUT90*) and some relevant sites were rejected due to low confidence measurements *high p-value*.

**YAP6:CIN5 and SKN7:CIN5**  Figures 7-10, 7-9, 7-8, 7-7 show the binding ratios for YAP6:CIN5 and SKN7:CIN5 and their respective components.
Figure 7-2: Genes appear in abscissa, proteins and complexes in ordinates. The map displays the logarithm of the equilibrium constant for each complex-site pair, which can be interpreted as an energy measure. The color-scale indicates magnitude in log-scale above background noise. Darker values indicate higher-valued equilibrium constants. The dendrograms at the left and the top show hierarchical clusters of complexes and genes respectively. Clusters were computed using the average cosine distance between energy profiles. Where the energy profile of, for instance, a gene $j$, is the vector of equilibrium constants of all complexes for that site in logarithmic scale ($\{\log K_{cj}\}, \forall c$). Note the sparsity of the structure learnt; white values indicate null equilibrium constants.
Figure 7-3: The figure shows genes (triangles) and regulators (circles for complexes and squares for individual proteins) that our algorithm predicted in the upper right cluster of Figure 7-2. Edge color and width indicates the strength of the equilibrium constant for each binding.
Figure 7-4: Complexes predicted to bind in a selection of the 12 conditions profiled in (Harbison et al., 2004). Each node in the network is a protein, edges indicate complex formation. Edge color and width indicates the number of genes where the joint binding of the two proteins was best explained by a single chemical reaction involving the two proteins. That is total amount of binding for each complex across sites. (d-f) Comparison between conditions. In (d) blue edges indicate increased complex bindings in $H_2O_2$Hi over YPD, red indicates decreased complex bindings, similarly in (e,f). Note that comparative networks compare binding summed across sites; complex reallocations between sites, that do not change the overall amount of binding, are not apparent in this representation.
Figure 7-5: Ste12-Tec1 bindings across sites for each condition. Note that percentages shown refer to bindings across all sites, for a site-detail refer to figure 7-6

Figure 7-6: Ste12-Tec1 bindings at each site where either protein or the complex were predicted to bind. Sites are sorted by the value of the equilibrium constant of the complex
Figure 7-7: Yap6-Cin5 bindings across sites for each condition. Note that percentages shown refer to bindings across all sites, for a site-detail refer to figure 7-8. The low amount of binding of protein and complex is a byproduct of the large selection of sites where either one is found to bind.

Figure 7-8: Yap6-Cin5 bindings at each site where either protein or the complex were predicted to bind. Sites are sorted by the value of the equilibrium constant of the complex.
Figure 7-9: Skn7-Cin5 bindings across sites for each condition. Note that percentages shown refer to bindings across all sites, for a site-detail refer to figure 7-10. The low amount of binding of protein and complex is a byproduct of the large selection of sites where either one is found to bind.

Figure 7-10: Skn7-Cin5 bindings at each site where either protein or the complex were predicted to bind. Sites are sorted by the value of the equilibrium constant of the complex.
Part III

Conclusions
Chapter 8

Extensions

We have used the theory of abstract economies to present a new abstraction for DNA binding that capitalizes on the interactions local to each agent (i.e., proteins and sites) to predict average states of regulation for the DNA as a whole system. To abstract DNA binding as an economy we focused on features such as resource allocation and global constraints on resources, which DNA binding shares with abstract economies. We then introduced learning to solve the inverse problem of inferring the economic structure from aggregate observations of the regulatory state of DNA.

Broadly speaking, we used economic theory as a means to abstract a problem in terms of its local structure, and propagate that local structure globally using the concept of competitive equilibrium. In doing so we avoided having to define explicitly an expression for the aggregate behavior of the system, which would have required a considerable effort to define all the possible interactions. We presented algorithms to make forward predictions about the aggregate behavior and algorithms to learn the local structure from noisy observations of that aggregate behavior. The key features of the economic approach are its scalability, its ability to infer structure from limited observations of the emerging aggregate behavior and that it can make causal predictions to experimental perturbations.

Our results for DNA binding demonstrate that this approach can be used to successfully describe a problem for which an economic structure is not immediately apparent. Similar economic structures underly a broad array of problems in science
and engineering, for which we know the structure of local interactions but describing
the global behavior is otherwise challenging. Some examples of problems are signaling
and molecular control in biology, drug development, neural structures, distributed
control, recommender problems, social networking, as well as market dynamics.

Here we review some extensions to the DNA binding economies and potential
applications of the economic approach within biology and other fields.

8.1 Extensions Within Biology

We believe that this technology may offer substantial advantages in the pharmaceu-
tical and healthcare industries. For example, learning and understanding competing,
coordinating, and compensating roles of molecules in signaling cascades could have
a strong impact on drug development. On the other hand, by extracting economic
structures underlying side effects and interactions of drugs may lead to successful
therapies involving multiple drugs in different combinations.

8.1.1 Tissue Regulation and Connectivity Map

The current approach to learning an economic structure from binding assays can
be extended to use expression arrays as input. An immediate advantage to using
expression arrays is the broader availability of experimental data. Introducing new
assumptions about transcription, however, poses a challenge. These can be explicit, a
functional relationship between binding and expression, or introducing a new agent;
or implicit, adding a level of abstraction interpreting sites and proteins as latent
variables.

With expression arrays as input, the economic approach can be used to learn an
economic structure that explains different tissue regulatory profiles using data from
Novartis’ Gene Atlas (Su et al., 2004, 2002). Causal economic predictions can then be
used to help determine factors (or combination thereof) that are responsible for the
differences and commonalities across tissues, and to explore methods to induce tar-
geted regulatory changes in tissues. Similarly, functional measures can be introduced
to categorize tissues and regulatory programs, to complement previous probabilistic work in functional tissue identification (v., (Gerber et al., 2007)).

Helping complete the connectivity map of molecules, gene-expression levels, and diseases (v., Lamb et al. (2006)), is another potential application of the economic approach using gene-expression data. In this case, one may use the promoter specific economy as a latent model, associating proteins with latent factors responsible for activating certain regulatory processes and grouping regions of DNA by their joint expression patterns. The economic approach can then be used in drug-design, to make causal predictions about the effect of combining drugs, and design genetic-therapies that, for instance, minimize the number of side-effects.

8.1.2 Augmenting the Economies with Sequence Information

When sequence information is available, it might be desirable to connect sequence information explicitly with the notions of affinity and equilibrium constant. As noted earlier, one such connection may be a valuable complement to sequence-based similarity metrics, to quantify the chemical effect of base-pair deletions, additions, and permutations. Sequence information will also enable the economic approach to be used for motif detection, and has the interesting side effect of reducing the number of parameters required to describe the economic structure.

The following section sketches an approach to extend the DNA binding economies with sequence information

Motif Specific derivation

We want to unify affinities and binding parameters under a common representation. We represent protein specific binding motifs with a $4 \times k$ probability matrix

$$w^i(n, \ell) = \begin{bmatrix} p(e_0 = A) & p(e_1 = A) & \ldots & p(e_{k-1} = A) \\ p(e_0 = T) & p(e_1 = T) & \ldots & p(e_{k-1} = T) \\ p(e_0 = G) & p(e_1 = G) & \ldots & p(e_{k-1} = G) \\ p(e_0 = C) & p(e_1 = C) & \ldots & p(e_{k-1} = C) \end{bmatrix}$$ (8.1)
where \( e_k \) is a shorthand for \( k^{th} \) element (i.e., a nucleotide A,C,T, or G) of the motif, and the columns of the matrix add up to one. The matrix defines a probabilistic motif for protein \( i \). Let \( d_j \) be the \( k \)-nucleotide-long sequence of DNA that starts at position \( j \). Then we use \( w^i \) to define the probability of any element in that sequence being in the motif \( b \) for protein \( i \) as:

\[
P(d_j(\ell) \text{ is } \ell^{th} \text{ element of motif}) = w^i(d_j(\ell), \ell)
\]  

(8.2)

And the probability that a sequence \( d_j \) is a motif for protein \( i \) is taken as the product of these probabilities:

\[
P(d_j \text{ is motif}) = \prod_\ell P(d_j(\ell) \text{ is } \ell^{th} \text{ element of motif}) = \prod_\ell w^i(d_j(\ell), \ell)
\]  

(8.3)

We want to represent affinities and equilibrium constants as a function of \( w^i \) and \( d_j \), \( (E_{ij} = E(w^i, d_j), K_{ij} = K(w^i, d_j)) \). Where \( d_j \) stands for a \( k \)-nucleotide sequence starting at DNA position \( j \), and we identify it with site \( j \)_1. We postulate that the binding energy of a protein-site pair, denoted \( H_{ij} \), is proportional to the log-ratio of the probability that the site is a motif to a given background probability \( b \),

\[
H_{ij} \approx \sum_{\ell=j}^{j+k-1} \log \frac{w^i(d_j(\ell), \ell-j+1)}{b(d_j(\ell))}
\]  

(8.4)

With sites indexed by the starting nucleotide, \( p^i_j \) is the protein \( i \) allocated near motif candidate \( d_j \), and \( s^i_j \) is the expected binding of protein \( i \) at candidate motif \( d_j \). The utility functions are slightly modified to account explicitly for neighborhoods of sites in the following manner. The protein utility (transport mechanism) becomes

\[
u(p^i_j) = \sum_j p^i_j E_{ij} \left( 1 - \sum_{\ell \in N(j)} \alpha^\ell \right) + \lambda \sum_j p^i_j \log \frac{1}{p^i_j},
\]  

(8.5)

where \( E_{ij} \) is a measure of energy proportional to \( H_{ij} \) (i.e., we assume that motif iden-

\footnote{Note that sites \( j \) and \( j+1 \) overlap in all but two nucleotides.}
tification is highly specific, and so energetic terms due to repulsions and attractions in the surroundings are negligible); \( N(j) \) defines the surroundings of site \( j \); and as usual \( \alpha^j = \sum_i s^j_i \). We note that we have imposed that

\[
\sum_j p^j_i = 1. \tag{8.6}
\]

This states that we do not account for any remaining protein in the nucleus/cell that is not involved in the DNA-binding economy. Hence, \( f^i \) is just the amount of protein that is close to DNA (either near sites or bound to them.)

At the site, the chemical equilibrium equation satisfies,

\[
\frac{s^j_i}{(f^i p^j_i - s^j_i)(1 - \alpha^j)} = K_{ij} = e^{\beta H_{ij}+C}, \tag{8.7}
\]

where \( \tilde{p}^j_i = \sum_{\ell \in N(j)} p^\ell_i \) is the total availability of protein \( i \) in the neighborhood of \( j \), and \( \beta \) and \( C \) are two parameters to capture the relationship of the equilibrium constant and the energy.

In addition, we need to specify a new binding constrain. Suppose a protein is bound to a \( k \)-nucleotide-long-sequence starting at \( j \), then sequences starting anywhere from \( j - k + 1 \) to \( j + k - 1 \) have to remain unbound. We can write the constrain in terms of the site specific neighborhood as:

\[
\sum_{\ell \in N(j)} \sum_i s^\ell_i \leq 1 \quad \forall \text{ site } j. \tag{8.8}
\]

It is also possible to introduce this constraint directly into the chemical equilibrium equation using balance of mass.

### 8.1.3 Theoretical Extensions

For the nucleus specific economy we showed that an equilibrium always exists and is unique. Equilibrium uniqueness, however, remains open for the site and promoter specific economies. Empirically the equilibria attained by our site and promoter spe-
specific algorithms appear to be unique, at least for biologically meaningful values of the economic parameters. Proving uniqueness formally, however, remains a challenge. This is also, in general, the case for supermodular games. Previous work in supermodular games has only been able to assert uniqueness for problems that can be shown to exhibit diagonal dominance; as is the case for the nucleus specific economy.

Equilibrium uniqueness is a desirable property because it simplifies the learning task. A possible extension to this work is to study the site and nucleus specific economy capitalizing on the biological intuition, to try to prove uniqueness formally; and hopefully, derive more general results for supermodular games.

8.1.4 Experimental Extensions

Throughout this thesis we have emphasized scalability and timeliness of predictions as two key features of the economic approach to empower experiment design. A potential avenue for future cross-disciplinary research is to steer some of the previously cited extensions of this work jointly with an experimental laboratory to take advantage of new measuring techniques in molecular biology and the application to medicine.

8.2 Extensions to Other Fields

The modelling framework developed in this thesis can be readily applied to problems already studied as an economy. These include problems in market dynamics, portfolio theory, etc. The learning paradigms developed here are likely to be relevant there to bridge the gap between the theoretical economic models and the vasts amounts of empirical data available about market trends. We do not discuss these cases further. Instead we focus on two examples in telecommunications and recommender systems, for which the economic abstraction may also prove valuable to simplify the problem.

A common problem in wireless telecommunication networks is to adapt power allocation as a function of call volume. There are already economic models that successfully derive power allocation as a function of demand for single antennas.
However, scaling these models up to an entire wireless network is challenging because demand fluctuates significantly in time and is location-dependent. Existing network optimization algorithms, however, are generally not designed to incorporate location-dependent signal strength information even if that information can already be obtained from the network. They require instead extensive field testing and costly calculations to take into account the changing geometry of the environment. The economic approach presented in this thesis can be extended considering users as consumption units and antennas as production units, information about location, signal attenuation, can be incorporated in the utility functions to determine the best power distribution for each antenna to minimize dropped calls, and ensure that the wireless space is covered under global power constraints. A learning approach following the guidelines introduced in this thesis for DNA binding would enable faster response to changes in the actual aggregate power demand being measured.

Recommender systems are widely used to facilitate the access of customers to products they are likely to buy (movies, music, books, and so on). The decisions concerning what to present to a customer are based on the collective behavior of other customers in addition to the previous actions of the customer in question. The wealth of information provided by other customers is typically summarized and filtered linearly to obtain predictions. Existing methods, however, do not possess a mechanism to understand how customers arrive at their decisions, i.e., they miss the underlying economic structure. Customers have diverse and competing interests at varying levels, their interests are tied to types and collections of products, and there is considerable (typically externally driven) coordination among customers as in selecting popular movies. The main thrust of the work presented in this thesis is precisely to recover and exploit this economic structure in a manner that scales to hundreds of thousands of users and products.
Chapter 9

Discussion

This thesis has introduced three abstract economies that describe DNA binding at different scales (site, promoter, nucleus) and explore the trade-off between purely competitive processes and processes that explore explicit coordination. First, we introduced a mathematical abstraction for DNA binding, then presented results that demonstrate the predictive ability of the economic approach alone, and, finally, enhanced with learning algorithms.

9.1 Mathematical Abstraction

The economic abstraction of DNA binding presents sites as production units, regulators \(i.e.,\) proteins and protein complexes) as consumption units, and associates the average arrangement of proteins along DNA with the competitive equilibrium. The DNA binding economies predict average binding arrangements as regulatory states. Transcription dynamics can be introduced as a time dependence in protein availability, leading to a dynamic representation of binding as a succession of regulatory states.

For each economy we have presented utility functions that encode the subproblems of binding and transport; a fixed-point equation that can be interpreted as balancing supply and demand; proofs of existence of competitive equilibria in pure strategies under any setting of the economic parameters (existence of equilibria in the DNA
binding economies does not derive immediately from previous work); and an algorithm to compute a competitive equilibrium for each economy. In the case of the nucleus specific economy, we have presented additional mathematical results that prove that the equilibrium is unique.

The learning algorithms introduced in Chapter 5 bridge the gap from the mathematical abstraction to high-throughput observations. We introduced optimization algorithms to infer the local structure of the economy from noisy observations of binding alone. In the motivation of our learning approach we emphasized that our algorithms learn structure from noisy and narrow slices of data. Microarray binding assays illustrate what we mean by narrow slices of data. While the problem of DNA binding has a dynamic structure due to the mechanisms of transport and binding, Microarray assays only provide observations about a final binding state.

9.2 Experimental Results

We validated the economic approach in two real cases, λ-phage (small scale), and yeast regulation (large scale).

9.2.1 Small Scale Results

We used a small scale illustration (λ-phage) to present regulation as a succession of regulatory states. We showed that our results reproduce the known order of events and agree with stochastic simulation results on the protein levels at which events occur in λ-phage.

We compared small-scale results on the site and promoter specific economies to illustrate the ability of the economic approach to capture effects of scale via the transport mechanism. The comparison illustrates how to interpret the parameters defining the economic structure. In the site-specific economy, affinities can encode proximity and help capture coordinate behavior implicitly as a byproduct of transport and competition. In the promoter-specific economy, affinities capture interdependence between DNA subsystems mediated by protein availability: propagating
limiting reagent constraints, and reallocating excess reagent.

We also used λ-phage to illustrate how our models can enhance the trial and error process in experiment design. We introduce two predictive examples: causal predictions following perturbations, (simulating knockout experiments); and, learning from a small batch of experiments in a hypothetical experimental scenario to determine the range of experimental conditions for the next assays. The fast convergence of our algorithms, (which we illustrate in the experimental section), enables economic models to be used to test hypotheses efficiently in a computer when planning the next batch of experiments. New experimental observations can then be incorporated into the economic structure using the learning algorithms from Chapter 5.

Using computational methods in experiment design, to predict the effect of experimental perturbations is not new. Scalability challenges, however, have prevented mechanistic methods from being adopted as an integral part of experiment design in the study of genetic regulation. The economic approach introduced in this thesis addresses the scalability challenges and enables timely predictions about average regulatory behavior; even when the underlying biological mechanisms are only partially known.

We emphasize that the algorithms presented in this thesis do not rely on the existence of a prior representation of the underlying biological process. However, they can be combined with prior knowledge about the underlying biology when available. For instance, the learning algorithms can combine empirical observations with fixed equilibrium constants set in accordance with prior knowledge.

9.2.2 Large Scale Results

We demonstrated the scalability of the economic approach learning an economic structure for yeast regulation from high-throughput observations from Harbison et al. (2004). Our results show that the economic approach can be used to complete the regulation map in yeast.

Validating large scale results is challenging because our knowledge of the underlying biological processes is partial. For instance, while our algorithms can be used to
trace the regulatory changes induced by smooth changes in the experimental conditions, there is no previous work describing such changes to which we can refer. Thus, an additional challenge is to identify measures to compare large-scale results to prior knowledge about regulation in Yeast. We show how to overcome this challenge with three new kinds of predictions based on the structure of the nucleus specific economy: functional similarities based on reactive profiles, network representation of the economic structure, and predictions about coordinate behavior.

**Reactive Profiles**

Our results using reactive profiles to cluster genes and regulators demonstrate the use of the economic structure as a functional similarity measure. Comparing the reactive profile of two genes is equivalent to comparing the set of binding reactions that take place up or downstream of that gene as part of a regulatory process. The reactive profile of a gene is a measure of how the gene is regulated, and similarly for regulators. Unlike sequence-based similarity measures, however, reactive profiles measure chemical activity, so they can be combined with sequence similarity measures to quantify the relevance of base-pair additions, permutations, deletions etc.

Specifically, we showed that the reactive profiles to cluster genes and regulators successfully grouped genes and regulators in agreement with their reported biological function. Most notably, our results show that the economic structure learned from high-throughput observations also assigns clusters of regulators to clusters of genes in agreement with what is known about their function. For instance, we showed that genes primarily involved in protein-synthesis are mostly bound by a set of proteins known to have function in RNA-synthesis. Our results would imply that genes in the protein-synthesis group are mainly defined by their chemical affinity for proteins and complexes known to have a role in RNA-synthesis, that is the first step in the flow of sequence information from DNA to protein.
Network Representation

The economic structure can be represented as a network with edges indicating the strength of the equilibrium constant of the binding reaction between a site and a protein or complex. We show one such network representation for one of the clusters detected in the extended set of genes and proteins.

The network representation is causal in that it incorporates information about the chemical mechanisms of binding. It can be seen as a coarse representation of a genetic circuit, in which kinetic information has been removed and only information about equilibrium dynamics is present.

Cooperative Behavior

Finally, we present a method to identify cases of cooperative behavior. We illustrated the method showing that it predicts a known complex \textit{STE12-TEC1} and a known case of coordinate binding \textit{YAP6-CIN5}. We also showed that our method can be used to detect previously unreported cases of cooperative behavior. Specifically, we showed that in some sites the action of \textit{SKN7} and \textit{CIN5} is best explained by a single chemical reaction involving both proteins than by two independent chemical reactions.

Predictions about cooperative behavior are particularly relevant because, generally, the design of Microarray binding assays does not differentiate between a protein binding a site alone or jointly with other proteins. Our methods can resolve between competitive and cooperative behavior comparing the behavior of the same protein across different conditions.

9.3 Advantages of the Economic Abstraction

Game-theoretic and economic concepts reside at a level of abstraction that eases the interpretation of biological concepts in computational terms. Approaches that require constructing a global energy function to minimize tend to rest at lower levels of abstraction, which make the computational task harder. For large-scale problems constructing a global function requires introducing new assumptions about how DNA
subsystems interact and the dynamics of transport. The concept of resource allocation under global constraints seems more adequate to express these interaction terms, and can be naturally expressed using the theory of abstract economies.

Modeling the transport mechanism as a resource allocation problem allows us to model the interdependence between DNA subsystems explicitly as a consequence of shifts in local availability of sites and proteins. This interdependence has been observed and can only be explained using the concept of chemical equilibrium in combination with the notions of limiting reagent and reallocation of excess reagent. From a chemical standpoint, the transport mechanism provides a way to define the inhomogeneities in the chemical mix of the nucleosol that arise from the high specificity of binding, dynamic changes in site availability, and conformational changes in DNA. Understanding how proteins are distributed to the well mixed regions of the nucleosol where binding takes place can only be accomplished otherwise by tracing the diffusion dynamics of transport.

The economic formalism provides a conceptual connection with statistical mechanics that is convenient to reason about chemical systems. Like chemical equilibrium, the concept of competitive equilibrium is a dynamic concept that results when the exchange of commodities has reached a steady state (i.e., supply=demand). The competitive equilibrium summarizes the average distribution of commodities when all the agents in the economy have no incentive to change their average allocation of resources. Similarly, at chemical equilibrium the average concentration of species remains constant because the reaction has reached a steady state and happens at the same rate in both directions. In either economy or chemistry, the choice to study population kinetics over equilibrium dynamics is intimately connected with the purpose of the analysis. For instance, in chemistry, equilibrium dynamics are convenient to study the evolution of the system at large time-scales as a function of changes in the concentration of some species.

The economic formalism emphasizes the local structure of the problem by making explicit the preferences of each agent in their utility functions. The concept of competitive equilibrium helps propagate the constraints imposed by each agent globally.
This abstraction suits the notion that proteins are affected by their local environment, rather than controlled by an unobserved mechanism to remain at a global energy minimum.

Taken together, transport mechanism and the notion of competitive equilibrium provide the economic abstraction with greater flexibility to model protein-DNA binding at different scales. For instance, using model selection, to determine the scale at which to model each region of DNA (choosing over site, promoter and nucleus specific economies locally), or viewing the problem of protein-DNA binding as involving aspects of all the three economies.

9.3.1 Generalization to Other Problems

Our abstraction of game theory and economic theory as inference tools has broad applicability in machine learning. Economic structures as introduced here, underlie a surprisingly broad array of problems including signaling and molecular control in biology, drug development, neural structures, distributed control, recommender problems, social networking, as well as market dynamics.

The problems listed above have several features in common. They are dynamic; generally, only aggregate observations of the state of the system are available; and, while expressing the problem via a function that describes the aggregate is challenging, understanding the local interactions between its parts is considerably simpler.

We believe that it is possible to extend the analysis done in this thesis to problems in the above domains. That is define the commodities, production units and consumption units that explain what’s known about local interactions, and propagate these globally using the concept of competitive equilibrium to balance supply and demand.

9.4 Summary of Contributions

This thesis has key explicit contributions to computational biology. The application to computational biology illustrates a new modelling framework with broad applicability
in machine learning. We overview the contributions to computational biology and the key distinctive features of the new modelling framework.

**Computational biology.** We develop a new abstraction of genetic regulation to infer causal regulatory structure from narrow observations of binding arrangements, and predict regulatory states genome-wide in response to experimental perturbations and as a function of protein availability.

- We define three classes of economies that differ by the scope of the transport mechanism, and explore the trade-off between purely competitive processes and processes that involve explicit coordination. We prove that an equilibrium in pure strategies exists for any setting of the parameters of the economies.

- We map each economy to a fixed point equation and prove monotonicity properties. We also indicate a connection with the theory of supermodular games.

- We develop fast iterative algorithms based on iterated dominance to predict regulatory states genome-wide. Each iteration of our algorithms is polynomial in the number of sites and proteins. Convergence rate of the iterative process depends, however, on the values of the parameters and the specificity of the transport mechanism. The algorithms provided in this thesis can be seen as instances of constraint propagation or message passing algorithms.

- We develop learning algorithms to infer the structure of regulation from high-throughput binding observations.

- Our approach enables three new key predictions:
  
  - predictions to guide experiment design in an effective *trial and error* manner,
  
  - functional identification of complexes and joint binding events,
  
  - a network representation of the predicted binding structure, and
  
  - new measures of similarity to discover gene and protein function based on the predicted binding structure.
• we illustrate the uses of the economic approach in a small scale example and a genome-wide example.

**Machine learning.** The economic models developed in this thesis illustrate a new modelling paradigm. Specifically, we show by example how to map a problem onto an economy and learn an economic structure from data. The economic structure learned can then be used to make causally meaningful predictions in response to perturbations. This use of game theory and abstract economies is new to machine learning; the economic theory of revealed preferences is the closest theoretical analog to our approach. We enumerate the distinctive features of the new framework

• We associate observations with equilibrium and try to solve the inverse problem of determining the parameters of the utility functions that gave rise to such observations. Our learning approach assumes rationality of the agents. We say that we learn an economic structure owing to the formalism of graphical games and economies.

• The new framework is scalable, and suitable for inference from limited observations. While limited in depth, observations may come from high-throughput experiments, our framework assumes that observations may be partial and noisy. The type of structure learned is, however, restricted by the mapping of the problem onto an economy. Specifically, it is restricted by the choice of utility functions, which encode prior knowledge about the problem.

• The framework is a new application of economic theory in an experimental setting. It is specifically suited to model problems where the global state of the system is not easily described with global considerations about energy or efficiency, but can be easily cast as emerging from the structure of local interactions.
Appendix A

Appendix: Equilibrium existence proofs

A.1 Brouwer’s Fixed Point Proof of Existence of a Site Specific Competitive Equilibrium

We start by recalling theorem 1

**Theorem.** 1 Every site based DNA binding economy has an equilibrium.

Parts of this proof follow closely the arguments in Nash (1951). To simplify the proof, we introduce some additional notation. We denote by $\alpha$ the column vector whose $j$th component is $\alpha_j$. For all $i$ and $j$, let

$$g_j^i(p^i, \alpha) \equiv E_{ij}(1 - \alpha_j) - \sum_j p^j E_{ij}(1 - \alpha_j).$$

We now view $\alpha$ as a variable and consider the mapping defined as follows: for all $i$ and $j$,
\[ p_j^i \leftarrow \frac{p_j^i + \max(g_j^i(p^i, \alpha), 0)}{1 + \sum_j \max(g_j^i(p^i, \alpha), 0)} \]

\[ \alpha_j \leftarrow \left(1 + \frac{1}{\sum_i \frac{K_{ij}}{1 + K_{ij}(1 - \alpha_j)p_j^i f_i}} \right)^{-1} \]

where \( \alpha_j \in [0, 1] \), \( p_j^i \geq 0 \) and \( \sum_j p_j^i = 1 \). Because this is a mapping from a compact set to itself and is continuous, we can apply the Brouwer fixed point theorem to show the existence of a fixed point.

We claim that any assignment to \( p \) and \( \alpha \) is a fixed point of the mapping above if and only if \((p, s)\) is an equilibrium of the site structured DNA binding economy, with \( s \) such that

\[ s^j_i = \frac{K_{ij}(1 - \alpha_j)}{1 + K_{ij}(1 - \alpha_j)} p_j^i f_i. \]  \hspace{1cm} (A.1)

Equation (A.1) follows from the chemical equilibrium given in (3.4), after some algebra and using the definition of \( \alpha_j \).

It is easy to verify that any equilibrium of the game corresponds to a fixed point of the mapping. Consider an equilibrium \((p, s)\) of the game. Note that it must satisfy (3.4) by construction, and therefore also (A.1). That the \( \alpha \) part of the mapping holds follows from summing both sides of (A.1) over \( i \). To see that the equilibrium satisfies \( g_j^i(p^i, \alpha) \leq 0 \), note that the mapping is a restatement of the utility function, if it weren’t satisfied \( p^i \) could not be optimal but that would contradict the assumption that \((p, s)\) is an equilibrium of the game. This implies that the \( p \) part of the mapping is also satisfied. Thus, \((p, s)\) is also a fixed point of the mapping. The same argument holds for every fixed point of the game.

To show the converse, assume that an assignment to \( p \) and \( \alpha \) is a fixed point of the mapping. We can obtain \( s \) from \( \alpha \) by using (3.10), which implies that \( s^j_i \) is optimal given \( p_j \). For the \( p \) part, we use Nash’s original argument. Fix \( i \) and consider any site \( j' \in \arg \min_j E_{ij}(1 - \alpha_j) \). Then, we have \( \max(g_{j'}^i(p^i, \alpha), 0) = 0 \), and since \( s^j_{i'} \) is a part of a fixed point, \( g_{j'}^i(p^i, \alpha) \leq 0 \) must hold for all \( j \). This implies that \( p^i \) is optimal given \( s \). The same argument works for every \( i \). Thus, \((p, s)\) is an equilibrium of the
A.2 Proof of Theorem 8, Promoter Specific Economy

We recall theorem 6

**Theorem 6** Every promoter based DNA binding economy has an equilibrium.

We recall that the existence of a competitive equilibrium (i.e. theorem 6) follows from lemma 7 and theorem 8.

**Proof of Theorem 8**

Recall the definition of $G$ and the theorem statement

$$G_j^r(\alpha, a^{-r}) = \sum_{i \in P} \frac{K_{ij}(1 - \alpha^j)}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k)} f_i p_i^r(a)$$

so that, conditionally on $a^{-r}$, we aim to find a fixed point $\alpha^j = G_j^r(\alpha, a^{-r})$ for $j \in N(r)$.

**Theorem 8** Let $\{\hat{\alpha}^1\}$ be the fixed point solution $\alpha^j = G_j^r(\alpha, a_1^{-r})$ and $\{\hat{\alpha}^2\}$ the solution to $\alpha^j = G_j^r(\alpha, a_2^{-r})$. If $a_1^l \leq a_2^l$ for all $l \neq t$ then $\hat{\alpha}_1^r \leq \hat{\alpha}_2^r$.

**Proof:** Set $a^l(\gamma) = a_1^l + \gamma(a_2^l - a_1^l)$ for $l \neq t$. Let $\alpha^j(\gamma)$ be the fixed point $\alpha^j(\gamma) = G_j^r(\alpha(\gamma), a^{-r}(\gamma))$ as a function of $\gamma$. It suffices to show that

$$\frac{d}{dr} \sum_{j \in N(r)} \alpha^j(\gamma) = \sum_{j \in N(r)} d_j(\gamma) \geq 0$$
To this end we differentiate the fixed point

\[
d_j(\gamma) = \sum_{i \in P} \left[ \frac{K_{ij}(-d_j(\gamma))}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k(\gamma))} \right] f^i p^i_r(\alpha(\gamma)) \\
+ \sum_{i \in P} \left[ \frac{K_{ij}(1 - \alpha^j(\gamma))}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k(\gamma))} \cdot \frac{\sum_{k \in N(r)} K_{ik} d_k(\gamma)}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k(\gamma))} \right] f^i p^i_r(\alpha(\gamma)) \\
+ \sum_{i \in P} \left[ \frac{K_{ij}(1 - \alpha^j(\gamma))}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k(\gamma))} \right] f^i \frac{d}{d\gamma} p^i_r(\alpha(\gamma))
\]

By defining

\[
\tilde{d}_j(\gamma) = \frac{d_j(\gamma)}{1 - \alpha^j} \\
w_{jk} = \sum_{i \in P} s_i^j(\gamma) \frac{K_{ik}(1 - \alpha^k(\gamma))}{1 + \sum_{k \in N(r)} K_{ik}(1 - \alpha^k(\gamma))},
\]

and noting that \( \sum_{i \in P} s_i^j(\gamma) = \alpha^j(\gamma) \) as well as expanding \( d/d\gamma \log p^i_r(\alpha(\gamma)) \) we get

\[
(1 - \alpha^j(\gamma))\tilde{d}_j(\gamma) = -\alpha^j(\gamma)\tilde{d}_j(\gamma) + \sum_{k \in N(r)} w_{jk} \tilde{d}_k(\gamma) \\
+ \sum_{i \in P} s_i^j(\gamma) \left[ \frac{d}{da^i} \log p^i_r(a) \sum_{k \in N(r)} d_k(\gamma) + \sum_{l \neq i} \frac{d}{da^l} \log p^i_r(a)(a^2_i - a^1_i) \right]
\]

\[
\geq -\alpha^j(\gamma)\tilde{d}_j(\gamma) + \sum_{k \in N(r)} w_{jk} \tilde{d}_k(\gamma) + \left[ \sum_{i \in P} s_i^j(\gamma) \frac{d}{da^i} \log p^i_r(a) \right] \sum_{k \in N(r)} d_k(\gamma)
\]

\[
= -\alpha^j(\gamma)\tilde{d}_j(\gamma) + \sum_{k \in N(r)} w_{jk} \tilde{d}_k(\gamma) - v_j \sum_{k \in N(r)} d_k(\gamma)
\]

where the inequality corresponds to

\[
\sum_{l \neq i} \frac{d}{da^l} \log p^i_r(a)(a^2_i - a^1_i) \geq 0
\]

168
which follows from lemma 7 and our assumptions. Similarly

\[ v_j = - \sum_{i \in P} s_i^j(\gamma) \left[ \frac{d}{da'} \log p'_i(a) \right] \geq 0 \]

Note also that \( w_{jk} \geq 0 \) and \( \sum_{k \in N(r)} w_{jk} < 1 \) for all \( j \in N(r) \). By rearranging terms we finally get

\[ v_j \sum_{k \in N(r)} d_k(\gamma) \geq -\tilde{d}_j(\gamma) + \sum_{k \in N(r)} w_{jk} \tilde{d}_k(\gamma), \quad j \in N(r) \]

We can now set up a linear program to minimize \( \sum_{k \in N(r)} d_k(\gamma) \) with respect to \( d_k(\gamma) \)'s subject to the above constraints. By rewriting the objective entirely in terms of \( \tilde{d}_k(\gamma) \)'s we find

\[
\min \sum_{k \in N(r)} (1 - \alpha^k(\gamma))\tilde{d}_k(\gamma) \quad \text{subject to} \\
v_j \sum_{k \in N(r)} (1 - \alpha^k(\gamma))\tilde{d}_k(\gamma) \geq -\tilde{d}_j(\gamma) + \sum_{k \in N(r)} w_{jk} \tilde{d}_k(\gamma), \quad j \in N(r) \]

The linear program has a feasible solution in the interior since \( \sum_{k \in N(r)} w_{jk} < 1 \) (consider constant positive values for \( \tilde{d}_k(\gamma) \)'s). By the Slater’s conditions there is no duality gap. The dual LP objective to maximize is

\[
L^*(\lambda) = \begin{cases} 
0, & \text{if } (1 - \alpha^k(\gamma))(1 - \sum_{j \in N(r)} \lambda_j v_j) - \lambda_k + \sum_{j \in N(r)} \lambda_j w_{jk} = 0, \quad \forall k \in N(r) \\
-\infty & \text{otherwise}
\end{cases}
\]

subject to \( \lambda_k \geq 0 \). We will construct a feasible solution with value 0 as follows. Let \( W \) be the matrix with elements \( w_{kj} \), \( \lambda \) the vector of \( \lambda_k \)'s, and \( 1 - \alpha \) the vector with components \( 1 - \alpha^k(\gamma) \). In the new notation, the value of the dual objective is zero whenever

\[
(1 - \alpha)^T - (1 - \alpha)^T (\lambda^T v) - \lambda^T (I - W) = 0
\]
Define $\hat{\lambda}$ as

$$\hat{\lambda}^T = (1 - \alpha)^T(I + W + W^2 + \ldots).$$

This is a vector with non-negative elements and is a solution to $\hat{\lambda}^T(I - W) = (1 - \alpha)^T$. The power series converges because $W$ is strictly sub-stochastic. The solution to the dual is now obtained simply by rescaling: $\lambda^*_k = \hat{\lambda}_k/(1 + \hat{\lambda}^Tv)$. The dual maximum – and the primal minimum – is therefore zero. □
Appendix B

Appendix: Learning Derivation

B.1 Site based Learning Algorithm

We recall the optimization problem:

\[
F(\hat{o}_{tij}, s_{tij}, K_{ij}, E_{ij}, f_i) = \sum_{(t,i,j) \in T} L(\hat{o}_{tij}, s_{tij}) + \lambda \sum_{ij} K_{ij} + \lambda \sum_{ij} E_{ij} \tag{B.1}
\]

subject to

\[
s_{tij} = \frac{K_{ij} \beta_{tij}}{1 + K_{ij} \beta_{tij}} f_{ti} P_{tij} \tag{B.2}
\]

\[
P_{tij} = \frac{e^{E_{ij} \beta_{tij}}}{\sum_l e^{E_{il} \beta_{tl}}} \tag{B.3}
\]

\[
\beta_{tij} = 1 - \sum_i s_{tij} K_{ij} \geq 0 \tag{B.4}
\]

\[
E_{ij} \geq 0 \tag{B.5}
\]

\[
f_i \geq 0 \tag{B.6}
\]

We now show how to simplify the optimization problem to reach an unconstrained expression for the gradient. It is useful to think of the procedure to solve this optimization as tracing a path in the \((s, \beta, K, E, f)\) space. That is, at any given point \((s^o, \beta^o, K^o, E^o, f^o)\) the gradient \((\dot{s}, \dot{\beta}, \dot{K}, \dot{E}, \dot{f})\) defines the direction of maxi-
mum growth in the cost function as a function of some affine parametrization of the optimization path. Dotted variables represent derivatives with respect to the path parameter. Since we are not interested in this parameter we may proceed without naming it.

Taking the derivative of the cost functional (B.1) with respect to the path parameter yields an expression for its gradient

$$\dot{F}(\dot{o}_{ij}, s_{ij}, K_{ij}, E_{ij}, f_{ti}) = \sum_{(t,i,j) \in T} \frac{\partial L(\dot{o}_{ij}, s_{ij})}{\partial s_{ij}} \dot{s}_{ij} + \lambda \sum_{ij} \dot{K}_{ij} + \lambda \sum_{ij} \dot{E}_{ij} \tag{B.7}$$

Similarly, differentiating the constrains imposed by the game-equilibrium algorithm we can rewrite the constrains to (B.7) in terms of the basic variables of the problem. Differentiating the first constrain yields

$$\dot{s}_{ij} = \left[ \frac{K_{ij} \dot{\beta}_{ij}}{1 + K_{ij} \beta_{ij}} \right] f_{ti} P_{tij} + \frac{K_{ij} \dot{\beta}_{ij}}{1 + K_{ij} \beta_{ij}} \dot{f}_{ti} P_{tij} + \frac{K_{ij} \dot{\beta}_{ij}}{1 + K_{ij} \beta_{ij}} f_{ti} \dot{P}_{tij} \tag{B.8}$$

The first term in (B.8) can be further simplified into

$$\left[ \frac{K_{ij} \dot{\beta}_{ij}}{1 + K_{ij} \beta_{ij}} \right] = \frac{\dot{K}_{ij} \beta_{ij} + K_{ij} \dot{\beta}_{ij}}{[1 + K_{ij} \beta_{ij}]^2} = \frac{K_{ij} \dot{\beta}_{ij}}{1 + K_{ij} \beta_{ij}} \left[ \frac{\dot{K}_{ij}}{K_{ij}} + \frac{\dot{\beta}_{ij}}{\beta_{ij}} \right] \frac{1}{1 + K_{ij} \beta_{ij}}$$

the second term in (B.8) is already expressed in terms of one of the variables of the problem $\dot{f}$. It only remains to simplify the third term which involves differentiating
replacing these two terms back in (B.8) we reach the final expression for the first constrain.

\[
\dot{s}_{tij} = s_{tij} \left[ \left( \frac{\dot{K}_{ij} + \dot{\beta}_{tj}}{K_{ij} + \beta_{tj}} \right) \frac{1}{1 + K_{ij} \beta_{tj}} + \frac{\dot{f}_{ti}}{f_{ti}} + \sum_{l} \left( \delta_{(j,l)} - P_{tli} \right) E_{il} \beta_{tl} \left( \frac{\dot{E}_{il}}{E_{il}} + \frac{\dot{\beta}_{tl}}{\beta_{tl}} \right) \right] 
\]

We rewrite here the two constrains in their final form

\[
\dot{s}_{tij} = s_{tij} \left[ \left( \frac{\dot{K}_{ij} + \dot{\beta}_{tj}}{K_{ij} + \beta_{tj}} \right) \frac{1}{1 + K_{ij} \beta_{tj}} + \frac{\dot{f}_{ti}}{f_{ti}} + \sum_{l} \left( \delta_{(j,l)} - P_{tli} \right) E_{il} \beta_{tl} \left( \frac{\dot{E}_{il}}{E_{il}} + \frac{\dot{\beta}_{tl}}{\beta_{tl}} \right) \right] 
\]

We define the following changes of variable

\[
\gamma_{tj} \triangleq \log \beta_{tj} \quad r_{tij} \triangleq \log s_{tij} \quad k_{ij} \triangleq \log K_{ij} \quad e_{ij} \triangleq \log E_{ij} \quad g_{ti} \triangleq \log f_{ti} \quad (B.11)
\]
and corresponding path derivatives

\[ \dot{\gamma}_{ij} = \frac{\dot{\nu}_{ij}}{\nu_{ij}} \quad \dot{r}_{ij} = \frac{\dot{s}_{ij}}{s_{ij}} \quad \dot{k}_{ij} = \frac{K_{ij}}{K_{ij}} \quad \dot{e}_{ij} = \frac{E_{ij}}{E_{ij}} \quad \dot{g}_{ti} = \frac{\dot{f}_{ti}}{f_{ti}} \]  

(B.12)

Rewriting the expression for the gradient (B.7), and constrains (B.10) in terms of these variables,

\[
\dot{F}(\hat{o}_{tij}, \dot{k}_{ij}, \dot{e}_{ij}, \dot{g}_{ti}, \dot{\gamma}_{ij}) = \left[ \sum_{(t,i,j) \in \mathcal{T}} \frac{\partial L(\hat{o}_{tij}, s_{tij})}{\partial s_{tij}} \right] s_{tij} \dot{r}_{tij} \\
+ \lambda \sum_{ij} K_{ij} \dot{k}_{ij} + \lambda \sum_{ij} E_{ij} \dot{e}_{ij}
\]  

(B.13)

subject to

\[
\dot{r}_{tij} = \frac{\dot{k}_{ij} \dot{\gamma}_{ij}}{1 + K_{ij} \beta_{ij}} + \dot{g}_{ti} + \sum_{l} (\delta_{(j,l)} - P_{til}) E_{il} \beta_{ul} (\dot{e}_{il} + \dot{\gamma}_{ul}) \\
- \beta_{ij} \gamma_{ij} = \sum_{i} s_{tij} \dot{r}_{tij}
\]  

(B.14)

To incorporate the first constrain into the gradient we define \( L_{tij} = \left[ \frac{\partial L(\hat{o}_{tij}, s_{tij})}{\partial s_{tij}} \right] s_{tij} \) and replace \( \dot{r}_{tij} \) in the first term of (B.13).

\[
L_{tij} \dot{r}_{tij} = L_{tij} \frac{\dot{k}_{ij}}{1 + K_{ij} \beta_{ij}} + L_{tij} \frac{\dot{\gamma}_{ij}}{1 + K_{ij} \beta_{ij}} + L_{tij} \dot{g}_{ti} \\
+ L_{tij} \sum_{l} (\delta_{(j,l)} - P_{til}) E_{il} \beta_{ul} \dot{e}_{il} + L_{tij} \sum_{l} (\delta_{(j,l)} - P_{til}) E_{il} \beta_{ul} \dot{\gamma}_{ul}
\]  

(B.15)

With some algebra we reach an expression for the gradient as a linear combination of \((\dot{k}, \dot{\gamma}, \dot{g}, \dot{e})\). Note the definition of \( M_{ij}^k, M_{ij}^\gamma, D_t^{-1}, N_{tij}, M_{ij}^g, M_{ij}^e, s_{tij} \) in teh following
expression.

\[ \dot{F}(\dot{k}, \dot{\gamma}, \dot{g}, \dot{e}) = \sum_{ij} \left( \lambda K_{ij} + L_{tij} \frac{\dot{k}_{ij}}{1 + K_{ij} \beta_{tij}} \right) \dot{k}_{ij} \]

\[ + \sum_{ij} \left( \sum_{i : (t,i,j) \in T} L_{tij} \right) \dot{\gamma}_{tij} \]

\[ + \sum_{ij} \left( \sum_{j : (t,i,j) \in T} L_{tij} \right) \dot{g}_{ti} \]

\[ + \sum_{ij} \left( \lambda E_{ij} + \sum_{t,i,j} L_{tij} \left( \delta_{(j',j)} - P_{til} \right) E_{il} \beta_{tl} \right) \dot{e}_{il} \]

Following the same procedure and replacing \( \dot{r}_{tij} \) in the second constrain

\[ -\beta_{tij} \dot{\gamma}_{tij} = \sum_{i} s_{tij} \dot{r}_{tij} \]

\[ = \sum_{i} \frac{s_{tij}}{1 + K_{ij} \beta_{tij}} \left( \dot{k}_{ij} + \dot{\gamma}_{tij} \right) + \sum_{i} s_{tij} \dot{g}_{ti} + \sum_{i,l} s_{tij} \left( \delta_{(j',l)} - P_{til} \right) E_{il} \beta_{tl} \left( \dot{e}_{il} + \dot{\gamma}_{il} \right) \]

rearranging and solving for \( \dot{\gamma} \)

\[ \sum_{l} \left[ -\delta_{(j,l)} \beta_{lj} - \sum_{i} \left( s_{tij} \left( \delta_{(j,l)} - P_{til} \right) E_{il} \beta_{tl} \right) \right] \dot{\gamma}_{lil} = \sum_{i} \frac{N_{ij}^{k} \dot{k}_{ij}}{N_{ijkl}} + \sum_{i} s_{tij} \dot{g}_{ti} \]

\[ + \sum_{il} \left[ s_{tij} \left( \delta_{(j,l)} - P_{til} \right) E_{il} \beta_{tl} \right] \dot{e}_{il}. \]

The equation above is a matrix multiplication. Defining \( D_{tij} = -\delta_{(j,l)} \beta_{lj} - N_{ijkl}^{\gamma} \) and writing the equation above in tensor notation helps identify the matrix multiplication
\[ D_{ij}^t \gamma_t = N_j^i \dot{k}_{ij} + s_{i j}^t \dot{g}_t + N_i^j \dot{e}_{ij} \]  

(B.17)

so \( \dot{\gamma}_{tj} \) can be found via matrix inversion. Note that the matrix to invert is of size \( N_{\text{SITES}} \times N_{\text{SITES}} \) and that it has the form of a diagonal matrix with a sum of perturbations of rank 1. The matrix can be inverted efficiently (in \( O(N_{\text{SITES}}^2) \) time) using the Sherman Morrison Formula (Press et al. (2002, Numerical Recipes in C++)). \( \dot{\gamma}_{tj} \)

becomes (back in summation notation)

\[ \dot{\gamma}_{tj} = \sum_l \left[ D_{t}^{-1} \right]_{jl} \left[ \sum_i N_{ij}^l \dot{k}_{ij} + \sum_i s_{tij} \dot{g}_t + \sum_{il} s_{tij} (\delta_{(j,l)} - P_{til}) E_{il} \beta_{tl} \right] \dot{e}_{il} \]  

(B.18)

Inserting the expression for \( \dot{\gamma} \) in the gradient has it become a function of \((\dot{k}_{ij}, \dot{g}_t, \dot{e}_{ij})\).

\[ \dot{F} = \sum_{ij} \left( M_{ij}^k + \sum_{ij'} M_{ij}^{ij'} \left[ D_t^{-1} \right]_{jj'} N_{tij}^k \right) \dot{k}_{ij} \]

\[ + \sum_{ti} \left( M_{ij}^g + \sum_{j'} M_{ij}^{ij'} \left[ D_t^{-1} \right]_{jj'} s_{tij} \right) \dot{g}_t \]

\[ + \sum_{ij} \left( M_{ij}^e + \sum_{ij'} M_{ij}^{ij'} \left[ D_t^{-1} \right]_{jj'} s_{tij} (\delta_{(j,l)} - P_{tij}) E_{ij} \beta_{ij} \right) \dot{e}_{ij} \]  

(B.19)

We can interpret this expression as the variation in \( F \) when following a path through \((k, g, e)\) space along the direction \((\dot{k}_{ij}, \dot{g}_t, \dot{e}_{ij})\).

\[ \dot{F} = \nabla F(k, g, e)^T \cdot \begin{bmatrix} \dot{k} \\ \dot{g} \\ \dot{e} \end{bmatrix} \]  

(B.20)

To minimize \( F \), we use a gradient descent method. For a given descent direction \((\dot{k}, \dot{\xi}, \dot{\beta})\) we obtain the change in the value of the parameter \( K \) with a multiplicative
update that results from applying an additive update on \( k \) as follows

\[
k_{ij} = k_{ij} + \gamma \dot{k}_{ij}
\]

\[
\log K_{ij} = \log K_{ij} + \gamma \dot{k}_{ij}
\]

\[
K_{ij} = K_{ij} e^{+\gamma \dot{k}_{ij}}.
\]

(B.21)

Note that \( \dot{k}_{ij} \) is already a descent direction, hence the positive sign. For instance, if we followed a path along the gradient then \( \dot{k}_{ij} = -\nabla k_{ij} F(k, \xi, g) \). The procedure to obtain the updates for \( R_i \) and \( f_{ti} \) is analogous.

If the gradient descent method of choice is steepest descent (Bertsekas, 2004), then the descent direction is chosen to be the gradient,

\[
\dot{k}_{ij} = - \left( M_{ij}^k + \sum_t G_{ij} N_{tij}^k \right)
\]

(B.22)

\[
\dot{g}_{ti} = - \left( M_{ij}^s + \sum_j G_{ij} s_{tij} \right)
\]

(B.23)

\[
\dot{e}_{ij} = - \left( M_{ij}^e + \sum_{tl} G_{lt} s_{tij} \left( \delta_{j,l} - P_{tij} \right) E_{ij} \right)
\]

(B.24)

where we have introduced \( G_{ij} = \sum_j M_{ij}^r \left[ D_t^{-1} \right]_{j,j} \) to simplify notation. In practice a conjugate gradient descent method (Bertsekas, 2004) may be a more robust choice.
B.2 Computation of the Hessian of $F(\xi, k)$ for the Nucleus Specific Economy

Here we detail the calculations to obtain the Hessian. For simplicity we compute the terms of the Hessian independently for each term in the cost function. We start by computing the derivatives of $s_{t,c,j}$ as they are needed in the rest of the calculations. Recall the expression for $s_{t,c,j}$

$$s_{t,c,j} = \frac{e^{k_{c,j} e^{\sum_{t \in c} \xi_{t}}} \times \left[ [\ell \in c] - \sum_{c' \in c_j} s_{t,c'}[[\ell \in c']] \right]}{1 + e^{k_{c,j} e^{\sum_{t \in c} \xi_{t}}}}$$

its derivatives are

$$\frac{\partial s_{t,c,j}}{\partial k_{\alpha \beta}} = \delta_{\beta j} s_{t,c,j} (\delta_{c\alpha} - s_{t\alpha \beta}) \quad (B.25)$$

$$\frac{\partial s_{t,c,j}}{\partial \xi_{t \ell}} = s_{t,c,j} \left( [\ell \in c] - \sum_{c' \in c_j} s_{t,c'}[[\ell \in c']] \right) \quad (B.26)$$

where we use $[[::]]$ to denote an indicator function.

We now show the gradient and Hessian terms for the biological regularizer. Recall the expression of the regularizer,

$$F_{BIO}(\xi_{ti}, k_{cj}, f_{ti}) = -\sum_{ti} f_{ti} \xi_{ti} + \sum_{ti} e^{\xi_{ti} + \sum_{tj} \log \left( 1 + \sum_{c \in C_j} e^{k_{c,j} e^{\sum_{t \in c} \xi_{t}}} \right)} ,$$

its gradient is,

$$\frac{\partial F_{BIO}}{\partial k_{\alpha \beta}} = \sum_{t} s_{t \alpha \beta} \quad (B.27)$$

$$\frac{\partial F_{BIO}}{\partial \xi_{t \ell}} = -f_{t \ell} + e^{\xi_{t \ell}} + \sum_{j} \sum_{c \in c_j} s_{t,c,j}[[\ell \in c]] \quad (B.28)$$
and its Hessian,

\[
\frac{\partial^2 F_{\text{BIO}}}{\partial k_{\alpha \beta} \partial k_{nm}} = \sum_t \frac{\partial s_{ta \beta}}{\partial k_{nm}} = \sum_t \delta_{\beta m} s_{ta \beta} (\delta_{\alpha n} - s_{tnm}) \tag{B.29}
\]

\[
\frac{\partial^2 F_{\text{BIO}}}{\partial \xi_t \partial \xi_{t'}'} = \delta_{tt'} e^{\xi_{tt}} + \sum_j \sum_{c \in c_j} \frac{\partial s_{tcj}}{\partial \xi_{t'}'} [[\ell \in c]]
\]

\[
= \delta_{tt'} e^{\xi_{tt}}
\]

\[
+ \delta_{tt'} \sum_j \sum_{c \in c_j} s_{tcj} [[\ell \in c]] \left( [[n \in c]] - \sum_{c' \in c_j} s_{tc'j} [[n \in c']] \right) \tag{B.30}
\]

\[
\frac{\partial^2 F_{\text{BIO}}}{\partial k_{\alpha \beta} \partial k_{\ell}} = s_{ta \beta} \left( [[\ell \in \alpha]] - \sum_{c' \in c_j} s_{tc'j} [[\ell \in c']] \right) \tag{B.31}
\]

We now show the terms of the gradient and Hessian of the Loss function. Recall that we chose to use the Kullbach-Leibler divergence as loss function, and that we have completed the distribution with an empty protein (∅).

\[
F_L(R_{ti}, K_{cj}, f_{ti}) = \sum_{tcj:(t,cj) \in T} \text{Loss} \left( \hat{s}_{tcj}, s_{tcj}(R_{ti; i \in c_j}, K_{cj}) \right)
\]

\[
= - \sum_{c \in C_j \cup \emptyset} s_{*tjc} \log \frac{s_{tcj}}{s_{*tjc}}
\]

\[
\text{s.t. } \hat{s}_{tji} = \sum_{c \in C_j; i \in c} s_{*tjc} \forall i
\]

\[
\sum_{c \in C_j \cup \emptyset} s_{*tjc} = 1
\]
its gradient is,

\[
\frac{\partial F_L}{\partial k_{\alpha\beta}} = -\sum_t s_{t\alpha\beta}^* [\alpha_{in}C_{\beta}] + \sum_t s_{t\alpha\beta} \tag{B.32}
\]

\[
\frac{\partial F_L}{\partial \xi_{tt}} = \sum_j \sum_{c \in C_j} (s_{tcj} - s_{tcj}^*) [[\ell_{inc}]] \tag{B.33}
\]

and its Hessian is,

\[
\frac{\partial^2 F_L}{\partial k_{\alpha\beta} \partial k_{nm}} = \sum_t \frac{\partial s_{t\alpha\beta}}{\partial k_{nm}} = \frac{\partial^2 F_{BIO}}{\partial k_{\alpha\beta} \partial k_{nm}} \tag{B.34}
\]

\[
\frac{\partial^2 F_L}{\partial \xi_{tt} \partial \xi_{t'n}} = \delta_{tt'} \sum_j \sum_{c \in e_j} \frac{\partial s_{tej} \ell_{\ell \in c}}{\partial \xi_{t'n}} = \frac{\partial^2 F_{BIO}}{\partial \xi_{tt} \partial \xi_{t'n}} - \delta_{tt'} \delta_{\ell \in c} \ell_{\ell t} e_{\xi t} \tag{B.35}
\]

\[
\frac{\partial^2 F_L}{\partial k_{\alpha\beta} \partial \xi_{tt}} = \sum_t \delta_{tt'} \frac{\partial s_{t\alpha\beta}}{\partial \xi_{tt}} = \frac{\partial^2 F_{BIO}}{\partial k_{\alpha\beta} \partial \xi_{tt}} \tag{B.36}
\]

The only remaining term that depends on the variables of the optimization of \(F(\xi, k)\) is the one-norm regularization of \(K_{cj}\). We recall its expression in terms of \(k_{cj}\),

\[
F_{L1} = \sum_{c, j} e^{k_{cj}}
\]

and its gradient and Hessian are

\[
\frac{\partial F_{L1}}{\partial k_{\alpha\beta}} = \sum_{c, j} \frac{\partial}{\partial k_{\alpha\beta}} e^{k_{cj}} = e^{k_{\alpha\beta}} \tag{B.37}
\]

\[
\frac{\partial F_{L1}}{\partial k_{\alpha\beta} \partial k_{nm}} = \frac{\partial}{\partial k_{nm}} e^{k_{\alpha\beta}} = \delta_{\alpha n} \delta_{\beta m} e^{k_{\alpha\beta}} \tag{B.38}
\]

\[
\frac{\partial F_{L1}}{\partial \xi_{tt}} = 0 = \frac{\partial^2 F_{L1}}{\partial \xi_{tt} \partial k_{\alpha\beta}} = \frac{\partial^2 F_{L1}}{\partial \xi_{tt} \partial \xi_{t'n}} \tag{B.39}
\]

Regrouping all terms we get an expression for the gradient and the Hessian of the cost function in the optimization of \(F(\xi, k)\).

\[
F(\xi_{tt}, k_{cj}) = F_L + \lambda_1 F_{BIO} + \lambda_2 F_{L1} \tag{B.40}
\]
\[
\frac{\partial F}{\partial k_{\alpha \beta}} = -\sum_t s_{t_\alpha \beta}^{*}[\alpha in C_{\beta}] + \sum_t s_{t_\alpha \beta} + \lambda_1 \sum_t s_{t_\alpha \beta} + \lambda_2 e^{k_{\alpha \beta}} \\
= -\sum_t s_{t_\alpha \beta}^{*}[\alpha in C_{\beta}] + (\lambda_1 + 1) \sum_t s_{t_\alpha \beta} + \lambda_2 e^{k_{\alpha \beta}} \tag{B.41}
\]

\[
\frac{\partial F}{\partial \xi_{i \ell}} = \lambda_1 (-f_{i \ell} + e^{\xi_{i \ell}}) + (\lambda_1 + 1) \sum_j \sum_c s_{tcj}[[\ell \in c]] \\
- \sum_j \sum_{c \in C_j} s_{tcj}^{*}[[lin c]] \tag{B.42}
\]

\[
\frac{\partial^2 F}{\partial k_{\alpha \beta} \partial k_{nm}} = (\lambda_1 + 1) \frac{\partial^2 F_{BIO}}{\partial k_{\alpha \beta} \partial k_{nm}} + \lambda_2 \delta_{\alpha n} \delta_{\beta m} e^{k_{\alpha \beta}} \tag{B.43}
\]
\[
\frac{\partial^2 F}{\partial \xi_{i \ell} \partial \xi_{i' n}} = (\lambda_1 + 1) \frac{\partial^2 F_{BIO}}{\partial \xi_{i \ell} \partial \xi_{i' n}} - \delta_{i i'} \delta_{\ell n} e^{\xi_{i \ell}} \tag{B.44}
\]
\[
\frac{\partial^2 F}{\partial k_{\alpha \beta} \partial \xi_{i \ell}} = (\lambda_1 + 1) \frac{\partial^2 F_{BIO}}{\partial k_{\alpha \beta} \partial \xi_{i \ell}} \tag{B.45}
\]
Appendix C

Appendix: Experimental Set-Up and background

In this appendix we provide a broad description of the Chip-Chip Microarray technique, the experiment design and how to process the resulting data for use with our learning algorithms.

C.1 Chromatin Immuno-Precipitation Microarrays

Chip-Chip Microarray (Chromatin Immuno-Precipitation Microarray Chip, cf. Lee et al. (2006)) data used in this thesis (simulated or experimental) follows the experimental criteria defined in (Harbison et al., 2004; Lee et al., 2002; Ren et al., 2000). Data is publicly available and can be downloaded from the web repository of the http://inside.wi.mit.edu/young/pub/download.htmlYoung lab (Lab).

C.1.1 Chip-Chip Microarray data

Chip-Chip microarray data is generally presented as binding ratios for each probe in a Microarray Chip. Broadly speaking, binding ratios compare the signal intensities produced by target protein bindings\(^1\) against a sample of non-specific DNA fragments.

\(^1\)Multiple cells from a single culture are harvested and their DNA fragmented by sonication. A sample of this solution is set aside as a representative aliquot of the original solution of DNA.
Both samples are biochemically amplified and labelled with a fluorescent marker prior to hybridization in a microchip array. The microchip array is washed and recorded by measuring the fluorescent intensity response to colored-light (one per marker used). The procedure resembles photography in that hybridization “exposes” the chip, washing exposes and fixates binding, and the result can be seen by shining light on the array. The binding ratios result from processing the intensities recorded for each of the two-color markers.

### C.2 Experiments on Learning Site-Specific Economies

#### C.2.1 Preprocessing Microarray Data

Our algorithms compare binding ratios (relative measures) to predicted per-site binding probabilities. Some preprocessing is required. Preprocessing is largely dependent on experiment design and a number of assumptions about how comparable are binding ratios from different microchip arrays. We determined the best procedure for the site-specific economy using cross-validation. Here we show our conclusions.

Let $\hat{o}_{tij}$ be the binding ratios obtained after processing the microarray assays. The experiment design is such that each microarray measures the binding profile of a single regulator $i$, $\hat{o}_{tij}$ is the result of averaging multiple repetitions (typically 3) of the same measurement. Each regulator is profiled against a number of experimental conditions indexed by $t$, and results are stored for each probe $j$ in the microarray. In this work we associate $j$ with sites although we note that a probe may well contain multiple sites and some overlap between probes is likely to occur. The work of (MacIsaac 2007) fragments; it is often referred to as whole-cell extract. Using the rest of the sample, target DNA fragments are isolated by immunoprecipitation. Target DNA fragments are those bound by the protein target of the experiment. Immunoprecipitation is a technique to precipitate target DNA fragments using protein-specific anti-bodies. The remaining DNA fragments remain in solution. Typically, this procedure is termed DNA enrichment. Readers unfamiliar with molecular biology terminology may find be mislead by the term DNA enrichment. DNA isn’t really enriched; it might be more adequate to say that the solution is enriched in target DNA fragments, which means that the concentration of such DNA fragments is significantly and artificially increased with respect to that of non-target DNA fragments by selective precipitation. The word enriched is used instead of isolated to emphasize that immunoprecipitation is not a noise-free procedure; the final mixture may contain small proportions of fragments other than the target fragments.
et al., 2006) resolves this overlap combining several conservation-based motif discovery algorithms. However in this thesis we choose to make the coarser association of probes and sites to remain as close to the original data as possible. We transform the measurements $\hat{o}_{tij}$ into a per-site binding probability measure $\hat{s}_{tij}$ by any of the following two methods depending on the use we give to $\hat{s}_{tij}$.

$$\hat{r}_{tij} = \frac{\hat{o}_{tij}}{\max_j \hat{o}_{tij}} \quad (C.1)$$

$$\hat{s}_{tij} = \frac{\hat{r}_{tij}}{\max_j \sum_i \hat{r}_{tij}} \cdot 0.99$$

$$\tilde{s}_{tij} = \frac{\hat{o}_{tij}}{N_{sites} \sum_j \hat{s}_{tij}} \cdot 0.99. \quad (C.2)$$

(C.1) assumes that signals may be compared across different microarrays (that is $(t, i)$ pairs) as a measure of binding relative to each microarray’s maximum observed binding. Each $\tilde{s}_{tij}$ is then normalized by the maximum total protein-binding observed at each site. Multiplication by 0.99 is a simple way to account for reversibility of site-reactions. This measure has proved useful to compute the initialization of the game parameters and to compute the Loss function. However, using (C.1) for initialization purposes solves the problem of feeding the economy with a set of parameters that incorporate some information about the overall scale of the binding distribution. This problem arises if parameters are initialized with random values. Then simpler methods are preferred to evaluate the Loss of a given prediction. The method in (C.2) does not need to assume that signals are comparable across microarrays. It can be used to compare overall binding in each microarray assay to the same average predicted by our algorithms. Generally, one such measure would be unable to discriminate between different scalings in prediction and observations, this is solved by setting the overall scale in the initialization.

### C.2.2 Learning Initialization

The choice of a starting point for the learning algorithm affects the final value of the parameters learnt for the site-specific economy. The algorithm we presented in Section 5.2.2 is only guaranteed to find a local minimum. the choice of initial param-
eters implicitly determines convergence to the closest local minima. As discussed in section 5.2.1 this is a consequence of the internal degrees of freedom that is normally resolved by choosing a scale (physically, this choice is implicit in the choice of units). Unfortunately, the relative nature of microarray binding ratios prevents us from setting this scale. The procedure of amplification erased the information about the true initial availability of proteins. It is nonetheless possible to recover partial information about the relative scaling across microarrays, as shown in equation (C.1) above, and use it to feed the iterative learning algorithm.

For all practical purposes, this limitation is only a true limitation to convert back and forth to physical units of concentration as that is the information removed. The effect on the learning and prediction abilities of our algorithms can be managed by choosing adequate initialization values. Consider the flow diagram summarizing the prediction process in Figure 6-4 (conveniently reproduced below); it assumes that we study the effect of varying $f$ on the binding profiles. The learning algorithm provided us with estimated values of $K, E, f$. For as long as the variation of $f$ in Figure 6-4 is interpreted in relative terms with respect to the learnt $f$ values rather than to an absolute scale, it will be possible to translate the predictions of the algorithm to the experiments. Conversely, incorporating known scale information into the initialization of $K, E, f$ is likely to yield learnt values for $K, E, f$ that are closer to being comparable to physically dimensioned constants. We hypothesize that scale information may contribute to increase robustness and reduce the number of iterations in learning. We expect the cost function to be sharper near values that

Figure C-1: Block diagram for predicting binding profiles in the site specific $\lambda$-phage economy.
correspond to true physical units.

When scaling information is unavailable we have found the following procedure to provide a reasonable starting point for the learning algorithm. It is based on equation (C.1) and several simplifications of the chemical equilibrium condition.

I \( K^0_{ij} = \frac{\sum_t \delta_{tij}}{N_{\text{experiments}} \max_{t,i,j} \delta_{tij}} \)

II \( f^0_{ti} = \sum_j K_{ij} \left( 1 - \sum_i \delta_{tij} \right) \)

III \( E^0_{ij} = \frac{1}{2} \frac{\sum_t \delta_{tij}}{N_{\text{experiments}} \max_{j} \delta_{tij}} \)

C.3 Experiments in Learning the Nucleus Specific Economy

C.3.1 Data Description

We overview the main details of data used in our experiments that originates from (Harbison et al., 2004). Data profiles 203 transcriptional regulators for cultures of Saccharomyces Cerevisiae. All regulators are profiled in a rich-media environment and at least 84 among these are additionally profiled in bacteria grown in 12 other different environmental conditions.

- Rich media. Cells were grown in YPD (1% yeast extract/2% peptone/2% glucose) to an OD600 of 0.8.

- Highly hyperoxic. Cells were grown in YPD to an OD600 of 0.5 followed by treatment with hydrogen peroxide (4 mM final) for 30 minutes.

- Moderately hyperoxic. Cells were grown in YPD to an OD600 of 0.5 followed by treatment with hydrogen peroxide (0.4 mM final) for 20 minutes.

- Amino acid starvation. Cells were grown to an OD600 of 0.6 in synthetic complete medium followed by treatment with the inhibitor of amino acid biosynthesis sulfometuron methyl (0.2 mg/ml final) for two hours.
• Nutrient deprived. Cells were grown in YPD to an OD600 of 0.8 followed by treatment with rapamycin (100 nM final) for 20 minutes.

• Filamentation inducing. Cells were grown in YPD containing 1% butanol for either 90 minutes or 14 hours (corresponding to an OD600 of 0.8).

• Mating inducing. Cells were grown in YPD to an OD600 of 0.8 followed by treatment with the alpha factor pheromone (5 mg/ml) for 30 minutes.

• Elevated temperature. Cells were grown in YPD at 30°C to an OD600 of 0.5 followed by a temperature shift to 37°C for 45 minutes.

• Galactose medium. Cells were grown in YEP medium supplemented with galactose (2%) to an OD600 of 0.8.

• Raffinose medium. Cells were grown in YEP medium supplemented with raffinose (2%) to an OD600 of 0.8.

• Acidic medium. Cells were grown in YPD to an OD600 of 0.5 followed by treatment for 30 minutes with succinic acid (0.05 M final) to reach a pH of 4.0.

• Phosphate deprived medium. Cells were grown in synthetic complete medium lacking phosphate to a final OD of 0.8.

• Vitamin deprived medium. Cells were grown in synthetic complete medium lacking thiamin to a final OD of 0.8.

The choice of which regulators to profile under different conditions is based on knowledge about the function of the regulators. The Microarray design included essentially all of the known intergenic regions in the yeast genome as given by the Yeast Intergenic Region Primers (Research Genetics/Invitrogen). Specifically, that is 6361 intergenic regions in the first generation of arrays and 400 additional intergenic regions in subsequent generations (as the Yeast Intergenic Region Primers were updated). In the resulting microarrays, the average size of the spotted PCR products was 480 bp, and the sizes ranged from 60 bp to 1500 bp.

Processed Data was obtained from (Lab), and is also available from ArrayExpress (http://www.ebi.ac.uk/aerep/result?queryFor=Experiment&eAccession=E-WMIT-10).
One of the main results of (Harbison et al., 2004) is a regulatory map of yeast, that has since been updated (MacIsaac et al., 2006) and can be consulted at http://fraenkel.mit.edu/improved_map/. Additionally, a summary of the behavior observed for the altered conditions is available from http://jura.wi.mit.edu/young_public/regulatory_code/Behaviors.html.

Data Limitations and Considerations

Our learning algorithms can work with observations from multiple experiments and repetitions. Data, however, is only available averaged over multiple repetitions according to a standard error-model (i.e. there is one observation per condition-regulator-site triplet). This limits our ability to control complexity via the biological and 1-norm regularizers. Additionally, conditions with at most 2 regulators profiled do not contain sufficient data to distinguish between the effects of protein availability and chemical equilibrium constant. Conversely, regulators that appear only in one condition represent a single data-point and are thus equally prone to overfitting.

The above considerations constrain the number of regulators and conditions that we can use. We initially restrict our attention to conditions with at least 2-3 regulators profiled and to regulators profiled in at least 3 conditions.

We follow the assumption from (Harbison et al., 2004) that regulators that were not profiled in a particular condition behave like in YPD condition. Our algorithms allow us to introduce this assumptions in two ways:

- filling unreported bindings with the corresponding YPD value.
- Instructing the learning algorithm to equal the initial availability of that regulator in the unreported condition to that of YPD.

Finally, sites that do not show distinct binding signals for any condition according to the criteria set forth in (Lee et al., 2006) are left out of the analysis.

In practice these considerations limit the data that can be used for the economic analysis. We consider two cases,
These cases, while restricted, are sufficient to illustrate the scalability of the economic approach. We emphasize that these restrictions are the result of adapting our economic approach to an experimental design that was intended for a different purpose. That is, these restrictions do not arise from limitations of the experiment design nor the algorithm alone. They result from finding an agreement between the purpose of our algorithms and the original intent of the experimental setting from (Harbison et al., 2004). For instance, data from multiple repetitions would ease our restrictions on the number of regulators profiled for each condition.

### C.3.2 Preprocessing

Our algorithms predict per-site complex binding rates (probabilities). Data provides binding ratios for each protein and site (we associate site with probe). Binding ratios compare the target signal intensity with a background intensity. Here we explain how to translate binding ratios to binding rates, in the next section we overview the steps in the nucleus specific economy that deal with complex formation.

The relative nature of microarray measurements and the differing sensitivity of the antibodies used in Immuno-Precipitation for each regulator pose an additional challenge to use microarray intensities as a quantitative measure of binding. Binding events generally show intensities 2 to 15-fold the background intensity. However it is not uncommon to see binding ratios of 80 and larger. Lee et al. (2006) single out the specificity of the Immuno-Precipitation step as the main cause for the wide range of values, and note that values are otherwise comparable across microarray assays and experiments (note that each microarray assay profiles a single protein.) The general preprocessing method introduced in Section 5.3.2 does not take care of immuno-precipitation specificity.

To prevent over-representing proteins with a highly specific immuno-precipitation

<table>
<thead>
<tr>
<th>set id</th>
<th># conditions</th>
<th># regulators</th>
<th># sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced set (for readability)</td>
<td>10</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>Extended set</td>
<td>8</td>
<td>27</td>
<td>414.</td>
</tr>
</tbody>
</table>
step we introduce a new preprocessing that takes into account the observations of Lee et al. (2006). Indexing $t$ experiments, $i$ proteins, $j$ sites, and denoting binding ratios with $\hat{o}_{tij}$ and binding probabilities with $\hat{s}_{tij}$, we used the following two methods to normalize observations with similar results.

The first method is extends the simple normalization strategy introduced in Section 5.3.2 with considerations about the range of binding signals. This normalization does not take the p-value into account.

\[
\hat{s}_{tij} = \frac{\hat{o}_{tij} - \min \left( \min_{t,j,2nd} \hat{o}_{tij}, 1 \right)}{\max \left( \max_{t,j,2nd} \hat{o}_{tij} - \min \left( \min_{t,j,2nd} \hat{o}_{tij}, 1 \right) \right)}
\]

\[
\hat{s}_{tij} = \min \left( \max \left( \hat{s}_{tij}, 0 \right), .99 \right).
\]

Where we use $2^{nd}$ to denote the second maximum or minimum, the factor 2 in the denominator emphasizes that binding events are expected to have at least a 2-fold binding ratio, and the factor 0.99 prevents infinite equilibrium constants. Note that maximisation over $t$, and $j$ assumes comparable assays and non-comparable proteins, which is consistent with the previous observations. This method works well for datasets selections based on signal strength and confidence.

The second method uses the p-value to set the signal-to-noise ratio and is more adequate for larger datasets.

\[
\varepsilon = 1.5;
\]

\[
\delta = .0001
\]

\[
\overline{o_{tij}} = \max_{t,j,2nd} \hat{o}_{tij}
\]

\[
o_{tij} = \max \left( \min_{t,j,2nd} \hat{o}_{tij}, \varepsilon \right)
\]

\[
\{\tilde{s}_{tij}\} = \left\{ \tilde{s}_{tij} = \frac{\hat{o}_{tij} - \overline{o_{tij}}}{\overline{o_{tij}} - o_{tij}} \left| pval_{tij} < \delta \right. \right\}
\]

\[
\hat{s}_{tij} = \min \left( \max \left( \tilde{s}_{tij}, 0 \right), .99 \right). \quad (C.4)
\]

If the second method is used, it is necessary to further verify that for each regulator-
site pair, there still exist a measurement in YPD condition. That is, that the measurement of the same regulator-site pair in YPD condition had a high confidence (small p-value). This restriction is imposed by experiment design to propagate YPD values to conditions where a given regulator is unreported. In practice this will further reduce the number of sites in the analysis and in some cases it might be advantageous to relax the confidence constraint for YPD measurements. In our experiments, relaxing the confidence of YPD measurements up to pval < 0.01 increased the scope of the analysis, but had otherwise no significant impact on the predictions obtained with more restrictive p-values.

Note that Microarray measurements are significantly affected by noise. The number of high-confidence probes can be expected to be only a fraction of the number of probes in the Chip, even for large p-values (e.g., p-value $\geq 0.01$). Further requiring intersection with YPD measurements will decrease the number of high-confidence probes that can be used in the analysis. In our experiments with p-value $\geq 0.001$ the number of high-confidence probes was reduced to about 600 from the initial 6000.

These preprocessing methods differ from the method introduced in Section 5.3.2 in that they do not assume that signal-to-noise ratio and measurement confidence have already been taken care of. If signal-to-noise ratio and measurement confidence are not taken into account, noisy and low-confidence observations tend to increase the number of strong binding events artificially. Normalization errors are, however, easy to detect because the economy will tend to over-predict complex bindings to explain the abnormally high levels of binding; this effect is particularly noticeable for proteins with less specific immunoprecipitation steps.

As noted in the analysis of the site-specific economy, additional information about the overall scale of the problem may greatly improve the choice of a preprocessing strategy.

### C.3.3 Complex Formation

The algorithm introduced in Section 5.3.2 takes care of inferring per-complex probabilities $s^*_{tcj} \in S$ from per-protein probabilities $s_{ti,j}$ given a set $S$. In all generality we
may avoid defining the set $S$ and find directly a feasible solution solving a family of linear programs of the general form,

\[
\begin{align*}
\text{minimize} & \quad \mathbf{c}^T \mathbf{\varepsilon} \\
\text{s.t.} & \quad \left| \hat{s}_{ij} - \sum_{c:j \in c} s^*_{tcj} \right| \leq \varepsilon_i \\
& \quad s^*_{tcj} \geq 0 \\
& \quad \sum_c s^*tcj \leq 1.
\end{align*}
\] (C.5)

In practice, however, this strategy does not scale well unless some restrictions are imposed on the number of proteins that may form a complex. For instance, limiting the search space to 2-protein complexes requires solving $\binom{n_2}{2}$ linear programs per experiment and site. Using a method like infeasible start newton method we can pass the task of finding a feasible solution to the minimization of the loss function and use a heuristic to grow $S$ to guarantee that one feasible solution exists. The heuristic is simply a greedy procedure to form first 2-protein complexes out of the maximally bound singletons until it is either no longer possible or we find that $\sum_c \min_i \hat{s}_{ij:i \in c} \leq 1$. If the condition is not satisfied we continue adding three protein complexes and so on. In the worst case scenario, when all singletons are fully bound ($\hat{s}_{ij} = 1$) this procedure will produce the power-set. In practice stopping the search of 2-protein complexes early greatly reduces that risk. For instance, stop forming 2-protein complexes when the remaining proteins have a probability of binding smaller than 10%. This procedure is based on the assumption that the proteins that are more likely to form complexes are those that bind the site more frequently.


Xi Chen and Xiaotie Deng. 3-nash is ppad-complete. *Electronic Colloquium on Computational Complexity (ECCC)*, (134), 2005.


Luis E. Ortiz. Luis E. Ortiz, personal communication.


Yuan Qi, Alex Rolfe, Kenzie D. MacIsaac, Georg K. Gerber, Dmitry Pokholok, Julia Zeitlinger, Timothy Danford, Robin D. Dowell, Ernest Fraenkel, Tommi S.


