Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-018 March 6, 2006

DNA Binding and Games

Luis Perez-Breva, Luis E. Ortiz, Chen-Hsiang Yeang, and Tommi Jaakkola

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

DNA Binding and Games

Luis Pérez-Breva[a,*], Luis E. Ortiz[a], Chen-Hsiang Yeang[b], and Tommi Jaakkola[a,*]
[a]MIT - CSAIL, Cambridge, MA
[b]UC Santa Cruz, Santa Cruz, CA

March 6, 2006

Abstract

We propose a game-theoretic approach to learn and predict coordinate binding of multiple DNA binding regulators. The framework implements resource constrained allocation of proteins to local neighborhoods as well as to sites themselves, and explicates coordinate and competitive binding relations among proteins with affinity to the site or region. The focus of this paper is on mathematical foundations of the new modeling approach. We demonstrate the approach in the context of the λ-phage switch, a well-known biological subsystem, and provide simulation results that successfully illustrate the predictions that can be derived from the model with known structure and affinities. Subsequent work will elaborate on methods for learning the affinities and game structures from available binding data.

1 Introduction

Effective transcriptional control relies in part on coordinate operation of DNA binding regulators and their interactions with various co-factors. Understanding such processes is challenging, however, since the role of interactions or binding sites associated with any specific genes may not be transparent if considered in isolation. The broader goal of our work here is to precisely capture the context provided by other mutually constraining processes.

We believe game theory and economic models provide an appropriate framework for understanding (and searching for) the context and constraints associated with interacting regulatory processes. In particular, the problem of understanding coordinate binding of regulatory proteins has many game theoretic properties. Resource constraints, for example, are critical to understanding who binds where. At low nuclear concentrations, regulatory proteins may occupy only high affinity sites, while filling weaker sites with increasing concentration. Overlapping or close binding sites create explicit competition for the sites, the resolution of which is guided by the available concentrations around the binding sites. Similarly, explicit coordination such as formation of larger protein complexes may be required for binding or, alternatively, binding may be facilitated by the presence of another protein. The key advantage of games as models of binding, however, is that they provide causally meaningful predictions, binding arrangements, in response to various experimental perturbations or disruptions.

Our approach deviates from an already substantial body of computational methods used for resolving transcriptional regulation. Much of the recent work has been statistical in nature as in identifying regulatory modules either by combining available binding data with mRNA expression profiles [3] or by relating mRNA levels of candidate regulators directly to their potential targets [12]. Our work is closest in spirit to more detailed reaction equation models [7, 1], while narrower in scope due to our preliminary focus on binding alone. Our conceptualization of the binding problem is nevertheless different, and the mathematical modeling approach is clearly distinct from reaction equation models, even in terms of the level of analysis.

*Corresponding authors: {lpbreva,tommi}@csail.mit.edu

We begin by clarifying the structure of the binding problem, followed by translating the problem into a game theoretic form. The formal approach, analysis, and associated algorithms for making predictions are presented in subsequent sections. We also provide an initial small-scale demonstration of our model in the context of the λ-phage lysogeny switch.

2 Protein-DNA binding

Before formalizing the game, we decompose the binding problem further into *transport* and *local binding*. By *transport*, we refer to the mechanism that transports proteins to the neighborhood of sites to which they have affinity. The biological processes underlying the transport are not well-understood although several hypotheses exist[14, 4]. We abstract the process by assuming separate affinities for proteins to explore neighborhoods of specific sites, modulated by whether the sites are available. This abstraction does not address the dynamics of the transport process and therefore does not distinguish (nor stand in contradiction to) underlying mechanisms that may or may not involve diffusion as a major component. We simply aim to capture the differentiated manner in which proteins may accumulate in the neighborhoods of sites depending on the overall nuclear concentrations and regardless of the time involved.

Local binding, on the other hand, captures which proteins bind to each site as a consequence of local accumulations or concentrations around the site. We assume that the neighborhood of each site constitutes a chemically well-mixed and closed system. Thus, we model the binding as being governed by chemical equilibria: for a type of protein i around site j, {free protein i} + {free site j} $\rightleftharpoons$ {bound ij}, where concentrations involving the site should be thought of as time averages or averages across a population of cells depending on the type of predictions sought. The concentrations of various molecular species around and bound to the sites as well as the rate at which the sites are occupied are then governed by the chemical equilibrium equation:

$$\frac{[\text{bound } ij]}{[\text{free protein } i][\text{free site } j]} = K_{ij},$$

where i ranges over proteins with affinity to site j and K_{ij} is a positive equilibrium constant characterizing protein i's ability to bind to site j in the absence of other proteins.

Broadly speaking, the combination of transport and local binding results in an arrangement of proteins along the possible DNA binding sites. This is what we aim to predict with our game-theoretic models. We emphasize that our predictions are indeed just binding arrangements, not how such arrangements are reached. The predictions can nevertheless be viewed as functions of the overall (nuclear) concentrations of proteins, the affinities of proteins to explore neighborhoods of individual sites, as well as the equilibrium constants characterizing the ability of proteins to bind to specific sites when in close proximity. Any perturbation of such parameters leads to a potentially different arrangement that we can predict. The game that we will now begin to characterize in detail provides the mechanisms for arriving at such predictions.

3 Mapping to Game Theory

To formalize the binding problem as a game we need to specify several key characteristics, discussed here informally at first. These include who the players are, available strategies to the players, and utilities specifying preference orderings over possible strategies. There are two types of players in our game, proteins and sites. A *protein-player* refers to a type of protein, not an individual protein, and chooses how its nuclear concentration should be allocated to the proximity of specific sites. Note that the protein-player is a game-theoretic expression of the transport process. The protein-players are assumed non-cooperative and rational. In other words, their allocations are based on the transport affinities and the availability of sites rather than through some negotiation process involving multiple proteins. The non-coopeative nature of the protein allocations does not, however, preclude the formation of protein complexes or binding facilitated by other proteins. Such extensions can be incorporated at the sites and will be discussed later on in the paper in the context of our empirical work.

Each possible binding site is associated with a *site-*

player. Site-players choose the fraction of time (or fraction of cells in a population) a specific type of protein is bound to the site. The site may also remain empty. The strategies of the site-players are guided by local chemical equilibria. Indeed, the site-players are introduced merely to reproduce this physical understanding of the binding process in a game theoretic context. The site-players are non-cooperative and self-interested, always aiming and succeeding at reproducing the local chemical equilibria.

The binding game has no global objective function that serves to guide how the players choose their strategies. The players choices are instead guided by their own utilities that depend on the choices of other players. For example, the protein-player allocates its nuclear concentration to the proximity of the sites based on how occupied the sites are, i.e., in a manner that depends on the strategies of the site-players. Similarly, the site-players reproduce the chemical equilibrium at the sites on the basis of the available local protein concentrations, i.e., depending on the choices of the protein-players. We provide quantitative definitions of the utilities in the next section.

The predictions we can make based on the game theoretic formulation are *equilibria of the game* (not to be confused with the local chemical equilibria at the sites). A game is at an equilibrium when each player, protein or site, is content with their current choice of strategies given the strategies of other players. Put another way, at an equilibrium players have no incentive to unilaterally deviate from their current strategy. Thus, at an equilibrium, no reallocation of proteins to sites is required and, conversely, the sites have reproduced the local chemical equilibria based on the current allocations of proteins. While games need not have equilibria in pure strategies (actions available to the players), our game will always have one.

Table 1 summarizes our mapping from biological concepts to game theory.

Biology		**Game Theory**
Type of protein "transport mechanism"	$\Leftrightarrow$	Protein-player "allocation to sites"
Binding site "chemical equilibrium"	$\Leftrightarrow$	Site-player "selection of who binds"
Binding arrangement	$\Leftrightarrow$	Equilibrium of the Game

Table 1: Conceptual mapping

4 The binding game

We abstract the notion of proteins and DNA binding sites by viewing them as rational agents or *players* competing non-cooperatively in a *game*. This allows us to build our model on solid mathematical ground by exploiting previous work and well-studied models from game theory and economics (see [8] and [2] for classical examples). We refer the reader to [5] for a more thorough introduction to game theory.

We proceed here to define players' strategies, their utilities, and the notion of an equilibrium of the game more formally. To this end, let f^i represent the (nuclear) concentration of protein i. This is the amount of protein that can be allocated to the neighborhoods of sites. The fraction of protein i allocated to site j is specified by p^i_j, where $\sum_j p^i_j = 1$. The numerical values of p^i_j, where j ranges over the possible sites, define a possible strategy for the i^{th} protein player. The choices of which strategies to play are guided by parameters E_{ij}, the affinity of protein i to explore the neighborhood of site j (we will generally index proteins with i and sites with j). The utility for protein i, defined below, provides a numerical ranking of possible strategy choices and is parameterized by E_{ij}. Each player aims to maximize its own utility over the set of possible strategy choices.

The strategy for site-player j specifies the fraction of time that each type of protein is actually bound to the site. The strategy is denoted by s^j_i, where i ranges over proteins with affinity to the site. Note that the values of s^j_i are in principle observable from binding assays (cf. [11]). $\sum_i s^j_i \leq 1$ since there is only one site and it may remain empty part of the time. The availability of site j is $1 - \sum_i s^j_i \leq 1$, i.e., the fraction time that nothing is bound. We will also use $\alpha^j = \sum_i s^j_i$ to denote how occupied the site is. The utilities of the site players will depend on K_{ij}, the chemical equilibrium constants characterizing the local binding reaction between protein i

and site j.

4.1 Utilities

The utility function for protein-player i is formally defined as

$$u_i(p^i, s) \equiv \sum_j p^i_j E_{ij}(1 - \sum_{i'} s^j_{i'}) + \beta H(p^i) \qquad (1)$$

where $H(p^i) = -\sum_j p^i_j \log p^i_j$ is the Shannon entropy of the strategy p^i_j and j ranges over possible sites. The utility of the protein-player essentially states that protein i "prefers" to be around sites that are unbound and for which it has high affinity. The parameter $\beta \geq 0$ balances how much protein allocations are guided by the differentiated process, characterized by the exploration affinities E_{ij}, as opposed to allocated uniformly (maximizing the entropy function). Since the overall scaling of the utilities is immaterial, only the ratios E_{ij}/β are relevant for guiding the protein-players. The strategies available for the protein-player i are given by the set

$$\mathcal{P}^i \equiv \{p^i : \sum_j p^i_j = 1, p^i_j \geq 0, \text{ for all } j\}.$$

The protein-player will always find a strategy that maximizes its utility over the set of possible strategies $\mathcal{P}^i$. Note, however, that since the utility depends on the strategies of site-players through how available the sites are $(1 - \sum_{i'} s^j_{i'})$, one cannot find the equilibrium strategy for proteins by considering s^j_i to be fixed; the sites will respond to any p^i_j chosen by the protein-player.

As discussed earlier, the site-players always reproduce the chemical equilibrium between the site and the protein species allocated to the neighborhood of the site. The utility for site-player i is defined such that the maximizing strategy corresponds to the chemical equilibrium:

$$\frac{s^j_i}{(p^i_j f^i - s^j_i)(1 - \sum_{i'} s^j_{i'})} = K_{ij} \qquad (2)$$

where s^j_i specifies how much protein i is bound, the first term in the denominator $(p^i_j f^i - s^j_i)$ specifies the amount of free protein i, and the second term $(1 - \sum_{i'} s^j_{i'})$, the fraction of time the site is available. The equilibrium equation holds for all protein species around the site and for the same strategy $\{s^j_i\}$ of the site-player. The units of each "concentration" in the above equation should be interpreted as numbers of available molecules (e.g., there's only one site). The utility function that reproduces this chemical equilibrium when maximized over possible strategies is given by

$$v_j(s^j, p) \equiv \sum_i s^j_i - K_{ij}(p^i_j f^i - s^j_i)\left(1 - \sum_{i'} s^j_{i'}\right) \qquad (3)$$

subject to the following constraints on the strategies

$$\begin{aligned} s^j_i &\leq K_{ij}(p^i_j f^i - s^j_i)\left(1 - \sum_{i'} s^j_{i'}\right) \\ s^j_i &\leq p^i_j f^i \\ \sum_{i'} s^j_{i'} &\leq 1. \end{aligned}$$

These constraints guarantee that the utility is always non-positive and zero exactly when the chemical equilibrium holds. $s^j_i \leq p^i_j f^i$ ensures that we cannot have more protein bound than is allocated to the proximity of the site. These constraints define the set of strategies available for site-player j or $\mathcal{S}^j(p)$. Note that the available strategies for the site-player depend on the current strategies for protein-players. The set of strategies $\mathcal{S}^j(p)$ is not convex.

4.2 The game and equilibria

The *protein-DNA binding game* is now fully specified by the set of parameters $\{E_{ij}/\beta\}$, $\{K_{ij}\}$ and $\{f^i\}$, along with the utility functions $\{u_i\}$ and $\{v_j\}$ and the allocation constraints $\{\mathcal{P}^i\}$ and $\{\mathcal{S}^j\}$.

We assume that the biological system being modeled reaches a steady state, at least momentarily, preserving the average allocations. In terms of our game theoretic model, this corresponds to what we call an *equilibrium* of the game. Informally, an equilibrium of a game is a strategy for each player such that no individual has any incentive to unilaterally deviate from their strategy. Formally, if the allocations $(\bar{p}, \bar{s})$ are such that for each protein i and each site j,

$$\bar{p}^i \in \arg\max_{p^i \in \mathcal{P}^i} u_i(p^i, \bar{s}), \text{ and } \bar{s}^j \in \arg\max_{s^j \in \mathcal{S}^j(\bar{p}_j)} v_j(s^j, \bar{p}_j), \qquad (4)$$

then we call $(\bar{p}, \bar{s})$ an *equilibrium* of the protein-DNA binding game. Put another way, at an equilibrium, the current strategies of the players must be among the strategies that maximize their utilities assuming the strategies of other players are held fixed.

Does the protein-DNA binding game always have an equilibrium? While we have already stated this in the affirmative, we emphasize that there is no reason *a priori* to believe that there exists an equilibrium in the pure strategies, especially since the sets of possible strategies for the site-players are non-convex (cf. [2]). The existence is guaranteed by the following theorem:

Theorem 1. *Every protein-DNA binding game has an equilibrium.*

The proof can be obtained either through Brower's fixed point theorem or, alternatively, on the basis of the algorithm we develop in the next section for finding equilibria of the game. In the interest of brevity, we will defer to the constructive proof provided by the algorithm.

The theorem guarantees that at least one equilibrium exists but there may be more than one. At any such equilibrium of the game, all the protein species around each site are at a chemical equilibrium; that is, if $(\bar{p}, \bar{s})$ is an equilibrium of the game, then for all sites j and proteins i, $\bar{s}^j$ and $\bar{p}^i_j$ satisfy (2). Consequently, the site utilities $v_j(\bar{s}^j, \bar{p}_j)$ are all zero for the equilibrium strategies.

5 Computing equilibria

The equilibria of the binding game represent predicted binding arrangements. It is therefore critical to be able to find equilibria on a genome-wide scale. While finding Nash equilibria of multi-person games is known to be hard, our game has special structure and properties that permit us to find an equilibrium efficiently through a simple iterative algorithm. The algorithm monotonically fills the sites up to the equilibrium levels, starting with all sites empty.

We begin by first expressing any joint equilibrium strategy of the game as a function of how filled the sites are, and reduce the problem of finding equilibria to finding fixed points of a monotone function. To this end, let $\alpha^j = \sum_{i'} s^j_{i'}$ denote site j occupancy, the fraction of time it is bound by any protein. α^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, then we can readily obtain the maximizing strategies for proteins expressed as a function of site occupancies:

$$p^i_j(\alpha) = \frac{\exp(E_{ij}(1-\alpha^j)/\beta)}{\sum_{j'} \exp(E_{ij'}(1-\alpha_{j'})/\beta)}, \tag{5}$$

where we view the maximizing strategies as functions of α. Similarly, at the equilibrium, each site-player achieves a local chemical equilibrium specified in (2). By replacing $\alpha^j = \sum_{i'} s^j_{i'}$, and solving for s^j_i in (2), we get

$$s^j_i(\alpha) = \frac{K_{ij}(1-\alpha^j)}{1 + K_{ij}(1-\alpha^j)}\, p^i_j(\alpha)\, f^i \tag{6}$$

So, for example, 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 $p^i_j(\alpha)$.

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

$$\alpha^j = \sum_i s^j_i(\alpha) = \sum_i \frac{K_{ij}(1-\alpha^j)}{1 + K_{ij}(1-\alpha^j)}\, p^i_j(\alpha)\, f^i \tag{7}$$

We have therefore reduced the problem of finding equilibria of the game to finding fixed points of the mapping $G^j(\alpha) = \sum_i s^j_i(\alpha)$. This mapping, written explicitly as

$$G^j(\alpha) = \sum_i \frac{K_{ij}(1-\alpha^j)}{1 + K_{ij}(1-\alpha^j)}\, \frac{\exp(E_{ij}(1-\alpha^j)/\beta)}{\sum_{j'} \exp(E_{ij'}(1-\alpha_{j'})/\beta)}\, f^i \tag{8}$$

has a simple but powerful monotonicity property that forms the basis for our iterative algorithm. Specifically,

Lemma 2. *Let α^{-j} denote all components α^k except α^j. Then for each j, $G^j(\alpha) \equiv G^j(\alpha^j, \alpha^{-j})$ is a strictly decreasing function of α^j for any fixed α^{-j}.*

We omit the proof as it is straightforward. This lemma, together with the fact that $G^j(1, \alpha^{-j}) = 0$, immediately guarantees that there is a *unique* solution to

$$\alpha^j = G^j(\alpha^j, \alpha^{-j}) \tag{9}$$

for any fixed and valid α^{-j}. The solution α^j also lies in the interval $[0, 1]$ and can be found efficiently via binary search.

5.1 The algorithm

We are now ready to define the algorithm. 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 proceeds as follows:

- Set $\alpha^j(0) = 0$ for all $j = 1, \ldots, m$.
- Find each new component $\alpha^j(t+1)$, $j = 1, \ldots, m$, on the basis of the corresponding $\alpha^{-j}(t)$ such that $\alpha^j(t+1) = G^j(\alpha^j(t+1), \alpha^{-j}(t))$
- Stop when $\alpha^j(t+1) \approx \alpha^j(t)$ for all $j = 1, \ldots, m$.

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 as discussed earlier. The algorithm generates a monotonically increasing sequence of α's that converge to a fixed point (equilibrium) solution.

5.1.1 The algorithm: analysis.

We provide here a formal convergence analysis of the algorithm. To this end, we begin with the following critical lemma.

Lemma 3. *Let α_1 and α_2 be two possible assignments to α. If for all $k \neq j$, $\alpha_1^k \leq \alpha_2^k$, then $G^j(\alpha^j, \alpha_1^{-j}) \leq G^j(\alpha^j, \alpha_2^{-j})$ for all α^j.*

The proof is straightforward and essentially based on the fact that α_1^{-j} and α_2^{-j} appear only in the normalization terms for the protein allocations and in these terms

$$\sum_{k \neq j} \exp(E_{ik}(1 - \alpha_1^k)/\beta) \geq \sum_{k \neq j} \exp(E_{ik}(1 - \alpha_2^k)/\beta)$$

as $\alpha_1^k \leq \alpha_2^k$ for all $k \neq j$. We omit further details for brevity.

On the basis of this lemma, we can show that the algorithm indeed generates a monotonically increasing sequence of α'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

$$\begin{aligned} \alpha^j(t+1) < \alpha^j(t) &= G^j(\alpha^j(t), \alpha^{-j}(t-1)) \\ &\leq G^j(\alpha^j(t), \alpha^{-j}(t)) \\ &< G^j(\alpha^j(t+1), \alpha^{-j}(t)) \\ &= \alpha^j(t+1) \end{aligned}$$

which is a contradiction. The second line follows from the induction hypothesis and lemma 3, and the third line derives from lemma 2 and $\alpha^j(t+1) < \alpha^j(t)$. □

Since $\alpha^j(t)$ for any t will always lie 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 $\bar{\alpha}$ such that $\bar{\alpha}^j = G^j(\bar{\alpha}^j, \bar{\alpha}^{-j})$ for all j.*

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

6 The λ-phage game

We use the well-known λ-phage infection [6, 13, 1] to illustrate the game theoretic approach. Viral infection by $\lambda-$phage is governed by a genetic two-state control switch that specifies whether the infection remains dormant (lysogeny) or whether the viral DNA is aggressively replicated (lysis). The components of the $\lambda-$switch are 1) two adjacent genes *cI* and *Cro* that encode cI_2 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_R1, O_R2, and O_R3.

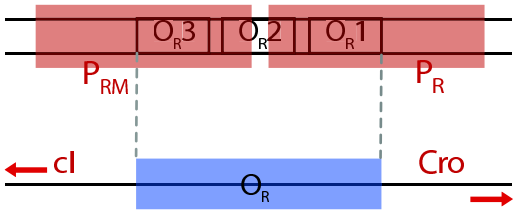


Figure 1: Representation of cI and Cro genes, promoters and operator sites as they appear in the λ chromosome. Arrows indicate direction of transcription. O_R1 overlaps with the *Cro* promoter P_R, and O_R3 overlaps with P_{RM}.

Figure 1 illustrates schematically the geometry of the switch. We focus on the lysogeny phase where cI_2 dominates over Cro. There are two relevant protein-players, RNA-polymerase and cI_2, and three sites, O_R1, O_R2, and O_R3. Since the presence of cI_2 in either O_R1 or O_R3 blocks the access of RNA-polymerase to the promoter region P_R, or P_{RM} respectively, we can safely restrict ourselves to operator sites as the site-players.

To fully specify the game we have to set the exploration affinities as well as the chemical equilibrium constants. The difficulty of this step arises from the ambiguity of affinities reported in the literature; affinities may refer to the activation energy of binding, energy of binding, or both. We begin by outlining the key considerations. RNA-polymerase can bind either promoter but does not bind O_R2. The affinity of cI_2 protein to bind the operator sites can be summarized as $O_R1 > O_R2 \simeq O_R3$. There are three phases of operation depending on the concentration of cI_2:

1. cI_2 binds to O_R1 first and blocks the Cro promoter P_R
2. Slightly higher concentrations of cI_2 lead to binding at O_R2 which in turn facilitates RNA-polymerase to initiate transcription at P_{RM}
3. At sufficiently high levels cI_2 also binds to O_R3 and inhibits its own transcription

The first two phases appear to contradict known affinities: cI_2 binds almost immediately at both O_R1 and O_R2 (in this order), and the presence of cI_2 in O_R2 results in increased transcription of cI_2. These effects

		Affinity for each site		
Protein	**P-P Inter.?**	O_R3	O_R2	O_R1
RNA-p	No	Low	none	High
	Yes	High	none	High
cI_2	No	O_R1's/10	O_R1's/10	High
	Yes	O_R1's/10	$=O_R1$'s	High

Table 2: Qualitative affinities (lysogeny).

are attributed to protein-protein interactions between cI_2 dimers at O_R1 and O_R2, and between cI_2 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_R1 and O_R3. Note that the spatial proximity of the sites places physical constraints on binding and that such constraints may be implicit in the affinities.

Table 2 summarizes the qualitative affinities with and without protein-protein interactions.

6.1 Game parameters

The game requires three sets of parameters: chemical equilibrium constants, affinities, and quantities of different protein species. We set the chemical equilibrium constants in accordance with the Gibbs' Free energies ΔG tabulated by [13],

ΔG (kCal)	O_R3	O_R2	O_R1
cI_2	-10.1	-10.1	-11.7
RNA-p	-11.5	0	-12.5

To incorporate these energies into our simulation, we have to analyze the units of the equilibrium equation carefully. Recall the chemical equilibrium equation (2),

$$K_{ij} = \frac{s_i^j}{(p_j^i f^i - s_i^j)(1 - \sum_{i'} s_{i'}^j)}.$$

To ensure the consistency of units, we have to define f^i as the total number of proteins i available, and arrange the units of K_{ij} accordingly:

$$f^i \equiv \widetilde{f^i}\ V_T N_A, \tag{10}$$

$$K_{ij} \equiv \widetilde{K}_{ij} \frac{1}{N_A V_S}, \tag{11}$$

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

$$\frac{s_i^j}{(p_j^i \widetilde{f^i} V_T N_A - s_i^j)(1 - \sum_{i'} s_{i'}^j)} = \widetilde{K}_{ij} \frac{1}{N_A V_S}$$

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

$$\begin{aligned} \frac{s_i^j}{(p_j^i \widetilde{f^i} \frac{V_T N_A}{V_S N_A} - s_i^j \frac{1}{V_S N_A})(1 - \sum_{i'} s_{i'}^j)} &= \widetilde{K}_{ij} \\ &= e^{-\Delta G/RT}, \end{aligned} \tag{12}$$

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

K_{ij}	O_R3	O_R2	O_R1
cI_2	.0020	.0020	.0296
RNA-p	.0212	0	.1134

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 [9, 10], summarized in Table 2:

E_{ij}	O_R3	O_R2	O_R1
cI_2	.1	.1	1
RNA-p	.2	.01	1

Note that the overall scaling of these values is immaterial; only the relative affinities will guide the protein-players. Note also that there is a ten-fold difference in cI_2 transport affinity to O_R2 and O_R1 since we have chosen not to incorporate any protein-protein interactions. The affinity of RNA-polymerase for O_R3 is similarly reduced. Whether RNA-p affinity to O_R2 is very small or exactly zero has no impact on the results.

p_j^i	f_{cI_2}/f_{RNA-p}	O_R3	O_R2	O_R1
cI_2	$\frac{1}{100}$	0	1	0
RNA-p		0.46	0	0.54
cI_2	2	0	.88	0.12
RNA-p		0.49	0	0.51
cI_2	10	0	.55	0.45
RNA-p		1	0	0
cI_2	100	0.32	.39	0.28
RNA-p		1	0	0

Table 3: Distribution of proteins near the sites. The numbers represent the fraction of proteins allocated to the sites.

We set $\widetilde{f}_{RNA-p} = 30nM$ (cf. [13]) which for a typical *E. coli* is equivalent to setting $f_{RNA-p} \simeq 340$ 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_{RNA-p}. We set $\beta = 10^{-5}$.

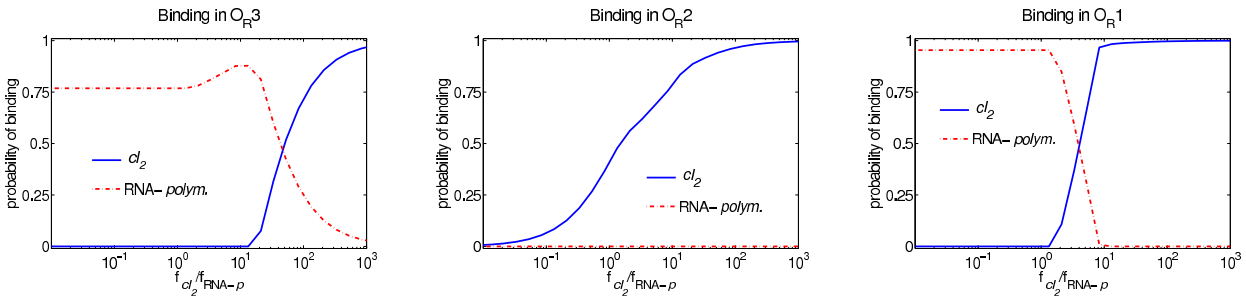
6.2 Simulation Results

The predictions from the game theoretic model 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 3 shows the distribution of protein expected near the sites at specific concentrations of cI_2. The amount of RNA-p is assumed to remain constant. Table 4 give the predicted fraction of time specific proteins are bound at the sites.

Tables 3 and 4, taken together, show 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 can only be explained by factors external



(a) O_R3 (b) O_R2 (c) O_R1

Figure 2: Predicted protein binding to sites O_R3, O_R2, and O_R1 for increasing amounts of cI_2.

s_i^j	f_{cI_2}/f_{RNA-p}	O_R3	O_R2	O_R1
cI_2	$\frac{1}{100}$	0	.001	0
RNA-p		0.77	0	0.95
cI_2	2	0	0.56	0.11
RNA-p		0.78	0	0.85
cI_2	10	0	0.83	0.98
RNA-p		0.88	0	0
cI_2	100	.78	0.97	0.99
RNA-p		0.19	0	0

Table 4: Probability of binding at each site.

to this analysis such as transcription time, or presence of additional proteins.

2. As more cI_2 becomes available, it accumulates near O_R1, O_R2, increasing the probability of finding cI_2 at either sites to nearly one.

3. Further increase of cI_2 leads to almost exclusive binding of RNA-polymerase to O_R3.

4. Finally, at high concentrations cI_2 blocks its own promoter and reduces RNA-p binding at O_R3. Figure 2(a) shows how increased cI_2 concentration sharply reduces RNA-polymerase binding at O_R3.

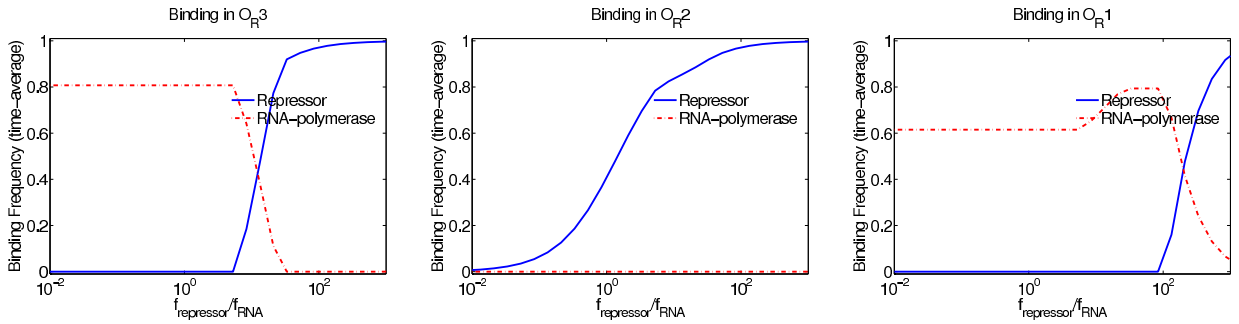
Figure 2 illustrates how the binding at different sites changes as a function of increasing f_{cI_2}. Although our model does not capture dynamics, and the figure does not involve time, it is nevertheless useful for assessing quantitative changes and the order of events as a function of increasing f_{cI_2}. Note, for example, that the levels at which cI_2 occupies O_R1 and O_R2 rise much faster than at O_R3. 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 O_R3 bumps up as the probability that O_R2 is bound by cI_2 increases.

6.2.1 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 2 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.

E_{ij}	O_R3	O_R2	O_R1
cI_2	.1	.1	01
RNA-p	.2	.01	.1
K_{ij}	O_R3	O_R2	O_R1
cI_2	.0020	.0020	.003
RNA-p	.0212	0	.0113

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 3 reproduces the qualitative behavior observed in knockout 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



(a) O_R3 (b) O_R2 (c) O_R1

Figure 3: Predicted protein binding to sites O_R3, O_R2, and mutated O_R1 for increasing amounts 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.

7 Discussion

We believe the game theoretic approach provides a compelling causal abstraction of biological systems with resource constraints. The model is complete with provably convergent algorithms for finding equilibria on a genome-wide scale.

The results from the small scale application are encouraging. Our model successfully reproduces known behavior of the λ–switch on the basis of molecular level competition and resource constraints, without the need to assume protein-protein interactions between cI_2 dimers and cI_2 and RNA-polymerase. Even in the context of this well-known sub-system, however, few quantitative experimental results are available about binding. Proper validation and use of our model therefore relies on estimating the game parameters from available protein-DNA binding data (in progress). Once the game parameters are known, the model provides valid predictions for a number of possible perturbations to the system, including changing nuclear concentrations and knock-outs.

Acknowledgments

This work was supported in part by NIH grant GM68762 and by NSF ITR grant 0428715. Luis Pérez-Breva is a "Fundación Rafael del Pino" Fellow.

References

[1] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected excherichia coli cells. *Genetics*, 149:1633–1648, August 1998.

[2] Kenneth J. Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3):265–290, July 1954.

[3] Z. Bar-Joseph, G. Gerber, T. Lee, N. Rinaldi, J. Yoo, B. Gordon F. Robert, E. Fraenkel, T. Jaakkola, R. Young, and D. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342, 2003.

[4] Otto G. Berg, Robert B. Winter, and Peter H. von Hippel. Diffusion- driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, 20(24):6929–48, November 1981.

[5] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, 1991.

[6] Ira Herskowitz and David Hagen. The lysis-lysogeny decision of phage λ: Explicit programming and responsiveness. *Annual Reviews Genetics*, 14:399–345, 1980.

[7] HarleyH. McAdams and Adam Arkin. Stochastic mechanisms in geneexpression. *PNAS*, 94(3):814–819, 1997.

[8] John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, September 1951.

[9] Mark Ptashne. *A Genetic Switch: Gene control and phage* λ. Cell Press AND Blackwell Scientific Publications, 3rd edition, 1987.

[10] Mark Ptashne and Alexander Gann. *Genes and Signals*. Cold Spring Harbor Laboratory Press, 1st edition, 2002.

[11] Bing Ren, Franois Robert, John J. Wyrick, Oscar Aparicio, Ezra G. Jennings, Itamar Simon, Julia Zeitlinger, Jrg Schreiber, Nancy Hannett, Elenita Kanin, Thomas L. Volkert, Christopher J. Wilson, Stephen P. Bell, , and Richard A. Young. Genome-wide location and function of DNA-binding proteins. *Science*, 290(2306), December 2000.

[12] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–76, 2003.

[13] Madeline A. Shea and Gary K. Ackers. The o_r control system of bacteriophage lambda. a physical-chemical model for gene regulation. *Journal of Molecular Biology*, 181:211–230, 1985.

[14] Neil P. Stanford, Mark D. Szczelkun, John F. Marko, and Stephen E. Halford. One- and three-dimensional pathways for proteins to reach specific DNA sites. *EMBO*, 19(23):6546–6557, December 2000.