ANALYSIS OF COOPERATIVE TRANSCRIPTION FACTOR BINDING AT THE SEQUENCE LEVEL

By

Jacob Clifford

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics

2015

ABSTRACT

ANALYSIS OF COOPERATIVE TRANSCRIPTION FACTOR BINDING AT THE SEQUENCE LEVEL

By

Jacob Clifford

Transcription Factor binding to DNA binding sites is one of the primary causes of gene regulation. A common representation of transcription factor binding sites is at the DNA sequence level, partly due to reoccurring patterns at the sequence level that occur throughout the genome for a given factor. The first chapter of this dissertation introduces gene regulation from the perspective of development. In addition the mathematical-physics foundation for performing calculations and for representations of the transcription factor binding sites at the sequence level is discussed in Chapter 1. In Chapter 2 I explore the possibility that two distinct sub-types of binding sites may co-exist within a population of functional sites. This leads to a model that can be used for prediction of transcription factor binding sites. In Chapter 3 I explore modelling of Dorsal Ventral early development Gene Regulatory Network, using the tools built up in Chapter 1 and 2, namely 'Position Weight Matrices' that allow for prediction of binding energies for genomic segments of DNA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

1.1 Mathematical Physics Introduction

This dissertation is a description of transcription factor binding that allows for prediction of the behavior of the Dorsal Ventral patterning gene regulatory network of Drosophila early development. Although Drosophila, the fruit fly, is much simpler than human, its molecular biology contains many similar to almost identical mechanisms for controlling genes, and hence is in the premiere league of modeling systems for understanding how human genes are molecularly controlled.

Genes, in a broad sense, are the particles that are passed on from parents to progeny that contain the information about the characteristics of the parent. The characteristics of the parents are their 'traits', like eye color, susceptibility to diabetes II, or fecundity (a proxy for 'fitness'). Hence knowledge of one's genes, or a population of people's genes, a gene pool, allows one to make predictions about what traits will exist in future generations.

The 'traits' of interest in this dissertation are not at the level of the adult, or even at the level of a recognizable animal. They are at the cellular level, where 'development' builds an adult by 'developing' different cell types that are arranged together to form an adult body plan. The initial steps of turning a totipotent cell (the zygote) into a ball of thousands of cells, the embryo, where each cell has its very own genome that becomes fated to be the brain, the heart, etc.. of the fly through gene regulation.

The aspect of gene regulation that I focus on is at the level of transcription. The first step in the 'central dogma', where DNA is copied to an RNA sequence. It is the control

of this that we will consider 'regulation', where control is in the sense of how many RNA 'transcipts' should be produced.

The reason we focus on transcription and on development is because 'development' produces in tandem with the production of RNA a set of controlled experiments. For example, the fly embryo produces sets of cells at gross anatomical positions, 'regions', where at each region the cells are all under the same conditions. Gross positions are like a binary north and south, top and back, or dorsal and ventral. Each of these regions provides a controlled experiment, allowing us to take advantage of the trusted reproducibility of animal development[45]. Hence each region in the embryo is under a strict set of controls, namely the physiological environment unique to that region (e.g. high or low doses of proteins and other factors); just as a scientist would set up a pitri dish full of cells under similar conditions in order to reliably cause gene expression. In a sense a colony of bacteria in a petri dish behave similar to the embryonic regions that fate certain tissue types. The effect of transcription allows for us to observe the emitted particles from a cell or from the genome, the transcripts or the proteins. It is these particles that allow for deciphering the mechanism of control of genes.

Maternally controlled molecules, controlled in the sense of being positioned at different locations of the embryonic shell, diffuse or are actively transported to the different regions of the embryo to eventually control specific genes in the genomes (in the cells) that will or does reside in that region. In early fly development the cells of the embryo, each with their own genome, form a monolayer around a spherical yoke, like corn on the cob. By observing both the 'morphogens' (transcription factors) and the emitted particles from the transcription factor's target genes, the mRNA and their translated proteins, we can decipher what is occurring at the gene regulatory level. The obvious inference is that the input molecules must somehow pass a message to the gene that is being controlled. The mechanism that I

study for gene regulation is where the input molecule 'binds' to a location specific segment of DNA, such as the DNA sequence 5'-GGAAAATCC, and thereby passing a message to the flanking sequence of the DNA 'binding site'.

Chemically, the message may begin by the bound protein adding a chemical motif to a protein that already was bound or wrapped up in the DNA like a 'histone'. This additional chemical motif may set a motion a cascade of further steps that ultimately lead to a clear modification of transcription levels.

The easiest message to observe is 'turn on' or 'turn off', which is seen through transcription, because we can easily observe proteins and RNA through common lab techniques. But broadly, regulation of genes by the 'binding' process means controlling the inheritable flanking sequences of the docking site [1].

How the morphogen binds to a location specific position of the genome, i.e. it finds the gene of interest, seems to be a complicated process. For example, random binding or sampling sites in the genome, in the time allocated during development, would not allow the morphogen sufficient time to find the target, even with the mass action of multiple morphogens.

To help understand how the binding process works, a central problem is understanding how the protein recognizes a specific location in the genome. By better understanding of how proteins or the morphogens recognize specific binding sites within the genome, problems such as the diffusion of maternal molecules or the active transport of the morphogens to a specific location in the genome may be better understood. In a broader sense, any cellular message that requires the modulation of transcription levels, will be better elucidated by a

[1] The controlled flanking sequence may not be a dogmatic 'gene' that encodes for a protein, it could be anything that is useful for the organism fitness, and is therefore selected by evolution. In this sense, the binding site itself is like a gene if it is under selection. Hence the molecular phenotype 'to bind' is expressed by the binding site sequence, the genotype.

well formulated understanding of the protein-DNA interactions, the recognition problem.

A central assumption in this work is that the recognition is encoded in the DNA. Hence the binding sites are more than just a surface upon which the protein deposits. Just as vapor deposition can be controlled by placing specific types of high affinity surfaces mixed with low affinity surfaces, so too, one could imagine the protein binding to regions of the genome solely due to high affinity surfaces that consist of material that is not DNA, such as binding to a histone or histone tails (histones are proteins always present that occupy a large fraction of the genome), or binding to other proteins that are already bound to the genome. I am solely interested in protein-DNA binding, and hence the differential affinity of the surface is only of interest for regions of the genome that have the DNA exposed and specifically protein-DNA binding that are under selection (i.e. it is functional DNA).

The natural mathematical physic's framework to discuss this problem, is through deposition of particles to a one dimensional lattice, a representation of the genome (see T.Hill's chapter on lattices for a general introduction to lattice statistics [53]). Hence, I will introduce the mathematical physics necessary for performing calculations of binding energies and occupation numbers of lattice sites. This machinery is very general, and is not specific to the recognition problem.

Once I have presented how 'binding' is represented I will introduce k-mers and the recognition problem, where we will account for the specific binding to ordered arrays of bases, sequences of DNA. This leads to applications in bioinformatic sequence alignment, where known binding energies to specific sequence of DNA can be used as a computational search for potential discovery of unannotated binding sites (i.e. not in a database) within a sequenced genome, and the inverse problem where known sequences of binding sites can be used to infer the binding energy. In particular, the first chapter of this dissertation will in-

troduce a 'mixture model', where a mix of binding sites for the same factor, Dorsal, are used for prediction of unknown sites. This is also used to explore the possibility of epistasis and physical cooperativity between Dorsal and cooccurring Twist binding sites. This interaction is encoded in Dorsal binding sites, where the cooperatively encoded binding sites form one component of the 'mixture model'. Given the foundation of protein-DNA recognition.

Chapter 2 considers further the prediction of unknown sites for the trio Dorsal, Twist, Snail; three of the dominant 'morphogens' or transcription factors in Dorsal Ventral patterning in early development of Drosophila. Here I explore the possibility that recognition is a function not only of the preferred k-mer sequence for a given factor, but also depend on different locations of the embryo (such as the neuroectoderm, or mesoderm) which contain different concentrations of these factors and therefore their recognition to specific sequences of DNA in those regions of the embryo is modulated, which is manifested by the differential expression of their target genes. Here I also explore the prediction of unknown sites as a function of the spacer between co-occurring k-mer sites. This is important in Drosophila, where it has been extensively documented as the primary mechanism of 'repression' utilized by Snail, a so-called short range repressor factor that turns genes off that would otherwise be activated by the activator transcription factor Dorsal that is in high concentrations in the ventral location of the embryo. Furthermore, this function that predicts binding sites as a function of the spacer also is explored in terms of the cooperativity between Dorsal and Twist factors, both known activators, that are known to act synergistically when their k-mer binding sites co-occur with a specific 'window' of spacer values (e.g. the sites must be about 2 to 30 base-pairs from each other).

The DV network of enhancer's occupancy for these factors is calculated as a subproblem in the optimization of a gene expression model for the DV network of genes that is location

specific within the embryo (thereby accounting for location specific concentrations). The input to the model is a two-dimensional profile of the concentration of the transcription factors along the Dorsal-Ventral position axis of the fly embryo, along with the corresponding profiles for the target genes whose mRNA expression is modulated due to regulation by the factor binding. The binding is accounted for by additionally inputting the cis-Regulatory Modules, or flanking sequence of DNA near the target genes where the factors are known to bind. The model contains unknown constants that are fit using root mean square error for the objective function F, where $F = \sum_i \sum_j (M_{ij} - O_{ij})^2$, here M is the model output for a given evaluation of the objective function (i.e. for a given values of the parameters) and the observed data O, where i runs over the positions of the Dorsal Ventral axis, and each j is a particular gene in the network. In addition, the objective function has an option of running a multi-objective that will add an additional objective to fit the occupancies of the factors to Chip-Seq or Chip-Chip data for the trans-factors, hence in the multiobjective case one has $F = \sum_i \sum_j (M_{ij} - O_{ij})^2 + \sum_k \sum_l (< N_{kl} > - I_{kl})^2$, where the new term calculates the model occupancy $< N_{kl} >$ for the genomic segment k for transcription factor l, which was observed with intensity I_{kl}.

1.1.1 Phase space, a real vector space

Classically one can calculate the statistical properties of many body systems, such as a gas with Avogodro's number of particles, by transforming Newton's second order differential equation for each particle to two first order differential equations for each particle, and then, upon solving the equations of motion, one can calculate time averages of quantities of interest. Another approach is through Hamilton's principle, which is a principle of parsimony, dictating that motion follows a path of 'least action' (the shortest path, where 'shortest' has

a special formulation). This approach also consists of two first order differential equations for each positional degree of freedom, and consist of constructing the 'Hamiltonian' of the system, which is effectively the total energy of the system (at least for systems that we are interested in). For example, for an isolated one dimensional system (such as a stretched out genome) of M particles (or M genomic units) one has:

$$H = \sum_{i}^{M} \frac{P_i^2}{2 * m_i} + U(X_1, X_2, \ldots, X_M) \tag{1.1}$$

H is the Hamiltonian, i indicates the particle label, and P is the momentum[2], and U is the potential energy of the system due to the interactions of the particles, which is a function of each particle's location X.

The 2M random variable joint distribution for the isolated system describes the occupancy of one point in phase space at any particular instant ($X_1 = x_1, X_2 = x_2 \ldots X_M = x_M, P_1 = p_1, P_2 = p2 \ldots P_M = p_M, t = 0$). Given some time t has elapsed, the distribution will be found to occupy some other point, ($X_1 = x_1(t), X_2 = x_2(t) \ldots X_M = x_M(t), P_1 = p_1(t), P_2 = p2(t) \ldots P_M = p_M(t)$), this is a delta distribution for each particle's position and momentum[103]. Hence this is a real vector space in R^{2M}. This construction is an example of the microcanonical ensemble. By loosening the isolation constraint (i.e. allowing energy of the system to vary), we have the canonical ensemble which has nonzero variance for many of the random variables.

If we label our phase space points with an index i, then the occupation of point i in phase space, n_i. could be normalized by the occupation of all states (points in phase space), we

[2]We use capital letters for the momenta and coordinates, since we will think of those as random variables, where we can translate Hamilton's deterministic equations to the Kolmogorov Chapman equations leading to a deterministic equation of motion for the joint distribution, a master equation, for example see page 10 of Van Kampen[103].

would find:

$$\frac{n_i}{n} = \frac{\exp \frac{-H_i}{kT}}{Q} \tag{1.2}$$

Here H_i, is the Hamiltonian evaluated at the phase space point i (just plug in the corresponding positions and momentum of each particle for that state into the Hamiltonian). The number of phase space points is determined by how well we can resolve our subspaces for each particle. For example, the configuration space, the vector space over the position coordinates X, would be meshed no finer than our ability to resolve different distances. Here, Q is the partition function, which can be shown by the maximum entropy principle to equal:

$$Q = \sum_i \exp -\frac{H_i}{kT} = \sum_i \exp -\frac{\sum_j^{2M} \frac{P_j(i)^2}{2*m_j} + U(X_1(i), X_2(i) \dots X_M(i))}{kT} \tag{1.3}$$

By using the size of a unit of the genomic biopolymer as the unit of our length scale for each particle's position subspace, we can mesh out phase space's 'configuration space' such that each base at each position along the polymer chain occupies a given location (mesh point) in 'configuration space'. The method to mesh out momentum space is irrelevant, since we are about to 'project out' or 'marginalize out' that portion of phase space.

By asserting that the biopolymer is stretched out, and its length and hence number of 'unit's is fixed, we can reduce configuration space to contain M mesh points (one for each unit of the polymer). This is a 'one dimensional' lattice, where the dimensionality is now in reference to the fact that each unit of the polymer lies along a linear array, like a thin spaghetti noodle that was still solid and was notched for each unit of the polymer.

We are interested in 'binding', where a distinct specie like a type of transcription factor 'binds' to the lattice. This requires more complexity, as we now must introduce more particles

than the original M units of the polymer. Before we introduce more particles, we will first get rid of the momenta. We define a reference system with no interactions (U=0), which has a corresponding partition function Q_o, then we would expect that for small interactions (U is small) the velocity distribution (Maxwell Boltzmann distribution), would effectively remain invariant. Hence we have an effective partition function, q, defined as:

$$q \equiv \frac{Q}{Q_o} \approx \sum_i \exp -\frac{U(X_1(i), X_2(i) \dots X_n(i))}{kT} \tag{1.4}$$

classically this is called the configuration integral. This is effectively a 'projection' or marginalization over the momentum coordinates.

Thermodyanmics

Our interest is not in isolated physical systems, as the M particle system above, rather we are interested in closed (energy is exchangeable, *but* particles are not exchangeable) and open (energy is exchangeable, *and* particles are exchangeable) systems. The Victorian founders of the field of open and closed systems, of thermodynamics, were contemporaries with Charles Darwin, which is very fitting seeing that their insights into heat and particle exchange (a form of work) supplies the essential mechanics to describe biological systems.

Conservation of energy for isolated systems is a consequence of Newton's second law. A more interesting statement is that the energy of a closed system in equilibrium can not spontaneously change, this is the statement of the first law of thermodynamics. Hence we must do work or heat the system to change its internal energy[3].

[3]The Clausius convention for the form of the first law: $dU = q - w$ = heat supplied to the system (q) - work done by the system (w)

$$\begin{aligned} dU &= q - w \\ dU &= TdS + PdV \end{aligned} \tag{1.5}$$

Here U is the internal energy, the internal energy of the system is the Hamiltonian, H, hence $H = U$. In (1.5) q is the heat (q has nothing to do with the partition function, which uses the same symbol), and w is the work (the work also accounts for particle exchange processes).

The system above only has one type of particle, for systems that allow transcription factor to bind to DNA segments it is necessary to introduce multiple particle species (the protein and DNA). By Gibbs phase rule, we know that for a binary system we will need four independent variables, one of which is intensive. Hence we have many options available for constructing binary particle ensembles. For the species that can vary particle number the natural variable is the chemical potential, while for systems that are closed the natural variable is just the particle number. Hence, for a binary system, composed of M components of S' (sequences) and N components of P' (protein), we could define the open open system with the following coordinates $(\mu_{S'}, \mu_{P'}, V, T)$; while the closed-closed system has the following form:

$$dU = TdS + \mu_{S'}dM + \mu_{P'}dN + PdV \tag{1.6}$$

We will think of binding sites interacting with proteins as solute solute interactions that are occurring in a solvent. The nucleoplasm represents a very complicated solvent, a mixture

in the liquid phase of all sorts of complicated biopolymers and water molecules and the many other inorganic substances in a biological nucleus. Of course, the protein that must find a specific binding site within the genome, must compete with all the other molecules in the nucleus for occupying any point in the spatial grid of the nucleus. However, again, my aim is to represent and describe the recognition of a particular binding site by a protein, hence I would like to make progress specifically on the recognition problem, without being hampered by the solvent properties. Hence, I will introduce an argument in T.Hill's text[52], and also discussed by Landua[63], the so-called 'dilute limit'. This will allow us to rigorously account for the fact that the protein and DNA are embedded in a phyiological solvent, while, to a large degree, hides the solvent molecular details and the solvent-solute details by introducing an effective molecular partition function q, which will be similar to Eq.**??** in two ways: first because the momentum will not be of interest and second because we will be taking ratio of two types of well-defined systems to define the 'effective' partition function. However, q is different than **??** in that a molecular partition function is about one type of particle (one molecule), where for large systems like a gas of n identical and distinguishable molecules, one commonly denotes the partition function as $Q = q^n$, hence lowercase is to denote 'one molecule', and capital case to denote large systems. By defining the effective partition function we will then proceed to the relevant problem of the solute-solute (protein-binding site) interaction, where the physiological environment (solvent) will be accounted for in both the DNA and protein by defining an effective partition funtion for each solute (e.g. q_{DNA} and q_{pro}).

Solvent Solute, dilute limit The solvent solute system is just a binary system. Here we will follow Hill's approach, which is to start off by organizing the liquidey solution with the

two component grand canonical system[4]

$$\exp\frac{PV}{kT} = \Xi(\mu_S, \mu_P, V, T) = \sum_{N_1}\sum_{N_2} Q(N_1, N_2, V, T)\exp\frac{(N_1\mu_1 + N_2\mu_2)}{kT} \tag{1.7}$$

Here we have relabelled S as 1 and P as 2, and denoted the two-component canonical ensemble as $Q(N_1, N_2, V, T)$[5]. Now we know the average solute particle number:

$$< N_2 >= \sum_{n_2} n_2 P(n_2) = \lambda_2\frac{\partial \log \Xi}{\partial \lambda_2} \tag{1.8}$$

By taking the derivative with respect to the absolute activity of the solute, $\lambda_2 = \exp\frac{N_2\mu_2}{kT}$, we can calculate the average number of solute particles. To take this derivative, first notice that in the dilute limit of solute, the summation over the solute may as well be neglected, since the solute activity approaches zero. If we define the pure solvent grand partition function as $\Psi_o = \sum_{N_1} Q(N_1, 0, V, T)\lambda_1^{N_1}$, and the solvent grand partition function $\Psi_1 = \sum_{N_1} Q(N_1, 1, V, T)\lambda_1^{N_1}$ as the grand for the solvent with one solute embedded in it[6]. Then the derivative will appear as:

$$\lambda_2\frac{\partial \log \Xi}{\partial \lambda_2} = \lambda_2\frac{\Psi_1 + 2 * \Psi_2\lambda_2 + 3 * \Psi_3\lambda_2^2 \ldots}{\Psi_0 + \Psi_1\lambda_2 + \Psi_2\lambda_2^2 + \Psi_3\lambda_2^3 \ldots} \tag{1.9}$$

[4]The grand canonical ensemble is related to the thermodynamic grand potential, by the legendre transform of Eq.1.1.1, the transform leads to: $dU = TdS + < M > d\mu_{S'} + < N > d\mu_{P'} + PdV$

[5]In the case of noninteracting solvent and solute $Q(N_1, N_2, V, T)$ factorizes into two 'ideal gas' ensembles, namely: $Q(N_1, N_2, V, T) = \frac{Q(N_1,V,T)^{N_1}}{N_1!}\frac{Q(N_2,V,T)^{N_2}}{N_2!}$, while for interacting solvent and solute the interaction energy is contained in the potential energy of of the Hamiltonian, and hence $Q(N_1, N_2, V, T)$ can not be factorized. Regardless of this interaction we can proceed to the 'recognition problem' and construction of an effective partition function being aware that explicit representation of the solvent solute interaction will require writing a potential between the solvent-solute in the Hamiltonian of Eq.1.3, which are details we wish to hide.

[6]Note that Ψ_1 will posses a solute solvent interaction inside of $Q(N_1, 1, V, T)$ only if the solute interacts with the solvent, otherwise $Q(N_1, 1, V, T)$ is simply $Q(1, 0, V, T)^{N_1}Q(0, 1, V, T)$ for the case that the solvent particles are identical and distinguishable.

As $\lim_{\lambda_2 \to +0}$, we find that:

$$< N_2 >= \frac{\Psi_1}{\Psi_0}\lambda_2 \tag{1.10}$$

Hence, just as in Eq.**??** where we defined the effective partition function, q, the configuration integral, here again we will define an effective partition function for a solution system. Hence, for a solute immersed in a solvent we will define the effective partition function for the solute as:

$$q(N, V, T) \equiv \frac{\Psi_1(\mu_1, N_2 = 1, V, T)}{\Psi_0(\mu_1, N_2 = 0, V, T)} \tag{1.11}$$

We can approximate the grand canonical ensemble partition functions on the right hand side of the above equation by using the maximum term of the partition function:

$$q = \frac{\Psi_1(\mu_1, N_2 = 1, V, T)}{\Psi_0(\mu_1, N_2 = 0, V, T)} \approx \frac{Q_m(< N_1 >, 1, V, T)}{Q_m(< N_1 >', 0, V, T)} \lambda^{-(<N_1> - <N_1>')} \tag{1.12}$$

here, $Q_m(N_1, 1, V, T)$ is the canonical partition function for one solute in a box of size V of solvent particles, where the subscript m indicates we have taken the largest term of the grand canonical ensemble. The maximum term's particle number is denoted as N_m, where $N_{m_1} =< N_1 >$ due to the Gaussian property that the maximum occurs at the expected value (where $< N_1 >'$ is a slightly different maximum for the constant volume case of no solute molecules (because the now evacuated space leaves more room for solvent to fill). The average solvent particle numbers are nuisance variables that we can, in a sense, transform out of the effective partition function. Hence instead of using the two component grand canonical system we will use the two component system with fixed pressure and fixed solvent particle

number[7]. Hence the new partition function is

$$\Xi'(N_1, \mu_2, p, T) = \sum_{N_2} \sum_{V} Q(N_1, N_2, V, T) \exp \frac{-pV}{kT} \exp \frac{N_2 \mu_2}{kT}, \tag{1.13}$$

where the summation over the volume only works for discrete physical spaces, such as lattices (see Eq.2.23 of Hill[52]). Now if we repeat the exact steps above from Eq.1.8 to Eq.1.12 using the new partition function (Ξ'), we will arrive at:

$$q = \frac{Q_m(N_1, 1, V_m, T)}{Q_m(N_1, 0, V'_m, T)} \lambda^{-p(V_m - V'_m)}, \tag{1.14}$$

where again we have used the maximum term method isolating the canonical ensembles that have the largest probability in the partition functions of Ψ'_o and Ψ'_1, where, for example the pure solvent grand partition function in this case is $\Psi'_o = \sum_V Q(N_1, 0, V, T) \exp^{pV/kT}$, where we denote the volume from the maximum term in this summation as V_m, and similarly V_m' is the maximum term in the partition function (which sums over volumes) of solvent with one solute particle present[8].

[7]For example, imagine a beaker filled with a fixed number of liquid water molecules, then if we drop a rock (one solute molecule) inside the beaker the water will rise - i.e. the volume will change); this is unlike the grand canonical case, where we have to dispose of the excess water molecules displaced by the rock (those that don't fit in the volume V), since we must keep the volume fixed (albeit this could be done by allowing the volume of interest to be the beaker filled to the brim, then any excess water molecules will simply flow over the rim; but we're interested in a simple mathematical tool, not a simple experimental design.

[8]We can think of $-p(V_m - V'_m)$ as the mechanical work w done by inserting a solute into a solvent, namely:$w = P\Delta V$, for example see figure 1.1 of Hill [52]. Furthermore, noting that the Helmoholtz is the free energy of the canonical $Q = \exp(-A/kT)$, we see that $q = \exp(\Delta A + p\Delta V) = \exp(\Delta G)$.

1.2 Derivation of the chemical potential: $\mu = \mu_o + \ln(c)$

On page 6 of T.Hill's text[52], he derives the dilute limit formula for the chemical potential $\mu = \mu_o + \ln(c)$ using a vacuum as a solvent. After his argument he points out that the derivation he presented could and does take different forms (depending on the text and purposes[9] however, he states that for a solute solvent system (biological system), that this approach (which I'll now recapitulate) is necessary (if you want to keep things simple).

In Eq.1.10 we have the absolute activity of the solute, which by definition is $\lambda_2 = \exp\frac{\mu_2}{kT}$, hence using the definition of the activity and along with 1.10, the average solute particle number, we can derive the standard formula for the thermodynamic chemical potential. First we must define the concentration c (particle density) as:

$$c = \frac{< N_2 >}{V}, \tag{1.15}$$

Now dividing both sides of Eq. 1.10 by V, we can rewrite Eq. 1.10 in terms of the density, which results in:

$$c = \frac{< N_2 >}{V} = \frac{\lambda_2}{V}\frac{\Psi_1}{\Psi_0}. \tag{1.16}$$

Now plugging in $\exp\frac{\mu_2}{kT}$ for λ_2, and taking the logarithm of both sides of Eq. 1.19 results in:

$$\log c = \log \frac{\exp\frac{\mu_2}{kT}}{V}\frac{\Psi_1}{\Psi_0}. \tag{1.17}$$

[9]For example, given that $dG = Vdp + SdT$, then for a constant temperature ideal gas system we have $G = G_O + \int V dp = G_O + \int \frac{RT}{p} dp$, where we have used the gas constant R in the ideal gas law in the last expression. This results in $G = G_O + \ln\frac{p}{p_O}$, where p_O is the reference (standard state) of one bar, which is usually omitted, furthermore using $G = \mu N$, we can divide the equation for dG through by N, now after integration this results in $\mu = \mu_O + \ln p$, but using the ideal gas law (and keeping in mind that we omitted the reference pressure) we can rewrite this in terms of a reference concentration of one particle per unit volume, resulting in: $\mu = \mu_O + \ln c$.

Now let us define q based on the right side of Eq.1.14, then if we rewrite the right side of Eq.1.17 as two terms we have: $\log c = \log \frac{\mu 2}{kT} + \log \frac{q}{V}$. Upon rearranging terms, and multiplying through by the thermal energy kT, and by writing $kT \log \frac{q}{V} = \mu_o$ (which acts as a reference or a standard state)[10], we have our desired result:

$$\mu = \mu_o + kT \ln (c), \tag{1.18}$$

Nucleoplasm genome ligand binding problem

The binding site is the main component of our physical system, we will let the number of binding sites be fixed in the genome (i.e. the system is closed with respect to number of binding sites). Let M be the number of binding sites in the genome, each site being of the same energy. Let the system be open with respect to factor binding. Hence, each particular locus (each site) is not just either bound or not bound, rather it will have an occupancy. In equilibrium, we can define the equilibrium binding constant as a function of the concentrations of the components of the system.

The change in free energy per particle, $\Delta\mu$, of the binding process is zero in equilibrium, recall each species in each phase has its own chemical potential:

$$\mu = \mu^o + \ln c, \tag{1.19}$$

here μ^o is the reference energy (standard state), and c is the concentration or density of the

[10]This statistical mechanics approach is in contrast to the thermodynamic approach of the footnote above, where one integrates $dG = Vdp$ *from* some standard state (an arbitrary reference) to a desired *point* (in thermodynamic space, with dimensions like pressure). Here the *point* we reference is, in a sense, a point in phase space (a finer level of detail than the coarse grained thermodynamic variables.

chemical specie relative to standard concentration of '1' in the units of interest, hence we also have:

$$\mu_{SP} - \mu_S - \mu_P = 0 \tag{1.20}$$

now if we group common standard states and concentrations, and rearrange:

$$\mu^o_{SP} - \mu^o_S - \mu^o_P = \ln(\frac{[SP]_e}{[S]_e[P]_e}) \tag{1.21}$$

Here the subscript e on the concentrations is to remind us that the concentrations are no longer a variable, but fixed by the equilibrium constraint. Our chemical potentials are linked to the molecular energies through the logarithm of the dilute limit partition function of Eq.1.14 (if the system is in equilibrium, hence we also have:

$$\mu^o_{SP} - \mu^o_S - \mu^o_P = \ln(\frac{q_{SP}}{q_S q_P}) \tag{1.22}$$

Here we will assume that $p\Delta V$ factors from Eq.1.14 that arise from the effective molecular partition functions from the righ side expression of Eq.1.22 all cancel. This is because the pressure is constant, and we will assume the volume of the complex SP is roufly the additive volume of the molecules S and P in isolation in the solvent. Now we see that the binding energy emerges from the ratio of partition functions, hence, we define a new partition function as:

$$q = \frac{q_{SP}}{q_S} = q_P \exp -\frac{E_b}{kT}. \tag{1.23}$$

Here, the binding energy, E_b is equal to the work done to separate the bound complex protein and DNA (denoted as the SP particle). It is the solute-solute interaction. It can

also be thought of as the energy to lift an adsorbate out of the potential well of depth E_b that describes the influence of the sequence on the adsorbate, or it could be thought of as the parameter σ in the pairwise potential of a Lennard Jones (the depth of the LJ potential). It determines the potential energy term $U(X_S, X_P)$ that we would have added to our Hamiltonian in equation (1.1). The emergence of E_b by taking the ratio of the effective partition functions is a consequence of the assumption that the molecular degrees of freedom, such as rotation and vibration are unperturbed by the binding process. For example, for the molecule S, we have $q_S \approx q_r^S q_v^S$, similarly for the molecule P. The complex SP contains all of these molecular states too, however the complex also contains an additional factor due to the interaction (such as an LJ potential). Assuming the complex is stable, then we can assume we are at the minima of the pair-wise potential, which we call the binding energy[11]

Linking the statistical mechanic's partition functions to the thermodynamic binding constant we have:

$$K = \exp -\frac{E_b}{kT}. \tag{1.24}$$

Experimentally this can be determined by binding titration curves, which allow one to transform the binding constant as a function of the fractional occupancy. As a consistency check,

[11] An example of the cancellations of the partition functions: Let q_r^S be the rotational partition function over the eigenvalues of the Hamiltonian for the rotational degrees of freedom, e.g. $q_r^S = \sum_i \exp(H_i^S)$, where i runs over the eigenvalues of the Hamiltonian for the S molecule, similarly for the other degrees of freedom (all the variables are assumed classical, hence we can work in a real vector space, as opposed to a complex vector space). Then $\frac{q_{PS}}{q_P q_S} = \frac{q_r^P \prod_d q_d^P q_r^S \prod_f q_f^S \exp(-U)}{q_r^P \prod_d q_d^P q_r^S \prod_f q_f^S}$, where d and f run over all remaining 'degrees of freedom' for the molecules S and P, where the form of each degree of freedom's Hamiltonian will determine the eigenvalues and hence the partition functions (the 'momenta' and 'position' random variables of the Eq.1.1 are seen as 'degrees of freedom'[54] in this context, hence the variables of Eq.1.1 can be seen as generalized coordinates in phase space, where the random variable X, for example, may represent a rotation). Whatever the form of these Hamiltonians, all of these partition functions cancel if they are unperturbed when P and S form a complex or 'bind', and all that remains is the interaction between S and P denoted as U, which at equilibrium has a value E_b.

we see that if the binding energy is zero (no interaction between sequence and protein, then the concentration of the bound complex is just as likely as the unbound complex, while complete binding requires the binding energy to be negative infinity, and for particles that repel such that the bound complex never forms the binding energy must be plus infinity), then we have:

$$q_{SP} = q_S q_P. \tag{1.25}$$

Hence, we find that partition functions behave almost identically as joint distributions. The beauty of partition functions, is that we maintain the molecular link to the Hamiltonian, and a link to thermodynamics. The above analysis lays the foundation for understanding the behavior of a transcription factor (protein) that binds in a solution with identical sequences (DNA oligos for example). One can imagine the solvent as the oligos, and use standard partition functions, or one can work in a frame where the oligos themselves are another solute particle (like the protein), where both solutes are bathed in a solvent like water or milk. I now extend the analysis to the case of a protein binding to a single site within the genome. First, we will make the observation that for a genome that contains an array of identical binding sites, the problem then effectively reduces to a protein binding to a solution of oligos.

For now we will assume that the n sites are independent of one another, hence, we can work with a system of just one site, and realize that to extend the system to all n sites, simply requires scaling the free energy by n, and raising the partition function to the power of n. Hence, although each individual site will have fluctuations between being bound and unbound, we can use the n sites as effective data to increase the power of our statistics for learning about the binding energy.

1.2.1 One binding site

For the case that the genome can be modeled as n identical binding sites, we can construct a system with n fixed sites where the binding protein number is allowed to vary, (open closed system):

$$\Xi = \frac{\xi_o^n}{n!} \tag{1.26}$$

Here, Ξ is the grand canonical partition function for the n sites. The independence of the sites means we can simply work with just the grand canonical partition function for a single binding site ξ_o, where the factorial is due to the indistinguishably of the n sites. Hence, we can work with a single site system, and simply note that extensive quantities (such as the binding energy) will simply be multiplies of the single site system (e.g. the binding energy of 10 bound proteins is simply ten times larger than the binding energy of single protein, and the one dimensional volume (i.e. length) of one site simply increases by a factor of 10 for ten sites).

The single binding site is fixed (closed) while the adsorbate is open, hence single site partition function is:

$$\xi_o = q_P + q_{SP}\lambda \tag{1.27}$$

here the q's are effective partition functions for solutes in solvent (the dilute limit). We can renormalize the partition function:

$$\xi = 1 + q\lambda \tag{1.28}$$

Here q is the effective partition function of the bound complex, $q = q_{SP}/q_P$, and c is the concentration of free adsorbate. Clearly ξ_o and ξ are different numerically. However, for relative probabilities the forms are irrelevant. Hence the occupancy (relative probability

between bound to unbound) is:

$$P_b = \frac{q\lambda}{1 + q\lambda}. \tag{1.29}$$

The absolute activity ($\lambda = \exp(\mu/kT)$) contains the chemical potential that is equal to the potential of both the free protein (the protein floating around in the nucleoplasm) and the the protein that is bound on the site (i.e. that is in our system). This is because in equilibrium, the chemical potential of the reservoir of particles (free protein) must equal the chemical potential of the bound protein. This is simply the definition of equilibrium. If the potentials are unequal, which certainly occurs in development, then there will be a net flux into or out of our system (the binding site), until the potentials equilibrate. Utilizing the fact that the potentials of the reservoir and system are equal gives us two potential equations, one in the form of the a controllable parameter (the free protein with concentration c_P that is related to the potential of Eq.1.19) and another in the form of partition function of the grand canonical ensemble (i.e. solve for the potential in Eq.1.29), relating these allows for us to rewrite the occupancy as:

$$P_b = \frac{c_p K}{1 + c_p K} = \frac{c_{PS}}{c_P + c_{PS}}, \tag{1.30}$$

which is utilized in chapter 2 of my dissertation for calculations of the occupancy of factors on DNA binding sites. For further details on this topic see also chapter 2 of Hill[52].

1.2.2 Two dependent binding sites

For a system with two identical independent binding sites we have:

$$\xi = 1 + 2qc + q^2c^2 \tag{1.31}$$

We will be interested in cooperativity between the two bound adsorbates, a dependency, which will modify the above function to:

$$\xi = 1 + 2qc + yq^2c^2 \tag{1.32}$$

Here y is the exponential of the work required to perform the following reaction 10 + 01 = 11 + 00, where 00 is the unbound unbound configuration etc.. This process requires no energy unless there is an interaction between the adsorbates. Using Hill's formalism we have in general:

$$\xi = y_{11} + qc(y_{10} + y_{01}) + y_{22}q^2c^2 \tag{1.33}$$

For example, $y_{22} = y$, and contains cooperativity for bound protein-protein interactions, while y_{11} refers to an interaction that occurs between the two binding sites (an interaction that occurs in the configuration 00).

1.2.3 A genome of n dependent binding sites

For n binding sites, where the the bound proteins interact, there are a total of 2^n different types of possible interactions that may be accounted for. For the case that only nearest neighbors interact, we have $\frac{\xi^n}{n!}$, where $\xi = 1+yqc$ such that y contains the interaction energy. For the case that all n sites are different, yet there are still nearest neighbor interactions of bound factors we have: $\prod_i^n \xi_i$, where $\xi_i = 1 + y_i q_i c$.

1.2.4 Highly correlated systems, the k-mer and recognition problem

We've been treating sequences S, as if they were simply particles. DNA sequences consist of units of bases: A,C,G,T. Proteins, like transcription factors, may prefer one base over another, and in general may prefer a specific ordering of specific bases, for example AAAT may not be equivalent to TAAA Notice these are not genetic complements (e.g. AAAT complements ATTT, where I will always write DNA sequences in the 5' to 3' direction).

Rather they are mathematical permutations of one another. For the case that one considers a k-mer, a binding site that contains k consecutive component sites that are all bound or all unbound, the k component sites can be aggregated into just one binding site (since they're completely correlated). This is the form I use for the representation of transcription factor binding sites, where each component of the k sites represents a DNA base[12]. For example, a specific 3-mer of DNA, is AAA. And rather than constructing a closed open system for the DNA and adsorbate for each component of the sequence, we rather construct a closed open system for the aggregate. Hence the configurations for the closed open system would simply be bound or unbound; identical to the problem of a single binding site with variable number of adsorbate. Another example is two dimers. For example, AA and AA, which consist of the sequence AAAA. This is simply considered as two binding sites, and hence has four configurations 00,10,01,11, where the 01 configuration indicates the first two bases of AA are unbound, while the last dimer is bound by the adsorbate. Hence this can be treated identically to how two binding sites were treated above.

An additional complexity will be to not only introduce each base as having a specific

[12]This is a form widely accepted possibly due to natural selection acting at the units of the bases (which are roughly the chemical functional groups)

binding energy (so 4 distinct binding energies), but each base *within* the k-mer as having a specific binding energy. Hence the binding energy will be based on a function of 4*k possible energies. This means that for a lattice of k sites, we treat each site independently in terms of their binding energies, yet in terms of the binding to the k-sites, the sites are completely correlated. Hence the binding energy of a k-mer S to a specific adsorbate is:

$$E_b = E(S) = \sum_i^k E(S_i), \tag{1.34}$$

and the binding constant for the k-mer to the adsorbate is:

$$K(S) = \prod_i^k K(S_i). \tag{1.35}$$

This complexity can be increased by considering a hierarchy of possible internal interactions, or cooperativity within the binding site, such that the top of the hierarchy has 4^k possible energies. This hierarchy is explored commonly in the interdisciplinary literature through different probabilistic models of sequences.

In 1987, Berg and von-Hippel (BvH) introduce their evolutionary selection model of protein DNA regulatory sequences, which effectively unites the idea of the highly correlated binding problem to Multiple Sequence Alignment. Staden, three years earlier, had introduced the idea of making a table to organize the count data from a MSA. At the time, Multiple Sequence Alignment was an emerging field (Simth Waterman's local pairwise alignment was only invented 3 years earlier), Blast doesn't appear until 1990, and the first named Hidden Markov Model applied to sequences, according to Sean Eddy, is 1994. The k-mer binding sites in the 1987 BvH paper, were modeled by what are called a Position Weight Matrices,

PWMs, which would eventually be recognized as a trivial HMM in Eddy's text, see chapter 5, where the PWMs are called Position Specific Scoring Matrices, PSSMs.

1.2.5 K-mers and PWM binding constants

We know that we can treat the binding of a protein to a k-mer using standard thermodynamics, for example we have:

$$K_o = \frac{[PS_o]}{[P][S_o]} = \exp -E_b/kt, \tag{1.36}$$

where the o symbol indicates a 'reference'. We can use the binding to highest affinity binding sequence (the 'concensus' sequence) as a reference point. From the perspective of Hill's perturbation theory (see page 15 of T.Hill[52]), we see that we could imagine perturbing this system by mutating the underlying sequence of the reference:

$$K(S) = K_o \exp -E(S)/kt \tag{1.37}$$

Here we have treated the sequence mutations (or differences between S and S_o) as a perturbation from the reference. Physically this is hard to imagine experimentally in real time for a bound protein-DNA system, as we're talking about changing just part of the genetic material bound to the protein while keeping in tack the bulk of the binding site. However, since the end result of the process is 'path independent', it is irrelevant the method used to cause the perturbation, hence the perturbation may even be an evolutionary mutation of a binding site.

Assuming that each position within the binding site is independent, we can then construct

a table of all the single mutation perturbations away from the concensus, thereby allowing us to estimate binding energy for all possible k-mers. This table contains the matrix elements of the so-called energy Position Weight Matrix, discussed more in Chapter 1, which is used in computational algorithms that 'search' for binding sites for transcription factors.

1.2.6 Single binding site protein systems in distinct environments c and u

We can further imagine perturbations to the binding energy due to the environment of the binding site. For example, if we have two distinct environments, c and u, we could construct two distinct binding constant tables, where the binding constant for any given sequence in environment c would be:

$$K(S)^c = K_o^c \exp -E(S)^c/kT, \tag{1.38}$$

and similarly in envirnoment u:

$$K(S)^u = K_o^u \exp -E(S)^u/kT \tag{1.39}$$

In environment c, we imagine all 4^k sequences S (all DNA k-mers) binding constants being measure to be $K(S)^c$. Similarly, in environment u, we imagine all 4^k sequences S binding constants being measured to be $K(S)^u$. Statistically we are assuming $P(S|c) \neq P(S)$, and similarly for environment u, where P(S) is the probability that sequence S is bound by the adsorbate when all possible 4^k k-mer sequences compete for binding with the adsorbate. Hence P(S) is the occupancy of sequence S and the adsorbate normalized by the sum over all possible occupancies of the 4^k k-mer sequences, where the occupancy is calculated for all

k-mer sequences under the same concentration of the adsorbate.

These environments could be considered at the cis-level, that is at the level of the genome. Hence, the environments c and u, could be determined by whether or not a cooccurring binding site is near the sequence S or not near. For example, one could imagine the environment c is due to cooperative interactions that have evolved between the cooccurring binding sites, while environment u is due to uncooperative or just plain independent binding to the cooccurring binding sites. We explore this problem in detail in chapter 1, where we consider the possibility that Dorsal binding sites have evolved as a mixture of distinct motifs due to a 'c' and 'u' cis-environment acting as a selective force to maintain to the component motifs. This is an example of 'epistasis' where multiple genes (i.e. the cooccurring binding sites) are selected for jointly (i.e. the sites are not evolutionarily independent.)

An example of two k-mer binding sites system using a mixture

Now we can imagine two k-mer binding sites adjacent on a lattice (there may be intervening nonspecific lattice sites between the two k-mer sites that act as 'spacers'), the important point is that we label two disinct locations (each of length k) on the lattice to be sites. For example two dimers separated by a spacer: $S' = AANNNNCC$, where AA is the first dimer, and CC is the second dimer, and NNNN is a spacer of DNA that the adsorbate does not recognize as a binding site. The two k-mer sites each bind distinct adsorbates, for example the adsorbate Dorsal binds AA and the adsorbate Twist binds CC. We can *explicitly* account for lateral interactions between the sites using the 'y' factors previously introduced. The lateral interaction can be adsorbate-adsorbate or sequence-sequence.

For example, there may be a sequence-sequence interaction between the particles AA and

TT. This can be accounted for by:

$$y_{11} = \exp(w(S')) = \exp(E(S)^c - E(S)^u) = E(AA)^c - E(AA)^u, \tag{1.40}$$

where we have introduced the energy terms from above that were from the environments, c and u. Here y_{11} is an exponential factor that contains the cooperative energy w(S') that is function of the underlying sequence S', and is defined by our equation (1.33), and is further defined on page 100 of Hill[52] for generalized binding site systems. If there is a sequence-sequence interaction, then the binding energies $E(S)^c, E(S)^u$ may be different, although it is not a necessary condition for a sequence-sequence interaction to manifest itself in this form.

An *implicit* form of cooperativity at the sequence-sequence level is to just use $E(S)^c$ for the case that Dorsal binding site k-mer cooccurs with a Twist binding site k-mer. This form contains the binding energy to the sequence along with a shift in the binding site energy due to the interaction with its cis-environment. The shift being modeled as a Kullback Leibler divergence between the cis-specific environment relative to the case that the case that the binding site is independent of its environment.

1.3 Biological Introduction

1.3.1 Tree of Life and the Theory of Life

The tree of life is both the organizing structure of life sciences (like it's predecessor- Linnean classification) and is a representation of the theory of life - evolution. On a short time-scale it is a picture of biological reproduction - or unfaithful cloning - a pedigree. On long time-

scales, the tree is a representation of the process of species evolution, which in many ways can be thought of as a pedigree of species, where each node represents a population of a particular specie, and descent down the branches represents time, and division of a node into two nodes (reproduction in a standard pedigree - parents having children..) represents speciation events from a common ancestor (reproduction at a population level on geological time-scales).

1.3.1.1 Comparative Anatomy and Physiology

Previously,in the eighteenth century, Linnean by brute force clustering techniques organized life according to anatomical similarities, leading to a classification system for life. This organization was not just a database that organized collections of living objects by comparative anatomy; the organization explained why the objects were different or why they were similar because comparative anatomy begs the question of 'function', of physiology, and physiology is a theory of life. In this sense, Linnaen systematic organization of life was a simple theory of life, in that it did explain life, as does all physiology, because it explained the purpose (function) of each anatomical feature[13]. For example, the purpose of a leg is locomotion, of a jaw is to bite, of a root is to stabilize a standing plant (among other functions), of a circulatory system (heart) is to circulate nutrients throughout the organism, of a stamen is for sexual reproduction etc.; and Linnaeus knew these trivial relationships between structure and function, which undoubtedly helped in his grouping of organisms. Interpreting the organization of life through a theory of structure and function is very powerful, as a theory intentionally simplifies complex patterns so that they can be understood and comprehended.

[13]In biology the 'function' of a trait or anatomical structure is what the trait does or accomplishes, hence the word 'purpose' seems to be a synonym to function. However, this may be confusing seeing that evolution has no 'purpose', hence 'purpose' should not be interpreted as if the population or lineage that evolved the trait had foresight and intended its creation.

Hence the theory of 'structure and function' has one of the most important aspects of biological theory, and that is to simplify life in a way that we can comprehend its rich diversity. It also has an obvious predictive power, for example if an animal loses its legs you can predict that it will lose locomotion. Although many of Linnean groups are based on reproductive anatomical features, a 'structure and function' theory is very short-sighted in terms of how life reproduces, as it can only explain how to maintain existing population of species (by having the reproductive structures do what they do). What's clearly missing is the origin of life, how to make life from basic physical chemical principles; and the origin of all the different kinds or species once basic life forms (i.e. specie meaning a group that can form viable offspring).

The theory of evolution by Charles Darwin, which I'll summarize in two pieces 'descent' and 'modification' added more structure to the theory of life. Darwin recognized (hypothesized) that common anatomical structures were due to common 'descent', indicating common features need not be derived anew possibly using different ingredients each time, the entire structure was passed on during reproduction. He also recognized that different structures between two groups of animals was due to 'modification' from a common ancestor between the two groups of animals.

1.3.1.2 Classical evolution

Darwin's theory in the late 1800s united life through one common lineage. However the biological mechanisms of how to produce life from molecules (e.g. how are completely novel and complex features derived in the first place) was not possible during Darwin's time, as the observational tools such as Leeuwenhoek's microscope seemed to be insufficient. Furthermore to show that one specie of organism could be related to another (through the common

ancestor) would require an evolution experiment that takes longer than one's lifetime (usually). Albeit, as suggested by Darwin's book's title, The Origin of Species, showing that one specie could be related to another was precisely his point.

In parallel with Darwin's work, Mendel's evolution experiment with peas would set in motion the molecular basis of evolution, and set the stage for two contributions from molecular biology that explain the molecular basis of inheritance. Before the advent of molecular biology, from a reductionist point of view, the proof (evidence) of the evolutionary tree of multicellular organisms can be thought of as a hierarchy of conservation of features or modules. The importance of conservation can not be stated enough, as Darwin's central tenant is 'descent', meaning common features between our ancestors and ourselves are due to conservation from descent: inheritance. At the highest level of modularity (and by far the most useful for a big picture resolution of the tree of life for multicellular organisms) is the anatomical structures (e.g. leg, eye, heart, nervous system). Comparative embryology during Darwin's time elucidated that all these anatomical structures from organs to gross features like a leg (containing multiple organs, like skin) are derived from possibly just three tissue types (germ layers). For example, the heart is derived from the mesoderm tissue, and all multicellular organisms without a mesoderm in their embryonic stage do not develop a heart. These three tissue types are composed of largely undifferentiated cells and during development they work together and sometimes work independently to fate the array of different cell types that make up the anatomical features that make up multicellular organisms (such as epidermal cells, hematopoietic cells, blood cells). Hence, before molecular biology (before the 'gene' picture), there was at least three basic conserved traits that could be used for evidence on resolving tree branching in multicellular organisms (anatomical features, germ layers, cell types).

1.3.1.3 Modern Synthesis of evolution

Molecular genetics, the first contribution from molecular biology, would show animals posses genes that are contained in chromatin and that those genes were passed on from parents to children during reproduction, this culminated in the 'modern synthesis' of evolutionary biology in the early 1900s. This would gain further support by R. Franklin's crystallization of a chunk of chromatin, DNA. DNA was found to be a polymer strand that would complement with another self assembling strand, which Francis and Crick saw could serve as a 'copy' for a replicating cell's progeny cell. The 'modern synthesis' is largely about the molecular dynamics of populations, 'population genetics', a population of organism's genomes (the frequency of genotypes) and how those change in time (in units of generations), and how they can be influenced by natural selection, mutation, gene flow, migration and isolation (E. Mayer type speciation-allopatry). Most significantly, the gene centred picture that had arisen in evolution now had given a fourth feature or module that was conserved: the gene, which encodes for a protein.

Comparative genomics

With the advent of fast 'sequencing' technology that can determine the genetic sequence of whole genomes, genetic conservation (by simply comparing genomes) has become a powerful tool in helping resolve the various branches of the evolutionary tree. With the knowledge of the vast array of processes of genome evolution (how genome's changes in time) we infer the enormous genomes of complex organisms like humans that share genes in common with the much smaller bacterial genomes, share these in common due to inheritance, through the process of natural selection that balances the randomizing forces of mutation, as generation of genes de novo (by random mutation) is far less likely (probabilistically) than the probability

of inheriting these modules (genes) (e.g. a gene of length 1000 nucleotide (hence about 333 amino acids) has $4^{1000} = 2^{2000} > 10^{300}$ possible sequences, which means an organism that reproduces 1000 times per year would still require vastly more time than the time available since the Big Bang to generate one gene (see J.Maynard Smith's chapter one[74]).). The cause of the genetic inheritance through natural selection can be due to standard reproduction along a lineage (i.e. meiosis and sexual reproduction), such as the human lineage, or it can be due to more exotic forms of homology (inheritance), where viruses or transposable elements can insert genetic material in host genomes resulting in genome expansion, where the newly inserted genetic material may, over time, become a beneficial resource to the host. Perhaps, the most significant source of genome evolution is gene duplication and whole genome duplication (the human lineage is thought to have had two whole genome duplications since the evolution of the deuterstomes - the 2 rounds (2R) hypothesis[14]) that can have enormous effects such as those seen in the mustard family of plants. These duplication events cause each original gene to have a copy whose purpose is in a sense unfated, and hence is free to evolve a new function. Hence although humans have about 25,000 genes, many of these are just slight variants of one another -a gene family - inherited through duplication events. Hence, the archibacteria that originated about four billion years ago, evolved gene templates (a basic design) of some of the gene families through natural selection for genes necessary to process the oxygen depleted environment, similarly their descendent cyanobacteria evolved process necessary for photosynthesis, thereby transforming the atmosphere to be amenable for evolution of gene networks in aerobic respiration found in bacteria. It can then be interpreted that most of our genes (the genes found in eukaryotes) that we share in

[14]This 2R hypothesis is supported by whole genome analysis between vertebrates and invertebrates. In particular the HOX genes as seen on four chromosomes occur on just one chromosome in the fly.

common with bacteria are inherited from the bacteria. Hence gene phylogeny has become the standard technique to resolve (determining the branching order) of the tree of life. Of course, it has been shown that a gene phylogeny almost always agrees with an anatomical grouping (such as Linnaeus groupings), because genes encode the anatomical structures[15].

An obvious question about gene templates (gene families) is how does one infer that a particular gene locus in human (for example) correspond to the same locus in say a fish. The gene of interest in the two organisms may be the direct descendent in each lineage since the most recent common ancestor of the organisms, or it may be the indirect descendent in each lineage due to gene and genome duplications or viral and transposable element insertions - this is the distinction of orthologs and paralogs[17]. Of course if the genomes are from a recent common ancestor, like chimp and human, it's relatively easy, since the 'order' of the genes is preserved (synteny), hence the genomes (chromosome) roughly match or align from start to end (barring the five largescale inversions and the famous fusion of two of the chimp's 24 chromosomes forming chromosome 'number 2' in human that has two centromeres and two additional telomeres at the fusion site of the chromosomes, hence humans have 23 chromosomes.). However, for more distantly related organisms, such as fish and human, this is more difficult due to chromosome inversions and the many indels since the most recent common ancestor.

[15]One must still be careful using just anatomical features for determining the topology of the tree, as 'mimics' commonly seen in the insect world in predator prey interactions to deceive their opponent clearly demonstrate that anatomical features can be evolved independently[16]. However, internal anatomy structures like the central nervous system are much stronger anatomical features for determining evolution, since these internal features are very complex and much harder to evolve de novo, and harder to co-opt due to high levels of epistasis and pleiotropy.

[17]Convergence indicates the organism generated the feature anew, which is of course a possibility for any material being obeying the laws of physics, but is so much less likely than inheritance that whenever inheritance can be demonstrated possible this is the most probabilistic origin.

Expression patterns

The answer is that we need to know where in the body (where in the anatomy) is the gene being expressed. If we are interested in a protein in the brain (say *chordin*), we would expect that the homolog (here i use homolog to mean direct common ancestry - so i mean ortholog) of that gene be expressed in the insect brain (say *dpp*). This expectation is in a sense obvious, as you may know through your own experience with food and taste that different organs and tissue of an animal or plant 'body plan' taste different due to the different proteins in those organs (of course some tissues are full of sugars and lipids possibly overwhelming any taste from the proteins). If the genes are indeed coexpressed in the same tissue, we next need to know if the gene functions the same way, is its protein being used for the same purpose. Given that the gene is coexpressed in the same tissue and functions the same, this, by parsimony, would suggest the genes are orthologs rather than paralogs.

How do we see 'where' a gene is expressed in a body (like is it expressed in the brain, heart or muscle etc.)? This question could be answered partly by classical genetics (breeding using Mendel's laws) by determining if a known genotype resulted in embryonic lethality (or when in development the fetus or larva died, or what environmental pressures would 'induce' death -presumably because of the mutant gene). However, using classical genetics alone to determine where a gene is expressed in a body is extremely tedious and indirect and has major limitations and requires a lot of interpretation. A much faster technique would be a visual tag on the gene's protein or mRNA (such as a fluorescent marker or dye), albeit classical genetics is still required to control for the correct allelic version of the gene of interest among other reasons (like genomic background). This microscopic technology became available in the 1980s and revolutionized are understanding of evolution by using

the technology coupled with time-shots of development, thereby also telling us 'when' a gene is expressed during an embryo's development.

1.3.1.4 The generalization of the theory of evolution; developmental genetics

Prior to the 1980s, embryologist and developmental biologist had carefully observed the different stages of development of many multicellular organisms. Maps were made of the different cell types that would be produced from the single fertilized egg, a cell lineage, as was mapped in detail for the first time in the development of the nematode. These cell lineages that were derived from the fertilized egg, and then germ layers, would mark the differentiation of cells (or cell specification). By determining 'where' a gene was expressed in a developing animal embryo it became evident that the process of cell differentiation, was linked with co-expression of whole sets of genes, where at each step of differentiation new sets of genes were being expressed (and other turned off)[18]. Hence development, in a large part, can be seen as a result of 'gene regulation'.

The second part of molecular biology that supports macroevolution, and possibly an extension of the 'modern synthesis', is **developmental genetics**. Developmental genetics would show that master genes (transcription factors) when their regulatory targets or binding sites evolved, then whole developmental networks (i.e. the genes necessary to build an

[18]Plant development is different than animals. Plants ubiquitously display 'phenotypic plasticity', which in development is called 'developmental plasticity'. Plant development is not a reproducible if the environment of the plant changes (e.g. change the temperature or light and the plant will respond). Plants respond to their environment, while animals do not, as S.Gilbert puts it: some organims are ruled by a tyrant genome, where the phenotype plays the passive permissive role of the ruled while other organims (like plants) are ruled by the environment, where the phenotype plays the passive role of the ruled. Animals have largely evolved plasticity as part of their development, this is a little different than homeostasis. Hence, 'developmental plasticity' seen in plants is incorporated into animal development. Animals have stably generated the *internal* environmental cues to turn genes on and off, where a plant may develop roots if cued by low light levels, an animal has put the cues for development as part of its internal system. For example, animals always develops say a neuroectoderm and mesoderm because the environmental cues that are a part of the embryonic environment are stable (such as in mother's womb that sets up concentration gradients of key transcription factors, or the egg casing of fly or chicken) and are processed by components of the genome (e.g. CRMs)[44]

anatomical structure) could be redeployed to a different position of a developing body (e.g. the position that expressed the activator master gene), or lead to modifications of current body parts[19].

Developmental genetics to large degree is the field of study that shows at a molecular level how whole anatomical features evolve, and how the diversity of anatomical features and their arrangement (body plans) has evolved. This is because the detailed molecular mechanisms of how animals develop has revealed the simplicity of anatomical scale evolution (sometimes crudely called 'macroevolution'). Developmental mechanisms are observed in the field of developmental genetics through powerful experiments such as transplantation experiments, and gain and loss of functions experiments, along with the powerful tools of genetics and microscopy techniques. In a sense developmental genetics, elucidates the 'Gene Reulatory Networks' that form the causal coarse grained molecular basis of self assembling anatomical features.

1.3.2 Development

1.3.2.1 The origin of multicellularity; the evolution *of* development

About a billion or two years ago, single celled algae started to cooperate by developing cell cell interactions, forming the first eukaryotic multicellular organisms, like algal mats. Such constructs eventually lead to the important innovation of multicellular organisms that is found in animals (and other kingdoms): sexual reproduction through meiosis and fertiliza-

[19] As suggested by the phrase, 'developmental genetics', this would seem to be a sub-discipline of 'molecular genetics' and hence not an extension of the modern synthesis, but rather a refining. However, the modern synthesis was a gene centred theory; it was about gene's that encode for proteins. Developmental genetics is a gene 'regulatory' theory, it is about the parts of the genome that turn traditional genes on and off; and gene regulatory elements (transcription factor binding sites) are not really apart of the modern idea of a gene. A gene encodes for a protein. A genetic regulatory element encodes for something altogether different.

tion. Like endosymbiosis, which is possibly the origin of eukaryotes from bacteria, where two cells would merge and partner to share their particular genes and hence traits (which may be different traits), in sexual reproduction cells not only merge, or fertilize (merging of two gametes into a zygote), but they first go through meiosis, which is significantly different from the results of endosymbiosis (which replicate through mitosis of each fused component) significantly different due to the 'recombination' of traits, or crossing over, leading to great diversity in progeny, thereby leading to faster evolution (by Fisher's Fundamental Theorem) and preventing Muller's ratchet (by creating gamete's free of detrimental mutations). Furthermore, the fusion of two genomes (e.g. the two gamete genomes) is a form of 'gene duplication', which is a source in 'gene families', where the copied gene leads to diversity in function of family of genes[20].

These early sexual reproducing algae are the origin of eukaryotic early development (which I define as eukaryotic cellular interactions that lead to a multicellular structure, such as an algal mat (a plain of cells), or a 'blastula' (a ball of cells)[105]). They are modern representatives, living fossils, of the evolution of multicellularity.

In the diverse domain of eukaryote the most famous groups are multicellular organisms, due to their gross anatomical features, in these multicellular organisms there is a remarkable common or conserved early development, yet there is also marked differences suggesting that multicellularity has independently evolved in plants and animals and fungus.

Examples of multicellular precursors to multicellularity can be even seen in bacteria in

[20]Two bacterial cells could fuse leading to a n $\rightarrow$ 2n genotype, similar to diploidy, and allowing for greater genomic diversity as one of the duplicate copies of the gene are now free to serve a new function. However, this genome duplication event is distinct from sexual reproduction (and hence is not diploidy, in my opinion), as the new bacteria with 2n genes, can be said to now have simply n' genes (a haploid with n' genes), when the n' genes are replicated (through mitosis), the progeny are clones (if mutation rate is slow enough), which distinguishes it from sex, in sexual reproduction the progeny are not clones due to recombination (assuming the mutation rate is high enough that there exists some genetic polymorphisms in the population, such that the cross over pieces are not identical).

quorum sensing (chemical communication/signalling) and the primitive fungus yeast through their form of sex using 'mating types' (mating 'types' are analogs of males and females)[105]. An array of protists (eukaryotes that are not animals, plants, or fungus) show multicellularity, notably volvox and slime molds. All these 'primitive' organisms are good starting points for the study of multicellularity.

'Development', to a large degree, focuses on multi-cell types - cellular differentiation (a skin cell 'functions' differently than a germ cell and functions different than a stomach cell that is a digestive cell that can 'eat' absorbed surroundings). That is the division of labor through specialized cells, which is not seen in many primitive organisms, rather these simple organisms are displaying 'colonies', which are the aggregates of cells due to mitosis resulting in progeny cells being proximal to one another (which is physically important in development, but it doesn't display the division of the aggregate into different cell types).

The starting point of 'development' is the innovation of meiosis, invented by the protists, and passed on to its progeny lineages of plants, animals, and fungus. Hence, plants and animals don't derive sex anew for themselves, they were the benefactors of it from a protistan ancestor. By the union of meiosis with fertilization (the opposite of meiosis, in a sense) the ability to always have an extra copy of a gene, and therefore the ability of a gene to evolve a new function becomes a stable component of life forms (regardless of whether some life forms display a 'haploid dominant' life form, like bryophytes in plants (e.g. moss and liverwort), where the plant one normally sees is the haploid, this is because most of the bryophytes' *life cycle* exist in the haploid stage).

The great lineages of animals, fungus, and plants all develop from the fertilized 'egg' (the fusion of the two gametes). 'Egg' means oocyte here, one of the gametes, while in development literature, egg may also mean the structure that encases a baby, like a chicken

egg. Initially in evolution, possibly, there was no distinction between gametes, egg and sperm, both gametes were equal in form, just haploid cells, like in yeasts. Regardless of the origin of meiosis, we see here a clear case of different cell types and an abstract case of the division of labor at the cell level, the emergence of multi-cell types (which, again, is different than simply cell aggregation). The division of cell types for the first multi-cell type organisms is speculative, but it seems the haploid cell's role in the life cycle of the first meiotic cells was to provide a means to generate unique progeny that were not clones of the parents (through recombination).

Cell aggregation in the form of colonies or even in the form of complex structures (the rudiments of a body plan) such as seen in the protist volvox or the 'transitional form' between fungus and animals in the organism choanoflagellates possibly evolved independently of meiosis (which i consider the origin of different cell types). Cell aggregation is caused by gene products that connect cells to together such as cadherins, actin, and microtubles. Hence one would suspect that the cell adhesion genes would be conserved among the multicellular lineages, however these great multicellular lineages of plants animals and fungus do not conserve some of these genes, such as receptor tyrosine kinase, and hence they have each evolved development 'by scratch' (by convergence), which suggests that the origin *of* development can not simply be tied into the origin and evolution of body plans (plants have body plans in the form of features like roots, shoots, and leaves; which can be laid out in many ways. Are plant body plans related to the layout of animal body plans (like head, thorax, and tail)? For an excellent developmental comparative anatomy between plants and animals see Alberts[4],demonstrating similarity of germ layers to the fundamental tissues in plants: epidermal cells, ground tissue cells, vascular tissue cells. Independent evolution of body plans between plants and animals and fungus does not mean we should not bother

comparing their different life cycles and body plans or compare their anatomies and physiology. The potential fact that these great lineages are all independent simply means that the powerful inferences (predictions) that can be made based on common ancestry are not valid (since their common ancestry is irrelevant if each one independently evolved their anatomical structures, hence the only thing constraining their differences would be basic physics and chemistry and the constraints imposed by descending from their common protistan ancestor). However, their potential multicellular independence is not as independent as some may suggest, there is still the deep homology in that they all still use meiosis and they both must deal with the complexities of transcription in the face of chromatin. Albeit, homology in meiosis may be of limited help in understanding their origins because plants have a 'cell wall' that is rigid, unlike animal's permeable and agile cell membrane, suggesting that the production of gametes, and the form of the gametes themselves is vastly different between plants and animals (and indeed pollen and how it 'grows' into the female stamen leading to the ovum is very different than sperm fertilizing an egg in animals).

1.3.2.2 Fly Development

The coarse-grained in space and time molecular dynamics of how a fly is self assembled from a maternally laid egg, a single cell, is coarsely understood at the molecular level thanks to developmental genetics. This does not explain or prove anatomical evolution, however if gain or lose or modification of a master genes (which encode transcription factors acting in early development) or of gene regulatory binding sites that interact with the master genes did result in modification or gain of loss of anatomical features, then this would convincingly explain the evolutionary mechanism of anatomical features[21]

[21]This type of huge leaps in modification (leaps) of an organism was observed in the fossil record by Jon Conway (among others), a palaeontologist, which he called the Cambrian Explosion, and is in contrast

In the 1950s J. Monod working with bacteria discovered that transcription factors can influence gene expression by activating or repressing a gene, where the transcription factor itself was activated by environmental cues (such as sugar). This would mark the start of the field of gene regulation, which would be the central theme in developmental genetics.

In *Drosophila* development it is known that after fertilization of the egg and after cleavage (mitosis without a G phase (cells don't grow)) within the blastocyst maternally laid transcription factors (such as Bicoid) zygotically regulate other transcription factors (such as gap genes (which are transcription factors) which in turn regulate pair-rule genes (which are transcription factors) which in turn regulate effector genes (like HOX genes some of which are transcription factors)) that ultimately feed in to signalling networks and paracrine factors that lead to the construction of anatomical features.

The initial maternally laid transcription factors are not ubiquitously expressed throughout the egg chamber, rather, like a Fourier series, the first maternal transcription factor protein form a coarse pattern across the embryo, where the protein concentration is like a square wave of concentration as a function of space that forms a term in a Fourier series (i mean discrete summation of a few signals). This transcription factor protein may act alone to activate a gap gene (like *hunchback*) or may act with another maternally laid transcription factor whose pattern is also like a square wave (with a phase shift), thereby *adding* another term to the series. The addition of the two input signals results after a bit of time in an additional new pattern across the embryo in the form a gap gene's protein concentration. Hence as time progresses more complex patterns appear as the gap gene's interact (primarily

(contrast does not mean conflict or contradictory) to the ubiquitously accepted 'gradual accumulation' of beneficial mutations that results in adaptations between particular species of organisms (this was even known by plant and animal breeders before Darwin's theory). The idea of saltation was called 'punctuated equilibrium' by Steven Jay Gould, where periods of gradual accumulation are punctuated by bursts of major anatomical feature novelty.

like *addition* of waves in a Fourier series) resulting in patterns like a sine wave (where, again, the amplitude is the amount of protein at a particular time and the horizontal axis is a spatial axis of the embryo (the wave is in space not time), where the embryo's axis length is fixed in time). Hence, after cleavage in development, the totipotent cells of the blastocyst are in a sense transforming into more specialized cell types, where the cell is defined by the amount of each specific gene product's protein concentration at a particular **position** of the embryo, where position is emphasized to stress that the concentration patterns are over space (the location of the embryo). These differentiated cells will divide further and differentiate further as the embryo starts to take the form of an adult segmented fly. The chain reaction of gene interactions that begins with the maternal transcription factors that activate other transcription factors, which in turn activate other factors, gives a molecular dynamics description of early development, and hence answers the question of how to build 'from scratch' a body plan (the arrangement of anatomical features).

'From scratch' is misleading and ambiguous when discussing body plans of animals. There are two issues the phrase needs to invoke, ontogeny (development) and phylogeny (evolution). Disentangling the ambiguity we see first, how an adult arises from a single fertilized egg using the model system Drosophila. Second, how did single celled organisms evolve as the diversity of all multicellular organisms. It is in phylogeny and evolution that the question of how to evolve from scratch an animal is misleading. Evolution rarely builds from scratch (i.e. independent evolution such as convergence or parallelisms), body plans are thought to derive through the reconstructions of a set of modules (developmental networks of genes) that each encode for anatomical features (hence all eyes, for example, are homologies at a developmental level through the master gene (transcription factor) *pax6* that binds to a set of genes that set a 'totipotent' or partially programmed cell to start to 'develop' the

eye imaginal disc (the set of cells that form the basic outline of an eye, which when induced, will form an eye (at any part of a body that 'induction' occurs (even in the tail region if so induced)[43].

The idea of a totipotent cell, and cell specification, are necessary to explain development, and hence necessary component of an extension to the modern synthesis. Cell differentiation and cell lineages I will briefly discuss below in the origin of multicellularity. Of course, the initial body with many of the modules (urbilateria, the ancestor of all animals) or its fungal ancestor, still needs to be explained as how one builds it from scratch (this is being done using choanoflagellites and brine shrimp -artemia), but most people are satisfied making inferences with a hypothetical urbilateria that can explain the diversity of animals by the evolution of gene regulatory networks. Furthermore, the statement that anatomical features are modular at the genetic level is not obvious, as many believe complex features (anatomical structures) contain highly correlated genetic networks, such that any genetic mutation would be deleterious and probably lethal. The discovery of the extent of modularity (or the extent of nonmodularity through 'induction' by signalling and paracrine factors) I will discuss briefly in the section below on 'modularity'.

1.3.2.3 The crowning jewel of evo-devo

The work of Ed Lewis (a PhD student of Alfred Sturtevant) on body-transformative genes that were 'saltationary'[22] helped provide evidence and a theoretical framework (extending evolution theory) for the statement that diversity in the animal kingdom (in particular the *segmented* insects) is due to evolution of gene reguatory networks. The "Hox" genes were a

[22]Saltation is the idea that within just one generation major evolutionary transformations could occur, possibly even speciation, taking the idea to its fictional extreme would be like an ape having a baby human, a 'hopeful monster', hence in one generation a speciation event occurred.

group of genes that caused 'Homeotic transformations', which means a transformation in a body part of an animal 'into' another part, like changing a pair of legs into a pair of antennae. These transformations that were known to some biologist such as Bateson (who also coined the word 'genetics') as early as the late nineteenth century, who had catalogued these in various animal groups such as crabs and flys etc.. *The* exemplary gene in the HOX group is 'bithorax', which was discovered in Thomas Hunt Morgan's fly lab in 1915, and a lineage of these mutants has been preserved since then. The gene is named after its mutation that lead to a mutant fly with two (bi) thoraxes (like an abdomen), which consequently doubles the number of wings of the fly (since the thorax is where the wing's developmental source (or pool of cells that each have the right genes turned on and off- a form of programming - leads to the developmental source of cells of the wing, these 'programmed' cells during early development are the so-called 'imaginal disc', and in this case, the wing imaginal disk[23] occurs, and hence is a homeotic transformation. The Morgan lab didn't know the molecular genetic basis behind bithorax (what DNA sequence had changed (or possibly networks of DNA sequences) from the wild type fly to the mutant), but they did know the mutation was inheritable, and hence was genetic. They also were able to 'map' the location of these genes (before they even knew genes were made of DNA) to locations on the chromosome due to a technique developed by Morgan's undergrad student (Alfred Sturtevant).

Armed with the knowledge of the location of the 10 genes that contributed to the so-called bithorax complex (due to Sturtevant's mapping technology), Lewis around 1960 created a model of how the 10 bithorax HOX genes evolved from an ancestral gene through tandem gene duplications (which is largely thought correct) which culminated in his theoretical

[23]place a 'wing imaginal disc' in the location of the head, and you get wings growing out of the head, or place a 'leg imaginal disc' - the cells 'programmed' to build a leg where a wing is supposed to occur and you'll see leg's 'grow' or further develop where the wing's were supposed to be.

paper in the late 1970s. Lewis knew through his own genetic gain and loss of function assays the behavior of HOX genes when crossed with various mutant fly lineages (due to classical genetics Lewis could create hybrid flies due to recombiantion during meiosis to create crosses of known mutant HOX lineages of flies, such as bithorax line). From these experiments he constructed a Wolpert-like gradient model of how the HOX genes would interact like a Fourier series, one HOX gene would be coarsly expressed, thereby turning on another HOX gene in the complex, then the two of these would work in tandem to turn on a third HOX gene, then the three would all work in unison to turn on the fourth HOX gene. Central to his hypothesis was not just the consecutive combinations of the HOX gene products activate new HOX genes in the complex, but also that each new actively generated HOX gene would repress the most recent in time activated gene, thereby creating fine patterns of gene expressions within the cells of the embryo. These unique gene products across the embryo were isolated in segments in space (he could literally see the segments, as these were gross anatomical features each containing 1000's of cells in early development).

Hence, Lewis had proposed a developmental mechanism for how each segment contained different sets of gene products, and thereby explained how each segment would set in motion the signalling and paracrine factors that would eventually cause the segment to form an anatomical feature like a leg, wing, head, or tail. The HOX genes are transcription factors (master genes) that target the genes necessary in signalling pathways and paracrine induction. In short, Lewis had a hypothesis for how anatomical features were built from the segmented larva, but more importantly, he saw an evolutionary mechanism for saltation or macroevolution by his model of the HOX genes; as the basis of HOX genes is that if you mutate them, then the segments of the fly would change in such a way to suggest that the anatomical features that decorate the segments (antennae, wings, legs, eye, abdomen

etc..) were all just one ancestral anatomical feature like a leg for locomotion, that could be adapted or modified for further purposes, like an antennae for sensing the environment, or a mandible or claw for killing prey.

It is now known that there are just three HOX genes in the bithorax complex. While, Lewis had proposed more than this because of various mutations in different locations of the same gene. The HOX genes display various splice sites, and various gene regulatory elements, where mutations in a particular exon or particular transcription factor binding site would cause different phenotypes. Although Lewis didn't get everything right, to a large degree he had proposed correctly one of the greatest theories of biology, how genes 'regulate' development (by master genes turning on other genes necessary for building the anatomical features), and how these genes, when mutated, can lead to major rearrangements of anatomical features.

The question of how the segmented larva developed from the fertilized egg was being elucidated in parallel with Lewis' work through experiments on oogenesis and early development. Partly through the extensive genetic mutation experiments of Christiane Nullsein Volhard and Eric Wieschous it became clearer that there were maternal laid and zygotic genes (transcription factors) that caused Anterior Posterior segmentation in the larva from the humble beginning of the fertilized egg, these genes are known as the gap genes and the pair-rule genes due to the mutant patterns that the would cause in the early embryo and larva.

Although a genetic description was emerging of how the fly develops, it was not possible to isolate the gene products at the time of Wieschous, Nullsein Volhard, and Lewis due to the gene's protein being inside of a multicellular environment. Lewis knew the locations on the chromosome of what gene's caused homeotic mutations, and what segments of a larva

were effected by the mutation; but Lewis simply didn't know to what extent that gene's product caused the mutation (it obviously could have been through some convoluted web of interactions). Work by McGinnis and Levine in Walter Gehring's lab and others, in the early 1980s, possibly motivated by Lewis's hypothesis that **all** the HOX genes are paralogs, lead to the development of a molecular optical microscopic staining technique that allowed one to visually confirm and isolate the region of expression of each HOX gene in specific segments of the fly. As Lewis had hypothesized, every HOX gene were almost identical (they were paralogs), all of which contained a DNA binding domain called the homeobox, a stretch of about 60 amino acids that encodes the binding domain.

A combination of the new microscopic technique, along with pattern recognition techniques has now allowed for developmental biology to quickly advance by elucidating each gene expressed in each cell, and if that gene's product is a transcription factor, where that gene binds within the genome to determine the targets of the factor.

1.3.2.4 Evolution of body plans in animals

The diversity of the animal kingdom can not be explained by standard gene molecular evolution (evolution of gene's encoding proteins); the differences between proteins in animals is not sufficient to explain the diversity in animal phenotypes[61]. It is thought there is about 20,000 proteins, and these proteins (for example, hemoglobin) are largely conserved across the entire kingdom. The diversity is now thought, to a large extent, to be due to the way these proteins are used in different amounts and in different combinations in different cells and body parts of different multicellular organism. To change the amount or create combinations of various proteins in a particular cell in an organism is accomplished by gene regulation, along with expansion of genome sizes (from the early bacterial and protist) through gene

duplication allowing the roughly *same* protein to adapt to a new niche in the organism. Hence, in short, the diversity is due to the evolution of the elements of gene regulatory networks, where the transcription factor binding site is THE fundamental unit.

1.3.3 Gene regulation

1.3.3.1 Conserved Gene Regulatory Networks

Anterior Posterior axis formation

AP HOX genes are conserved. However the maternal element Bicoid is derived, which seems contradictory to the idea that for master genes the earlier they are turned on in development the more effect they will have on downstream components of developmental pathways. The Bicoid protein gradient is formed during oogenesis (the maternal process of producing the unfertilized egg), where maternally laid mRNA becomes localized to the anterior portion of the egg casing (which is a multicellular structure, which upon completion of oogenesis, all cells but one (the egg) dissipate and form part of the yolk surrounding the egg's nuclei, all of which is inside the egg casing[24].

Master gene's such as Bicoid are usually thought to be highly pleiotropic (form network hubs) in that they causally interact with many target genes (which in turn will interact with further genes). The fact that Bicoid is not a conserved element of development across the animal kingdom suggests that even the upstream elements of developmental pathways are modular (can be substituted) or that there are multiple ways (paths) that lead to the same developmental outcome.

Although Bicoid is not conserved, some of the downstream elements of AP developmental

[24]I'm considering the 'nurse' and 'follicle' cells as a part of what I'm calling the 'egg casing'.

pathway are conserved most notably the HOX genes ,which have shown to be present even in chordates (which includes the vertebrates, and some weird invertebrates like echinoderms). The conservation of the HOX genes in mammals and insects was confirmed not only by sequence similarity (i.e. homeobox sequence conservation), but also in that the HOX genes expression patterns are similar during development and play the same functional role in laying out the location of the anatomical features such as head and tail. Indicating that the genetic instructions for how the embryo 'knows' where to form the body structures is an ancient conserved network that has been so useful to the animal kingdom that has been preserved since the cambrian explosion[25]. The seminal experiment that confirmed the genetic role of the genes that shape body plans in the animal kingdom was with *eyeless* and *pax6* in 1994 in Walter Gehring's lab, where an insect's gene early in development that is known to cause loss of compound eye formation (*eyeless*) was replaced by a similar gene found in mouse (which controls where the mouse eye forms, which is a different form than the fly's compound eye). Using molecular biology techniques the endogenous insect gene was replaced with the mouse gene, where it was found that the fly still formed compound eyes using the mouse gene, suggesting that the *eyeless* gene is ancient gene that instructs the embryo to form an eye. What kind of eye? That is not a question answered in body plan genes. The developmental genes necessary to instruct the embryo where each anatomical feature goes, the master genes, are fundamentally distinct from the genes that actually are critical components of the actually anatomical structure. For example, the gene *eyeless* is found (expressed) during the development of the eye (i.e. it is expressed during development in locations of the embryo where an eye will form), but it is not found (not expressed) in the

[25]Using the construction of a home from a master builder (maybe an architect) as a metaphor, the master genes (i.e. the HOX genes) don't tell the developmental program 'how' to build a wing, it tells the program 'where' to put the wing in the home. Hence 'master builder' is a bit misleading of a metaphor, they don't build anything, they simply give crude locational cues of where certain features should begin to be built.

fully functional eye (the adult structure). Hence the body plan genes, are similar to bankers or stock brokers in a financial system, who don't actually produce any tangible products like music or a cup of coffee, rather they help set in motion the complicated web of interactions that can bring these commodities together in a harmonious manner[26].

Further evidence of this significance was corroborated by modifying the expression pattern of this gene, which would cause eyes to form at different locations of the body. Hence, the HOX genes, are all examples of master genes (transcription factors) whose role is in regulating how the animal is to build itself into the fixed arrangement of body parts that help define each animal specie.

Dorsal Ventral axis formation

The early steps necessary in forming the Dorsal Ventral axis of the fly embryo, much like the AP network, are set in motion during oogenesis, where in the case of DV, at the same time the Bicoid mRNA is producing a gradient in the unfertilized egg, in parallel a gene called Gurken (which is transcribed from the egg's nucleus, unlike Bicoid, where the mRNA were maternal) polarizes DV axis of the encasing egg by localizing to one side of the future embryo (this side becomes the 'end' of the Dorsal axis, and directly opposite of this side becomes the Ventral side of the future embryo. Shortly after Gurken localization and DV axis polarization, a maternal gradient of Spatzel is created from a series of protein complex degradations or splits induces the 'Toll pathway', an ancient pathway conserved between humans and bacteria used in innate immunity, where Spatzel binds to the Toll transmembrane receptors that signal to a maternal complex of Cactus and Dorsal protein that in turn causes the complex to

[26]Some HOX genes, like *dpp*, are master genes that play the role of putting disparate object together in harmony, but they also play other roles latter in development, such as in the literal construction of body parts (such as the nervous system for *dpp*).

split and the protein Dorsal now displays a nuclear localization signal, that allows it passage into nuclie, where Dorsal acts as a transcription factor. Hence the more Spatzel the more Dorsal that is unlocked from the complex, leading to a Dorsal nuclear concentration gradient that mirrors the Spatzel gradient. Dorsal acts as a transcription factor (hence Dorsal is a master gene), which unlike most master genes, Dorsal in turn is activated by a signalling network (not other upstream master genes). Furthermore, *dorsal* and *cactus* are 'maternal effect genes', a set of genes discovered by Nusslein-Volhard and Anderson in 1984, where the initial egg's Dorsal and Cactus protein is transcribed from maternally laid mRNA of *dorsal* and *cactus* maternal genes[27]

Dorsal acts as a coarse grained regulator of the DV axis by forming roughly three different cell types from the totipotent blastula. These three cell types are components of the 'germ layers' (three primitive tissues seen in all 'tripoblasts' - the animals with bilaterial symmetry that have both a mouth and anus (i.e. go through gastrulation). Dorsal helps form the mesoderm, neuro-ectoderm and dorsal-ectoderm; where the endoderm is partly constructed by AP genes. The endoderm is the first evolved germ layer that distinguished inside from outside of primitive 'diplblastic' organisms; animals with radial symmetry like echinoderms that only have one hole (no gastrulation) that serves as both a mouth and an anus (like jellyfish)), where the ectoderm served as the 'outside' of these primitive organisms (hence the mesoderm and the 'triploblast' clade are a more recent evolutionary invention than the 'diploblasts').

Just as Bicoid is a derived feature in flys, similarly Dorsal's use as master gene in fly is a derived feature. The seminal experiment of DV axis polarization in chordates was

[27]It is possible that zygotic *dorsal* is also transcribed in the embryo, as there are putative binding sites for Dorsal nearby the *dorsal* gene.

through experiments by Spemann and Mangold that revealed the source of the germ layers was localized to one side of the vertebrate embryo (they used translucent newt embryos), which was discovered by Spemann artificially splitting the embryos in half by putting a thin belt of string around the tiny embryo and squeezing the string tight, in a way acting like the contractile ring of cytokinesis. Their famous transplantation experiment revealed that the analog of the fly's germ layers constructed from the Dorsal transcription factor's gradient, where Dorsal binds to CRMs of DV target genes, was occurring in a small localized region of the newt embryo (completely different that insects). Spemann called this region the 'organizer'. What Spemann had discovered was that to a large degree, vertebrate embryos are almost entirely yolk in early development. And in early development, the embryo forms in a very small region of the 'egg', the organizer, where the majority of the peripheral region of the egg is just 'yolk', which supplies necessary molecules and energy late in development inside the egg.

Although there is no homolog of Dorsal active in early vertebrate development, the targets of Dorsal found in fly, are also found active in vertebrates. Twist and Snail are found active in the mesoderm (the Spemann Organizer) and most famous is *dpp* and *sog*, where their homologs form an antagonistic gradient in vertebrates that patterns the notochord (albeit in chordates there has been an 'inversion', which, for example, leads to hearts being ventral in vertebrates while dorsal in most invertebrates). The notochord is a derived vertebrate feature that forms just below the neural tube (the future brain and central nervous system that becomes decorated by the vertebra)[28].

The fact that Dorsal is not found to be the master gene that patterns the vertebrate

[28]Tunicates develop a gelatinous notochord in early development, suggesting these primitive sea organisms are closer to humans than flys are! Showing not all paths of evolution lead to complexity.

organizer (the germ layers) is very perplexing, just as confusing as why Bicoid is not found upstream of the developmental pathway that leads to activation of the HOX genes in vertebrates. Furthermore, vertebrate development, to a large degree, seems very different than the localization of transcription factors to specific regions of the embryo (protein gradients), where the different germ layer emerge at specific locations in space due to the various interactions of the protein (morphogen) gradients. Vertebrates use 'induction' (cell cell interactions) early on in development as a patterning mechanism, where, for example, the endoderm is first formed by the organizer, juxtaposed against a somewhat undifferentiated layer of cells (Nieuwkoop center), where the endoderm - Nieuwkoop interaction in turn induces the formation of mesoderm tissue. Hence, unlike in flys, where all three layers form somewhat simultaneously, in vertebrates the endoderm forms first, which in turn induces the development of the mesoderm (albeit, the mesoderm targets of Dorsal in fruit fly are activated first, where Snail's expression border helps differentiate neuroectoderm tissue from mesodermal tissue).

Perhaps these master genes (Bicoid and Dorsal) are a modification of mitosis' chromosome separation, which became fixed in some of the insect lineage. In mitosis the microtubles attach to centromeres to direct the movement of the duplicated chromosomes during nuclear division, while in oogenesis the actin attaches to Bicoid or Gurken mRNA leading to dynesin or myosin motors moving these mRNA transcripts to localized region of the egg. Perhaps the radical differences in insects and vertebrates suggests that urbilateria never existed; that the protostomes and deuterostomes independently evolved from a unicellular protist, albeit, the conservation of the HOX genes would seem to rule this out. Regardless, by strategically choosing organisms (transitional forms) in between vertebrates and invertebrates these questions can begin to be answered by looking at how, for example, the regulatory regions of

dpp and *sog* have evolved to be inserted in the developmental pathway of organsims in the vertebrate lineage. An 'insertion' event of *dpp* and *sog* in vertebrate genomes is not accurate, as the genes are conserved between invertebrates and vertebrates, rather the binding sites of the factors that regulate *dpp* and *sog* must have evolved to respond to the particular factors active in the respective lineages, the evolution of these binding sites is what would help reveal the origin of the event that led *dpp* and *sog* to be regulated by different activators in early DV patterning.

1.3.3.2 Modularity in gene regulatory networks

The time point of development where Dorsal begins to act like a master gene (a transcription factor) is the same point in time when Bicoid is causing its targets to become expressed. How is it that each developmental pathway for axis specification (DV and AP) is simultaneously operating in a given cell of the embryo? The modularity of the transcription factor binding sites allows for AP target genes to only have DNA binding sites for Bicoid, and similarly DV target genes to only have DNA binding sites for Dorsal. Similarly as Bicoid works in tandem with some of the genes that it activated in the first place (such as Hunchback and Giant) in feedforward circuits, where these AP transcription factors combinatorial act together by binding to their target DNA binding sites that co-occur in short segments of DNA (about 300 bp the *cis* regulatory modules (CRMs)) that are near the target gene. By excluding Dorsal and other DV master transcription factor binding sites from these CRMs, AP developmental pathway can proceed in a particular cell in the embryo independently of the DV developmental pathway that is also operating in that same cell. Hence, these circuits, or components of the developmental pathways that are 'patterning' the embryo (fating the cells through cellular differentiation) are modular. Late in development, once the nervous

system has grown, the modularity seen in laying out the body plan that has organized the anatomical features largely disappears, as the body becomes an integrated system highly interdependent and controlled by the nervous system.

1.3.3.3 Transcription factor binding sites

Gene regulation occurs at various temporal and spatial levels. Within a cell, the amount of a particular protein can be regulated at 'translation' level by regulating the amount of mRNA transcripts that are translated. However, transcription is the process that is predominantly regulated during development.

At the finest spatial level, transcription factor binding sites provide the docking area for transcription factors to temporarily bind and in various ways modify the transcription of nearby genes. The binding process physically depends on the concentration of the transcription factor and the number of binding sites that reside within a CRM. The flanking sequence of a CRM as a competing location for the binding of the transcription factor, hence knowing the size of a genome (i.e. the number of flanking binding sites) is equivalent to knowing the chemical potential. Thermodynamically a transcription factor follows the chemical potential gradient, hence unoccupied sites in the genome compete for binding of a given transcription factor. At the transcription level

CRMs have evolved strategies to recruit transcription factors by creating highly specific binding sites that reside within the CRM. These very specific binding sites provide a potential energy 'well' that captures transcription factor due to the very low binding energy of the well (compared to the high energy of docking at the flanking sequence of the CRM). Furthermore 'binding cooperativity' also provides a further drop of the potential well for a given transcription factor if a particular cooperating factor is bound nearby.

Chapter 2

2.1 Introduction

The 'particle' abstraction of classical mechanics reduces the many degrees of freedom of an extended material into a single point in space and time [63]. A similar abstraction is useful for treating the process of regulation by transcription factor proteins of gene regulatory networks. In such a model, the entire genome is seen as a one-dimensional lattice where each lattice 'site' is like a type of static particle with a coordinate along the genome, and where the site is a short sequence of DNA, ranging from a single base-pair to a coarse-grained extended sequence of DNA. Each such site can be defined by its specific logic given by the interactions that are relevant for regulating transcription [27, 5, 64]. This logic, encoded in the type of site, is an inheritable trait. Furthermore, evolution of regulatory sites changes the logic, which is known to cause major transformations on animal body plans [24, 106]. Understanding this logic, at a sequence level, has produced state of the art phylogenetic models for classification at the phylum level that allows us to better understand our deepest homologies with the rest of the kingdom.

2.1.1 Position Weight Matrices

Commonly, estimating the nucleotide frequencies of functional transcription factor binding sites is achieved by aligning experimentally confirmed functional sites of length s, and counting the frequency of each nucleotide at each position. These counts can then be used to infer the distribution of functional binding site sequences. The inferred distribution of functional

sequences is called a Position Weight Matrix [100]. For example, for a length s binding site, the probability that the binding site has the sequence S is:

$$P(S) = \prod_{ij}^{s,3} P_{ij}^{S_{ij}}, \tag{2.1}$$

where the sequence S is represented by the matrix of indicator variables $S_{ij} \in \{0, 1\}$ (Boolean variables), and P_{ij} is the probability (maximum likelihood estimate from the frequencies) to find base j at position i, with $i \in \{1, 2, \ldots, s\}$ and $j \in \{0, 1, 2, 3\}$, such that each integer represents a letter from the alphabet A,C,G,T.

Information theoretic and classification methods can then be used to relate these probabilities to linear (additive) logarithmic models or a discrimination function. For example, the energy PWM gives a bioinformatic score *E(S)*, for any sequence *S*. The energy of the sequence can be decomposed into a sum over each internal position of the sequence:

$$E(S) = \sum_{i}^{s} \sum_{j}^{3} E_{ij} S_{ij}, \tag{2.2}$$

where the binding site sequence S is again represented by the indicator variable S_{ij} for each position i in the sequence and base-pair j, which selects the appropriate transcription factor-DNA interaction energies E_{ij}. We define the interaction energies E_{ij} mathematically in Equation (2.10) below.

2.1.2 In-vitro Biophysical PWMs

The energy weight matrix elements used in Equation (2.2) can be determined for each of the $4s$ matrix elements using an affinity assay. This assay is purely based on physical principles,

completely blind to notions of "functional" (meaning adapted) binding sequences. The key measurement is the relative change in affinity to the transcription factor for all possible single mutation sequences from the highest affinity sequence [100, 38, 52, 101, 37]. Such an assay assumes that the highest affinity sequence (which we denote as S_0), is known. By choosing the highest affinity sequence as the reference DNA-transcription factor interaction, one can then construct the full set of relative affinities for full sequences (all 4^s affinities). Just as a key assumption of the PWM model was linearity in sequence, so too in this experiment we must assume that the binding energy is a linear function of the sequence. This assumption enables each of the three possible DNA mutations from the reference sequence at a particular position within the DNA binding site to be tested independent of the genetic background of the remaining positions within the binding site.

The theoretical justification that the binding energy is a linear function of the sequence is that the binding affinity constant $K(S)$ is equal to the exponential of the binding energy in units of kT, where k is Boltzmann's constant and T is temperature. The free energy, being a state function (i.e., exact differential), then would result in the following displacement reaction: $\log K(S) = \log K(S_0) - \Delta G$, where the transcription factor was originally bound to sequence S_0 (the reference sequence) and then (by any physical process) is displaced and binds to sequence S. If we set the energy scale such that the highest affinity sequence bound to the protein has zero energy, then all other bound complexes have higher energies $G(S)$, hence $\Delta G = G(S) - G(S_0) = G(S)$.

Using the physical approach above, one can treat each mutation of a base from the reference sequence (highest affinity sequence) as a perturbation of the reference sequence S_0.

By expanding the free energy in sequence space, we have

$$G(S) = \sum_{ik}^{s,3} \Delta G_{ik} S_{ik} + \sum_{i,j=1}^{s} \sum_{k,l}^{3} w_{ijkl} S_{ik} S_{jl} + \ldots \tag{2.3}$$

$$\approx \sum_{ik}^{s,3} G_{ik} S_{ik}. \tag{2.4}$$

The pairwise interaction term, w_{ijkl}, is a function of four indices, where indices i and j run over the positions of the sequence S, and the indices k and l run over the nucleotide bases. The indicator variables S_{ik} and S_{jl} select the appropriate pairwise interaction term w. The expansion in sequence space has a total of 2^s interactions, the final approximation assumes all these are negligible except the first order terms.

2.1.3 Evolutionary PWMs

Just as a phylogenetic analysis of genes can reveal subsequences that are important for the function or enzymatic activity of the protein, so too can phylogenetic analysis of binding sites reveal subsequences that are important for the binding function (affinity) of the sequence. Unlike cladistics, where a binding site alignment would only include a monophyletic group (sequences evolved from a common ancestor), and hence be hampered by patterns of conservation that are due to inheritance as opposed to adaptations, here we use a phenetic approach to alignment, based on Berg and von Hippel's phenetic approach [13], where both convergent sites, paralogs, and orthologs are used in the alignment to reveal conserved patterns in the DNA binding sites that are a consequence of the molecular properties that provide the binding phenotype.

A basic molecular evolution principle initially formulated by Zukerlandyl and Pauli and latter utilized by Dayhoff and refined by Kimura is that neutral DNA accumulates sub-

stitutions with a reliable rate, such that *neutral* DNA can be used as a molecular clock. However, *functional* DNA's mutation rate (what Berg and von Hippel called the "base-pair choices") are correlated with the functionality of a site [13]. Hence, functional DNA under purifying selection evolves slower (if at all) than neutral DNA, enabling a comparative analysis of regulatory sequences by screening conserved blocks of sequences, or "phylogenetic footprints" [97].

Berg and von Hippel used these assumptions in 1987 to relate the empirical nucleotide counts from an alignment to theoretical binding site sequences under mutation-selection balance [13]. Theoretically, they assumed a binding site was constrained by the binding affinity necessary for binding (*i.e.*, binding that influences gene expression) [15]. This constraint allowed them to use Jaynes's principle to derive a theoretical distribution known in physics as the Boltzmann distribution, which they then could equate to the empirical normalized counts from Equation (2.1). In this context, Jaynes principle states that the information content of the set of binding site sequences (*i.e.*, binding site sequence data in the form of Equation (2.1) and knowledge of the genome-wide frequencies–the prior, or GC content of the genome–should be minimized subject to the binding energy constraint [54].

For a simple example, consider a binding site of just one base-pair[1]. The "Lagrangian" for the constrained minimization problem can be written as (the sum is over the nucleotides that base B can take on)

$$\sum_B P(B) \log \frac{P(B)}{P_0(B)} - \lambda_0(\sum_B P(B) - 1) - \lambda_1(\sum_B P(B)G(B) - \langle G \rangle). \tag{2.5}$$

[1]Binding sites are frequently about 10 base-pairs long. A binding site of length one base-pair is not realistic for transcription factors, as most proteins would cover more space than one base-pair (about 1 nanometer). For an evolutionary argument for why binding sites are about 10bp in length see [99], and for a diffusion argument see [93]

The first term is the information content of the steady state probabilities $P(B)$ relative to the genome-wide frequencies $P_0(B)$, the prior. The second term represents the normalization constraint over the probabilities (where the prior is assumed fixed) and the last term is the constraint that the average binding energy be fixed. Minimizing the Lagrangian leads to the theoretical estimate of the steady state distribution, $P(B)$, which takes the form of a Boltzmann distribution (e.g., see Equation (2.9) for a definition of the Boltzmann distribution).

The equilibrium frequencies, $P_0(B)$, are those expected of neutral DNA (*e g.*, the frequencies estimated from the Jukes Cantor substitution model. Sites under selection are forced away from equilibrium, and form a steady state distribution $P(B)$. For a physical example, the relative frequency of a particular base B is like a concentration, which when in thermodynamic equilibrium will be equal to the concentration of this molecule in the background. Assuming the background can be modeled as chemically random bases (A,C,G,T) [104], then in thermodynamic equilibrium the base Bs concentration will equal the background concentration of the respective base. In a steady state, however, the base frequency is forced to to concentration unequal to the background. Similarly, in an evolutionary steady state, there is a flux of mutations driving the population of binding sites to the random frequencies, but this flux is balanced predominantly by the flux from the selective pressure. In the population genetics sense, the steady state frequencies are the result of mutation selection balance.

2.1.4 Relation between biophysical PWMs and evolutionary PWMs

As a consequence of Berg and von Hippel's hypothesis that the normalized frequencies from an alignment of binding sites could be equated to the theoretical distribution of sequences under mutation-selection balance (the Boltzmann-like distribution) [13]; Berg and von Hippel

were able to derive a simple relation between their information theoretic logarithmic score E(S) from Equation (2.2), and the known binding energies $G(S)$ of the binding sites to the transcription factor from Equation (2.3). Using the standard statistical mechanics relation: $\log \frac{K(S)}{K(S_0)} = \log \frac{P(S)}{P(S_0)}$, where $K(S)$ is the binding constant, and $P(S)$ is the Boltzmann-like distribution (see Equation (2.9) for details), and observing that $\log \frac{P(S)}{P(S_0)}$ can be replaced by the normalized frequencies from the alignment, and defining the information theoretic score from Equation (2.2) as $E(S) = \log \frac{P(S)}{P(S_0)}$[2]; one then obtains:

$$\log K(S) = \log K(S_0) - \frac{\Delta(E(S))}{\lambda_1}, \tag{2.6}$$

where $E(S)$ is estimated from an alignment. (*E(S)* is fully explained in our Methods section, where $\Delta E(S) = E(S)$, and similarly $\Delta G(S) = G(S)$ by choosing S_0 to be a reference.) The linear relation above is of the same form as the first-order thermodynamic perturbationEquation (2.3):

$$\log K(S) \approx \log K(S_0) - \Delta G. \tag{2.7}$$

This gives us a linear relation between the evolutionary substitution pattern (data from an alignment), E, and the free energy, G (in units of kT).

2.1.5 Shortcomings of PWMs

Analyzing typical functional binding site sequences for a particular transcription factor reveals signs of a conserved pattern of nucleotides at specific positions within the binding site.

[2]Here we our conflating our notation for $P(S)$, which in one case is the empirical normalized frequencies from the alignment (which Berg and von Hippel denoted as $f(S)$), while in the other case of statistical mechanics $P(S)$ is a theoretical distribution parameterized by the Lagrange Multipliers (which can be shown to be the thermodyanmic temperature for systems like an ideal gas[6]). Here we do keep the derived variables $E(S)$ and $G(S)$ separate, in order to clearly see the relation between the bioninformatic score $E(S)$ and the free energy ΔG.

However, because the sequences are short, false-positive matches to the pattern are expected to occur frequently in large genomes, too frequently than time available for the protein to find all the sites. This kinetic search problem was also analyzed by Berg and von Hippel using one- and three-dimensional diffusion models [14], which has since been reinterpreted several times. In particular, Sela *et al.* showed that symmetries in DNA sequences flanking functional binding site loci can dramatically affect binding [93], later verified experimentally [3]. In the same manner, bioinformatic searches for binding sites using only the conserved patterns in order to discover new binding sites often results in poor predictions on a genomic scale [18].

Another limitation of the model is that in development, heterotypic clusters of binding sites (rather than isolated sites) govern gene expression. Hence, binding site sequence matches to a motif, if occurring in an isolated locus within a genome (i.e., not occurring within a cluster of other binding sites) are incapable of recruiting the complexes necessary for transcription, and hence these isolated loci are unlikely functional. Hence the functional sequence distribution simply does not contain enough information to make a one-to-one map to the functional loci [89]. Furthermore, in eukaryotes, binding is modulated by the chromatin state of a locus and the cellular state that the genome resides in. These epigenetic cues and other external variables that influence binding are not usually encoded into the binding site sequences, and gives rise to departures from the linear assumption inherent the PWM model.

Evolution in development has repeatedly evolved new combinations of binding sites producing new types of logic regulating gene expression [41, 28, 29]. Traditional bioinformatic sequence tools to discover binding sites in developmental systems can discover the low resolution segments (500 bp) of regulatory DNA that contain clusters of coevolving binding sites,

CRMs, by simply using clusters of motifs [19]. However, determining what sequences within the CRM are functional is difficult. For example: is the spacing between sites functional, is the ordering of sites functional, what about 'half sites' or sites with mismatches, what is the number of mismatches allowable before a sequence is not functional? Tedious genetic experiments must be conducted in order to discover what sites significantly contribute to gene expression [29].

For example, the *in vivo* binding site contribution to gene expression can be understood by comparing the expression of a target gene driven by a wild-type CRM with a knock out of a putative binding site. However, this is complicated for a number of reasons: first, binding site turnover within CRMs leaves remnants of functional sites such as "half sites" that have partial matches to motifs [25], second the multiple half sites (that are easier to evolve) may be able to compensate for a strong full site. Therefore, even with a confirmed functional CRM, functional binding site discovery is a daunting task, due to vestigial sites that have fuzzy or poor matches to bioinformatic motifs.

2.1.6 Physical Shortcomings of PWMs

2.1.7 Dependencies within transcription factor bindings sites

The linear relation in Equation (2.6) becomes nonlinear if there are cooperative interactions between positions within a binding site (or if there are context dependent base-pair dependencies). For example, cooperativity at the biochemical level tends to cause the linear relationship between the first order Gibbs free energy and the binding constants to become nonlinear as a function of sequence, thereby decreasing the ability of linear models (or first order thermodynamic perturbations) to capture the relationship [52, 16]. Furthermore, some

DNA-protein interactions require specific nucleotides at various positions to jointly occur, such that the additive sum of the interactions of each nucleotide to the protein is not what would be expected under the linear model. In such cases it becomes important to consider higher-order interactions, such as via dinucleotides or other various joint occurring nucleotides [96, 94].

2.1.8 Dependencies between transcription factor binding sites

If the base-pair preferences for a particular transcription factor are contingent on a cooperating factor, then evolution will have filtered the co-occurring sites jointly. For example, the transcription factor NfκB is known to have a specificity that is dependent on co-occurring binding sites [67], and similarly the binding sites of the Glucocorticoid Receptor are specific to their context [75]. The NfκB homolog Dorsal's binding sites have also been shown to encode differences when active in different innate immunity pathways [22], or to signal Dorsal's role as an activator or a repressor [78].

2.1.9 Conditional PWMs based on co-occurring factor binding sites

Here we present a model that incorporates locus-specific information into PWMs that we call "conditional" PWMs, that improve binding site discovery within CRMs by incorporating flanking information of each binding site locus into the functional binding site sequence distribution. This is useful for transcription factors that display specialized behavior based on their *cis*-environment. Our PWM approach accounts for DNA-DNA epistasis (hard-wired cooperativity) that is a function of the DNA spacer between target binding site and a

putative cooperating transcription factor's site. The hypothesis is that base-pair preferences between known cooperating proteins will be a function of the spacer between the known sites (assuming that sites that are separated by large spacers are effectively non-interacting). If the base-pair preferences change as the spacer changes, then evolution will have filtered the co-occurring sites jointly rather than independently. As a consequence, we expect different PWMs for binding sites separated from a putative interacting site as a function of spacer size. This model is similar to the cooperative nucleotide model in Ref. [13], but now we effectively have a spacer model between binding sites.

Furthermore, Berg and von Hippel in Ref. [13] introduce a spacer dependent interaction energy, which similarly addresses that spacing between co-occurring transcription factor binding sites affects the total binding energy between the two separated sites. However, in their spacer dependent interaction energy, these authors kept the PWM for each binding site a constant, regardless of its interaction with co-occurring binding sites, and only focused on the spacing between the co-occurring binding site. Our model, in a sense, encodes the spacer dependent interaction energy into the different conditional PWMs constructed for different spacer windows.

2.2 Materials

2.2.1 Data for known Dorsal binding sites in *D. melanogaster* Dorsal-Ventral network

The initial development of the fruit fly is partly based on maternally laid morphogens that form a gradient across the blastoderm thereby causing differential target gene expression [66,

55, 68]. The Dorsal-Ventral (DV) network of genes active in the *Drosophila* embryo is largely conserved across the Drosophila genus, furthermore their coarse-grained expression patterns in terms of percent egg length along the DV axis are also largely conserved [86]. The transcription factor Dorsal regulates the genes responsible for patterning the DV axis of embryogenesis leading to gastrulation [77, 98, 107]. Hence Dorsal transcription factor binding sites within and across *Drosophila* species represent a large set of binding sites that are amenable to constructing a PWM.

We collected Dorsal binding sites active in the *Drosophila melanogaster* neuroectoderm region of the DV axis that cooperate with a bHLH (basic helix-loop-helix) dimer with Twist. These sites are the D_β sites of Table S2 of Crocker et.al. [25], the Dorsal sites from figure 2 of Crocker *et al.* [26], as well as the "specialized" NEE (Neurogenic Ectoderm Enhancers) and NEE-like Dorsal binding sites of Erives *et al.* and Crocker *et al.* [35, 26]). Those sites are specialized in the sense that they have been shown to evolve slower than flanking Dorsal binding sites in homotypic clusters of Dorsal binding sites in the NEE [25], and possibly specialized to the cooperative interaction with Twist (which we aim to characterize through information techniques).

There is ample evidence and a long-standing history in the literature for Dorsal sites cooperating with a bHLH dimer, see [73, 62, 108, 102, 59, 56] and references therein. In those cases, the bHLH dimer is likely a Twist:Daughterless heterodimer. Daughterless is a ubiquitously expressed and obligate partner in tissue-specific bHLH dimers, such as Twist. The 'specialized' Dorsal data set is labelled as $\mathcal{D}_{\mathbf{DC}\text{mel}}$, where $\mathcal{D}$ represents a data set, and the subscript **DC** means 'Dorsal Cooperative' and mel stands for the species *melanogaster*.

We also collected Dorsal binding sites from the REDFLY footprinting database [39] for target sites active in embryogenesis. This data set is labeled as $\mathcal{D}_{\mathbf{DU}\text{mel}}$, where *DU* means

'Dorsal Uncooperative'. We did not find Dorsal footprinted sites from REDFLY for the Dorsal target gene *snail* in the CRM of *snail*, hence these Dorsal binding sites were omitted from our data set (our CRM data are described below). The $\mathcal{D}_{\mathbf{DU}\text{mel}}$ is a subset of the full REDFLY Dorsal binding sites, where we filtered out any sites that had already been collected in our $\mathcal{D}_{\mathbf{DC}\text{mel}}$ data set, or sites that were not active in the DV network, or binding site loci that overlapped.

2.2.2 DNA sequence context of binding sites

Our aim is to characterize the Dorsal sites based on patterns in the loci's flanking sequence. The regulatory regions (the cis-regulatory modules) of DNA that contain the $D_{\mathbf{DC}\text{mel}}$ and $D_{\mathbf{DU}\text{mel}}$ binding sites consisted of the following *melanogaster* CRMs: *rho, brk, sog, sogS, vn, vnd, twi, zen, dpp, tld*. In that list the CRM is labeled by the gene it targets, and the *sog* gene had its Dorsal binding sites in two distinct CRMs labeled *sog* and *sogS* (where *sogS* is a 'Shadow' enhancer).

These CRMs have been collected in a centralized file by Papatsenko et al. [83]. Additionally, these authors collected known *melanogaster* modules from the literature and using a BLAST approach predicted the remaining 11 Drosophila orthologs of the known *melanogaster* regulatory regions (at that time there we 12 sequenced genomes for *Drosophila*). The orthologs were not 'known' with same certainty as the *melanogaster* data, however we will still classify these as known for our purposes, as conservation of synteny (order of sites) along with each module containing multiple conserved blocks where sequence matches to binding sites reside renders these predictions accurate. These modules are usually minimal modules that are on average about 300 base pairs in length.

We aligned the 12 orthologs of each CRM, and only extracted the aligned blocks that

contained our $D_{\mathbf{DC}\text{mel}}$ and $D_{\mathbf{DU}\text{mel}}$ binding sites, see Supplement section 2.9.1 for details. The enlarged set of combined data we call $D_{\mathbf{CB}} = D_{\mathbf{DC}} \cup D_{\mathbf{DU}}$, where the removed subscript mel on DC and DU, denotes that all 12 orthologs of a given binding site sequence are in the data set, and CB stands for combined.

2.3 Methods

2.3.1 Clustering Dorsal target loci based on co-occurring binding sites

Given the locations of the Dorsal binding sites within a given CRM (see Supplement section 2.9.5 for details) and the *predicted* sites of another factor (a putative cooperating factor), we are able to construct a distance matrix where each row 'i' is a known Dorsal locus (base-pair coordinate), and each column represents a predicted co-occurring factor's locus 'j' within the CRM. The matrix elements of the distance matrix are the spacer length (denoted as $d(i,j)$) in base-pairs between any row i (Dorsal binding site locus) and column j (co-occurring binding site locus), a difference of the coordinates z of the loci:

$$d(i,j) = z^i - z^j - w^i, \tag{2.8}$$

where we assume that the ith Dorsal site appears upstream from the jth co-occurring site, and that both sites are annotated as on the positive strand of the CRM, where w^i is the width (length) of the ith site, and z^i and z^j are the CRM coordinates of site i and j respectively. Here we define the spacer as the base-pair distance of neutral DNA between two binding sites (hence the internal positions within either site are not counted as part of the spacer).

For cases where the Twist and Dorsal site overlap, the spacer is valued at 0 bp regardless of the amount of overlap. For cases that a CRM did not contain a predicted co-occurring site, we set the spacer to a maximum value such that the corresponding Dorsal site for the spacer was guaranteed to be classified as "Uncooperative".

2.3.2 Classifying binding sites based on spacer window

We define a partitioning of the flanking sequence of any given Dorsal locus, hence we use the reference frame of the Dorsal locus with both upstream and downstream sequence. We partition the upstream flanking sequence by the minimum distance $d_{\min}$ and a maximum distance $d_{\max}$ away from the locus using Equation (2.8). Similarly, we define a symmetric partition of the downstream flanking sequence by the minimum distance -$d_{\min}$ and a maximum distance -$d_{\max}$ away from the locus. We then define a coarse-grained binning of all the flanking sequence into just two bins, where a 'spacer window' represents the bin that contains the interval $[d_{\min}, d_{\max}] \cup [-d_{\min}, -d_{\max}]$, and the other bin contains all the rest of the flanking sequence. Once the bin borders have been defined by the spacer window, we then define a Boolean class variable C, which classifies each Dorsal locus as $C = 1$ if the co-occurring binding site of interest is present *in* the spacer window, and $C = 0$ if the co-occurring binding site sequence of interest is absent *in* the spacer window. Hence, the class variable is entirely based on the patterns that occur within the spacer window, as the class value of each class is determined solely on co-occurring sites in the spacer window. Using Equation (2.8) we classify the Dorsal loci that fall within a defined window. Once each Dorsal binding site's locus is assigned a class, we then can align the loci of a class and estimate the conditional PWM.

2.3.3 Energy estimation of a base

The theoretical steady-state Boltzmann-like distribution is the solution to minimizing the Lagrangian with respect to $P(B)$ in Equation (2.5). The Boltzmann-like distribution in units of the second Lagrange multiplier is:

$$P(B) = \frac{P_0(B) \exp -E(B)}{Z}, \tag{2.9}$$

where the normalization Z is related to the Lagrange multiplier λ_0, and we have assumed calibration of the energy $E(B)$ by estimating the shift and scaling factors from Equation (2.6). Assuming our frequencies from Equation (2.1) is governed by the Boltzmann-like distribution, we then can construct an energy PWM by inverting the distribution, arbitrarily choosing the consensus base B_0 to be the zero of the interaction energy between transcription factor and bases. The consensus base is the base at a position with the most counts from the alignment, hence this choice of reference leads to all other bases contributing a higher energy (or zero for degenerate cases). We then can calculate the interaction energy of the remaining bases B as:

$$E(B) \approx -\log \frac{P(B_0)}{P(B)} = -\log \frac{n_{B_0} + \beta}{n_B + \beta}. \tag{2.10}$$

Here we have made the approximation that the degeneracy factors $P_0(B) = g(B)/L$ are negligible (this is the prior or background DNA frequency), where $g(B)$ is the multiplicity or number of times that base B occurs in a genome of length L [12], n_{B_0} are the counts of the reference base B_0 and similarly n_B are the counts of base B from the contingency table estimated from the alignment of n known sites, and β is a pseudocount $\beta > 0$. The

joint energy of a given base B with co-occurring flanking sequence S' (that may or may not contain a co-occurring binding site of another factor) is defined as $E(B, S') = E(B) + E(S') + w(B, S')$. By setting a spacer threshold (spacer window) and an energy threshold on the potential cooperating factor we effectively create a Bernoulli variable for the flanking sequence, such that S' is aggregated into the class variable C. Hence, we have $E(B, C) = E(B) + E(C) + w(B, C)$, where $w(B,C)$ is an interaction energy that is shared between the systems B and C. Once we have determined what class a Dorsal locus belongs to, we are then uninterested in the energy of the co-occurring site in sequence S'. Hence we define a conditional energy that is the standard PWM energy from Equation (2.2) for a particular position and base plus the interaction term:

$$E(B|C) = E(B) + w(B, C). \tag{2.11}$$

The interaction term shifts the standard energy of a sequence if $P(B|C) \neq P(B)$. We define our context C for Dorsal sites B based on proximity to Twist (the spacer window), thereby placing a class tag C, on each of Dorsal binding site bases B. We calculate the shift w as:

$$w(B, C) = -\log \frac{P(B, C)}{P(B)P(C)} = -\log \frac{P(B|C)}{P(B)}. \tag{2.12}$$

The shift w is simply the Kullback-Leibler divergence of the conditional distribution $P(B|C)$ and the marginal distribution $P(B)$.

2.3.4 Energy estimation of a sequence of bases

We now extend the model from a single site to a binding site sequence. The total shift for a particular binding site sequence S and its flanking sequence is: $w(S, S') = E(S, S') - E(S) - E(S')$, where the shift is calculated as:

$$w(S, S') = -\log \frac{P(S, S')}{P(S)P(S')} \ . \tag{2.13}$$

The sequence S is the Dorsal binding site at a particular locus, and is a sequence of bases B, while the sequence S' is effectively a Bernoulli variable C, which means the flanking sequence S' of the Dorsal site either has a Twist site (in which case C=proximal) or not, in which case C=distal. Hence, $w(S,S') = w(S,C)$, which we define as:

$$w(S, C) = \sum_{i}^{s} w(B_i, C) \tag{2.14}$$

where we have defined S as the sequence $\{B_i\}$, where $i \in \{1, 2, 3 \ldots s\}$, and s is the length of the binding site sequence S. Equation (2.14) uses a standard PWM to calculate $P(S)$ (as opposed to using the marginal of S over C), because the marginal distribution is a "mixture model" that cannot be factorized into a product of base specific probability factors [10]. Computationally, for energy PWMs there is a matrix w for each class value of C. By adding the w matrix to the energy matrix $\mathbf{E}$ (matrix elements defined by Equation (2.10)) we obtain a *conditional* energy. We define a *conditional detector*, or a *conditional energy PWM*, which we use for bioinformatic predictions and annotations of binding site sequences. The detector

trained from sequences of class C then will score each sequence S as:

$$E(S|C) = E(S) + w(S, C) \ . \tag{2.15}$$

Here *E(S)* is from Equation (2.2), where the matrix elements E_{ij} are equal to $E(B(j)_i)$ from Equation (2.10). The function *B(j)* is a map between base B's alphabet A,C,G,T and the values of the matrix index j: 0,1,2,3; where we define the 0 index to be the consensus base and therefore reference energy level (the ground state). The matrix index i denotes the position of the base, which we previously denoted as B_i in Equation (2.14), where it was clear which particular base B resided at position i of sequence S. Hence the conditional energy is $E(B|C) = -\log \frac{P(B_0)}{P(B|C)}$, where B_0 is the consensus base of the position independent PWM from Equation (2.9).

2.4 Model detectors

We define two types of Dorsal binding site sequence models ("detectors") that we use for detection and classification. The first detector is conditioned on flanking sequence motifs, and hence potentially can better resolve functional loci. The second detector is simply a standard (unconditional) PWM model, which we use as a baseline for model comparison.

First we define the detector that incorporates flanking sequence information. As we will see,the detector acts like a logic-like gate that we call the "OR gate", due to its similarity with a standard digital OR gate used in electronics. The input to the gate is a k-mer, and the output is a decision on whether the k-mer is a Dorsal binding site or just random background DNA. The detector's decision is based on the conditional energy PWM scores

from Equation (2.15) described above, that is, its output depends on the output of two distinct "subdetectors", which we call DC ("Dorsal Cooperative") and DU ("Dorsal Uncooperative"). The DC component of the OR gate scores all incoming k-mers based on the conditional energy for a sequence with class type 'proximal', while the DU component scores all incoming k-mers based on the conditional energy for the class type 'distal'. The "OR gate" detector fires (that is, predicts a Dorsal site), if either the DC or the DU detector (or both) fire. In general, any energy PWM model (and hence our conditional energy PWMs) can be used as a linear classifier for binding site sequences. This classification is based on the following linear equation for any given k-mer:

$$y(\mathbf{S}) = E_c - \mathbf{E} \bullet \mathbf{S}, \tag{2.16}$$

Here $\mathbf{E}$ and $\mathbf{S}$ are vectors from a $4k$ dimensional real vector space, where we elevated the matrix of indicator variables from Equation (2.10) to be a bona fide vector. E_c acts as bias that shifts the hyperplane that separates putative functional sites from non-functional sites. The Euclidean dot product between the two vectors, $\mathbf{E} \bullet \mathbf{S}$, is defined as the sum over element-wise multiplications, where the energy $\mathbf{E}$ is now another vector in the space that projects each k-mer $\mathbf{S}$ onto a line of length *E(S)* (i.e., Equation (2.2)). The so-called bias or energy threshold is a positive real number (E_c), and represents a partitioning of the line defined by y into positive and negative real numbers. Here all k-mers with a positive value of y have energy less than E_c, and are classified as a binding site. All k-mers with a negative value of y have energy greater than E_c, and are classified as random DNA sequence.

The OR Gate detector is partially defined once the flanking sequence feature (the co-occurring binding site motif) and the spacer window have been set (or optimized), as de-

scribed in the Methods section above. These settings allow us to estimate the conditional probabilities. Hence, using only Dorsal binding site sequences from the data set $D_{\mathbf{CB}}$ we are able to train and define an OR Gate that is not mixed with binding sites based on purely bioinformatic matches. The second model is the standard PWM, which we call the *CB* detector, where *CB* stands for the "combined" set (meaning the conditional and unconditional data sets combined), which we denote by $\mathcal{D}_{\mathbf{CB}}$). The CB model assigns an energy score $E(S)$ to each sequence S as in Equation (2.2), which has a corresponding probability $P(S)$ as in Equation (2.1).

2.5 Results

2.5.1 Optimal spacer window for the OR Gate detector

In order to calibrate our conditional detectors we must define an optimal interval of the spacer window by calculating the mutual information between the known Dorsal binding sites and the potential cooperator's binding site (*eg.*, co-occurring Twist sites). The spacer window that leads to the maximum mutual information determines an optimal clustering of the Dorsal loci into two classes, which we then can use to build the OR gate.

We predict 5'-CAYATG loci (putative Twist sites) within the CRMs by scoring the CRMs with an energy PWM and threshold that corresponds to exact matches of the Twist motif 5'-CAYATG, which we found to have the highest mutual information with Dorsal binding site sequences. In the Supplemental section 2.10 we show a similar analysis with the alternative Twist motif 5'-CACATG, and some results for the motif's restricted form 5'-CACATGT.

Upon construction of the spacer distance matrix we are able to classify all annotated Dorsal sites as 'Cooperative' or 'Uncooperative', based on whether any of the spacers for a

given Dorsal locus was within the bin border defined by $d_{\min}$ and $d_{\max}$. For example, a CRM annotated with one Dorsal site and three Twist sites will have three spacers. If any of those spacers are within the spacer window, then the Dorsal site is classified as 'Cooperative'. We define the spacer window as a 30 base-pair closed interval, which starts at [0,30]bp relative to each Dorsal coordinate within the CRM (not counting the body of the binding site as a part of the spacer).

All known Dorsal loci of a given class are then aligned (see Supplement section 2.9.9 for details) to construct the conditional Dorsal binding site sequence distribution (conditional PWM). Given the class labels on the Dorsal sites, we are able to estimate the probability of a given class as simply the fraction of Dorsal sites that belong to each class C. With these distributions we are then able to calculate the mutual information, *I(S;C)* between the Dorsal site sequence variable S and the class C as

$$I(S;C) = \sum_{S} \sum_{C} P(S|C)P(C) \log \frac{P(S|C)}{P(S)} \tag{2.17}$$

where $P(S|C)$ is the conditional PWM, and $P(S) = \sum_C \prod_i P(C)P(S_i|C)$ is the marginalized distribution of sequence over class labels C (note this is not the same as the CB detector's probability). As stated above, the initial $d_{\min}$ was set at zero and $d_{\max}$ at 30bp, and then both parameters are incremented by 30bp to shift the window to a new position. For each shift of the spacer window we classify all Dorsal loci, align each class to a length 9 motif, and then calculate the mutual information. The result is shown in Table 2.1 and implies that the information between sequence and class label is highest if the spacer is between 0 and 30 bps, as expected for binding sites that interact via molecular interactions. Furthermore we appended one nucleotide of flanking sequence on each binding site sequence to see if we

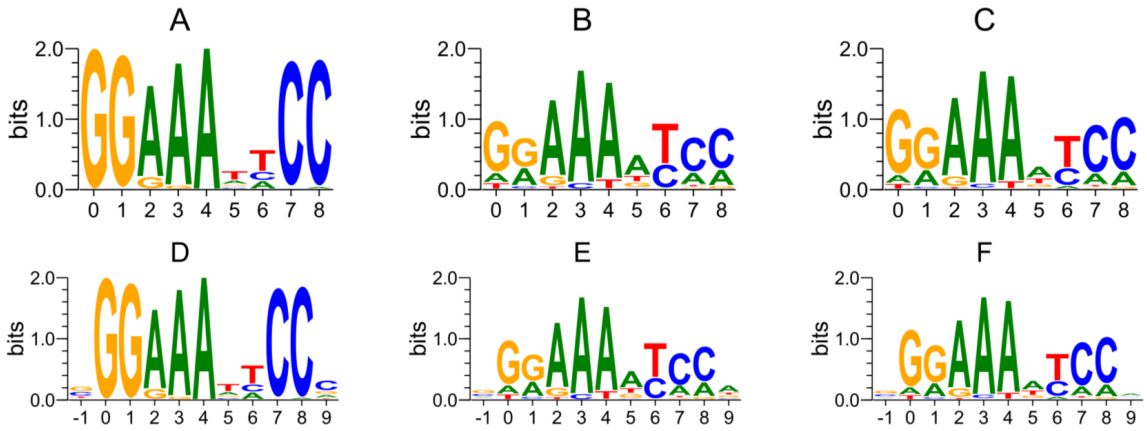

Figure 2.1: Logos generated for known Dorsal sites (the D_{CB} data) tested for adjacency to 5'-CAYATG used as the cooperative class if in the [0,30]bp distance. Logo A corresponds to the cooperative class, and displays the known 5'-AAATT core, with total information content 13.5 bits. Logo D is the exact same logo as A but with a single base-pair of flanking sequence at the start and end of the site (hence, this logo starts at position -1). Position 9 of this logo shows about two decibits of information relative to the background sequence in the nucleotide base 'C' (2 out of 10 functional DC sites have a 'C' at this position). Logo B is the 'uncooperative' class for the [0,30]bp window, which we calculated to have 9.1 bits information relative to the background (uniform distribution of bases), and logo E has the added flanking sites to the 'uncooperative' class. Logo C is the CB motif with 9.6 bits of information relative to the background, which looks similar to the 'uncooperative' class at position 6 due to there being many more sites that prefer A to a T at this position amongst all the Dorsal sites in the network. Logo F is the CB motif with the flanking sequence appended.

were missing flanking parts of the conditional binding sites.

spacer	[0,30]bp	(31,60]bp	(61,90]bp
Mutual Information, Equation (2.17)	0.49	0.29	0.04

Table 2.1: Mutual Information between functional Dorsal binding site sequences and putative Twist sites that match 5'-CAYATG using a sliding spacer window scheme.

We show the conditional Dorsal binding site sequence logos for functional binding sites generated for this first spacer window in Figure 2.1. The information content of each position of the binding site corresponds to the height of the logo, where we used a symmetric hyperparameter value of $\beta = 0.1$ as discussed in the Supplement section 2.9.11 and section 2.11.4.

2.5.2 The conditional and unconditional PWMs are significantly different

Here we test the optimal DC and DU detector's training data energy scores to see if the median energy of DC is significantly different than the median energy of DU. The optimal detectors were based on the 5'-CAYAGT Twist motif and the [0,30]bp window. The rank sum test rejected the null hypothesis that the median energies are equal with $p = 10^{-26}$. The median energy of the DC PWM was 0.27, while the median energy of the DU PWM was 2.7.

It is possible that any random partitioning of a set of binding sites that are used to build detectors using our technique would produce p-values consistent with significance. We used our original data set of Dorsal sites $\mathcal{D}_{\mathbf{CB}}$ to construct a sampling distribution of p-values for the rank sum test To calibrate the p-value we created a sampling distribution of the p-value from 1000 repetitions, where at each repetition the combined data $\mathcal{D}_{\mathbf{CB}}$ were randomly partitioned into two data sets. PWMs were constructed for each partition. The energy of each sequence within a partition was calculated as $E(S) + w(S, P)$, where P is the partition, S is a sequence in the partition, and $E(S)$ is the CB energy. We then determined the corresponding rank sum p-value between the random sets. We found that the p-value of the rank sum test between the DC and DU model fell well beyond the right tail of the random sampling distribution (shown in Figure 2.2), indicating that the median energies of DC data set and the DU data set are significantly different from *any* random partitioning of the combined data set. More details are in the Supplement section 2.10.1.

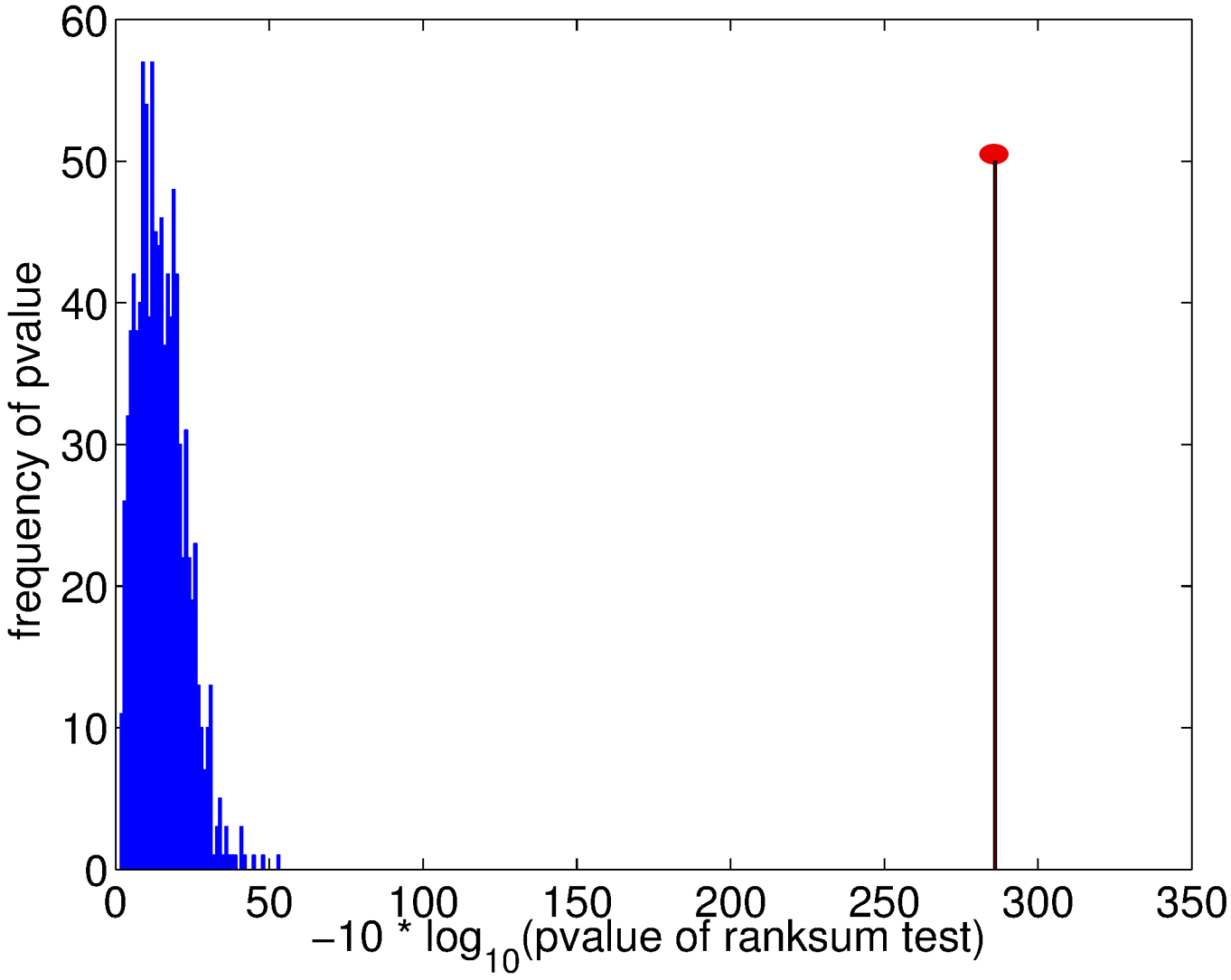

Figure 2.2: Histogram of p-values of a rank sum test of random partitions of the combined data set $\mathcal{D}_{\mathbf{CB}}$. The binning is in units $-10 \times \log_{10}$ of the p-value, rounded to the nearest integer. The p-value of the rank sum test between DC and DU energy data sets based on their energy PWMs was 260 in log base ten units (scaled by 10), which is indicated by the red bar of arbitrary height.

2.6 Performance of optimal classifiers (detectors)

All detectors were built from length 9 alignments (see Supplement section 2.9.9for details of the alignment procedure). The OR gate is based on the DC detector built from the data set $\mathcal{D}_{DC}$, which contains Dorsal loci from $\mathcal{D}_{\mathbf{CB}}$ that were tagged with class labels from the optimal spacer window of [0,30]bp with the 5'-CAYATG motif, and similarly, the DU detector is built from the data set $\mathcal{D}_{DU}$, which contains the remaining Dorsal loci from $\mathcal{D}_{\mathbf{CB}}$ that did not have the Twist sites in the spacer window. . The unbolded subscripts DC and DU on the data sets denote that these sets of Dorsal sites were based on our clustering scheme (not based on literature annotation).

We now present three experiments that tests the performance of our OR gate detector and the conditional detectors using the CB PWM as a benchmark.

2.6.1 The DC detector predicts sites proximal to 5'-CAYATG with better odds than the DU detector.

We expect that DC should predict Dorsal binding site sequences that are adjacent to Twist more precisely than DU (since we showed earlier that the Dorsal site sequences contain information about adjacency to Twist). In Table 2.2 we collected all the hits (all the positives) of the detectors. We test whether the DC conditional energy PWM is actually *predicting* Dorsal sites within the CRMs that have the correct flanking sequence feature (presence or absence of Twist motif) with better odds than the DU detector. The odds of DC for predicting binding site sequences that belong to the proximal class was $\frac{61}{39} = 1.6$. The odds of DU for predicting sequences of the proximal class is $\frac{280}{345} = 0.81$, hence the odds ratio is 2.0. The one-sided p-value for this table's log odds ratio test is $p = 0.001$ for the chances of

seeing a DC detector with better odds relative to DU at predicting correct flanking sequence features. Increasing the energy cutoff E_c increases the total counts of the table, and we obtain similarly significant tables up until about $E_c = 5$.

	proximal	distal
DC	61	39
DU	280	345

Table 2.2: Contingency table with the conditional detectors DC and DU represented along the rows and the class type distal and proximal represented along the columns. Each table element represents the number of sites predicted from each detector of each class type based on Twist sites (5'-CAYATG) and a CB energy cutoff $E(S) = E_c = 2.1$.

2.6.2 Both OR gate and CB detectors show high sensitivity with known sites as positives and CRM sequences as negatives

In order to test the sensitivity and the specificity of the detectors we used the Receiver Operator Characteristics (ROC), which displays the tradeoff between optimizing predictive performance for 'positives', while also optimizing for not detecting known 'negatives'. The True Positive Rate (TPR) is defined as $TPR = \frac{TP}{(TP+FN)}$, where the denominator is the total counts of True Positives (TP)and False Negatives (FN)). The False Positive Rate (FPR) is defined as $FPR = \frac{FP}{(FP+TN)}$, where the denominator is the total counts of True Negatives (TN) and False Positives (FP).

We use the data set $\mathcal{D}_{\mathbf{CB}}$ as our training set of 'positives' ($TP + FN$) for both the CB detector and the OR gate. The 'negative' data set ($TN + FP$) is the set of all CRMs that contained a known binding site (i.e., the CRMs associated with $\mathcal{D}_{\mathbf{CB}}$), where the bona fide sites (the functionally confirmed sites) are masked out. Furthermore, within the CRMs we also mask out overlapping predicted binding sites based on the algorithm in the Supplement section 2.9.6, hence the negative data (the CRMs with known sites masked and overlapping

hits masked) is at least nine fold smaller than the concatenated length of the CRMs due the binding sites being nine base pairs in length.

For a given energy threshold, $E_c = E(S)$, set by the CB energy PWM for both the OR gate and the CB detector, each detector 'scans' the CRM using a sliding window approach, where each 'hit' of the detector is classified as a TP if the hit overlaps a known binding site locus in $D_{\mathbf{CB}}$, and as a FP if the detector 'misfired' in the background of the CRM. Similarly, known sites (loci) from $D_{\mathbf{CB}}$ that were not called hits by the detector are classified as FN, while TN are the k-mers from the CRM background sequence that the detector did not call a hit.

The ROC of the OR gate (shown in Figure 2.3A) tends to perform better than the CB detector at low energies up until the energy reaches about $E(S) < 8$ (the last point (FPR, TPR) displayed in the figure), after which the CB detector tends to do better. The OR gate in the region of ROC space displayed shows better performance than the traditional CB detector (This is clearer quantitatively, where we found the OR gate had a higher area under the curve (AUC) integrated from the minimum energy to CB's energy cutoff of $E(S) < 8$ (which is the last point displayed in ROC space)). The OR gate and CB detector both perform well for strong sites (low energy sites), which is indicated by their good TPR (almost 80% before a noticeable fraction of negatives start to be detected as positive.

2.6.3 The OR gate performs better than CB at predicting known sites at lower energies

Another metric of performance of the classifiers is the mutual information between the type of k-mer (Dorsal or not Dorsal) and the classification by the detector. For example, if the

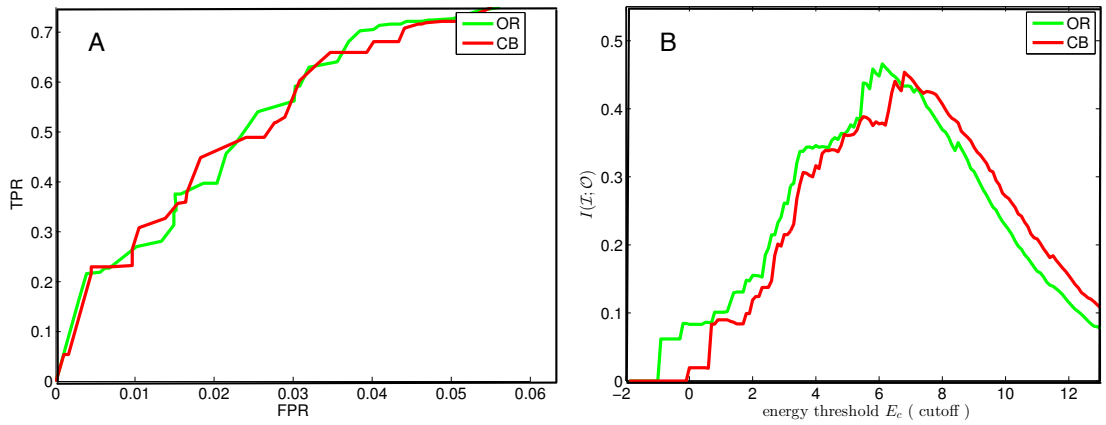

Figure 2.3: ROC and Information. (A) False positive rate (FPR) vs. True Positive Rate (TPR) when varying the energy cutoff E_c. (B) shows the mutual information $I(\mathcal{I};\mathcal{O})$ Eq. (2.6.3) between the input and output of the detectors as a function of the cutoff energy.

input is not a Dorsal binding site, the detector should stay silent, while it should fire if it is a Dorsal site (either adjacent to Twist or not). We can write this mutual information as

$$I(\mathcal{I};\mathcal{O}) = H(\mathcal{I}) - H(\mathcal{I}|\mathcal{O}), \tag{2.18}$$

where $\mathcal{I}$ is the binary random variable holding the true identity of the 'Input' k-mer received by the detector, while the 'Output' variable $\mathcal{O}$ is the binary variable given by the detector's decision. The entropy $H(\mathcal{I})$ is in principle given by the relative likelihood to find Dorsal binding sites within the ensemble of CRMs, which is of course heavily biased towards negatives (non-Dorsal sites). However, this Bayesian prior is not available to the transcription factor, in other words, for each decision to bind, the factor has its own Bayesian prior p, which we will set to $p = 1/2$ (maximum entropy Bayesian prior) below.

The conditional entropy $H(\mathcal{I}|\mathcal{O}) = -\sum_{i,o=0}^{1} p(i)p(i|o)\log p(i|o)$ quantifies the remaining uncertainty about the identity of the k-mer given the decision of the detector, and can be

calculated using the false positive and true positive rates introduced earlier. In particular, the conditional probability $p(i|0)$ is obtained as

$$p(1|1) = p(\mathcal{I} = 1|\mathcal{O} = 1) \qquad = TPR \tag{2.19}$$

$$p(1|0) = p(\mathcal{I} = 1|\mathcal{O} = 0) \quad = 1 - TPR \tag{2.20}$$

$$p(0|1) = p(\mathcal{I} = 0|\mathcal{O} = 1) \qquad = FPR \tag{2.21}$$

$$p(0|0) = p(\mathcal{I} = 0|\mathcal{O} = 0) \quad = 1 - FPR\,, \tag{2.22}$$

while $p(i)$ is the Bayesian prior (density of Dorsals/non-Dorsals in the CRM). Using an arbitrary prior p, we can rewrite the mutual information from Equation (2.18) as:

$$I(\mathcal{I};\mathcal{O}) = H[p] - pH[TPR] - (1-p)H[FPR]\,, \tag{2.23}$$

where $H[*]$ is the usual binary entropy function of a Bernoulli distribution characterized by *, so for example

$$H[TPR] = -\frac{FN}{TP+FN}\log\frac{FN}{TP+FN} - \frac{TP}{TP+FN}\log\frac{TP}{TP+FN}\,, \tag{2.24}$$

with a similar expression for $H[FPR]$. We show the mutual information $I(\mathcal{I};\mathcal{O})$ in Figure 2.3B using the maximum entropy Bayesian prior $p = 1/2$. Compared to the information the CB detector has about Dorsal sites, the OR gate's information is shifted to lower energies, implying that at fixed energy cutoff it knows Dorsal sites better than CB.

DC conditional detector is able to predict that Twist is nearby

The conditional detectors are expected to make predictions not only about what is a Dorsal site relative to the background, but also whether Dorsal is in the vicinity of Twist. By partitioning all the known sites into the two class types (e.g., 'distal' and 'proximal') as determined from the spacer window of [0,30]bp and Twist motif 5'-CAYATG, we can test how well each detector can resolve the class type of a Dorsal site (Dorsal with Twist or without).

For a given energy threshold we scanned the combined data set $D_{\mathbf{CB}}$ with the DC as well as the DU detector, and asked how much the detector knows about the class variable $\mathcal{C}$ (further details of this experiment are in Supplement section 2.10.3). We show this mutual information $I(\mathcal{C};\mathcal{P})$ in Figure 2.4 , where $\mathcal{P}$ is the binary random variable encoding the detector's decision about the context. We see that the DC detector has up to 0.3 bits of information about the proximity of Twist in any particular Dorsal site, while the DU detector has virtually no information about this variable.

2.7 Discussion

2.7.1 DC and DU Information logos and previous evidence

The binding site sequence logos display the information content of our binding site data relative to a uniform distribution. By inspection of the DC logo the consensus sequence (highest information scoring sequence) is partially consistent with Table S2 of Crocker et al. [25]. The 5'-AAATT core is reproduced as our DC consensus sequence, while the flanking sequence for the length 11 binding sites are not enriched with G at the start of the site and

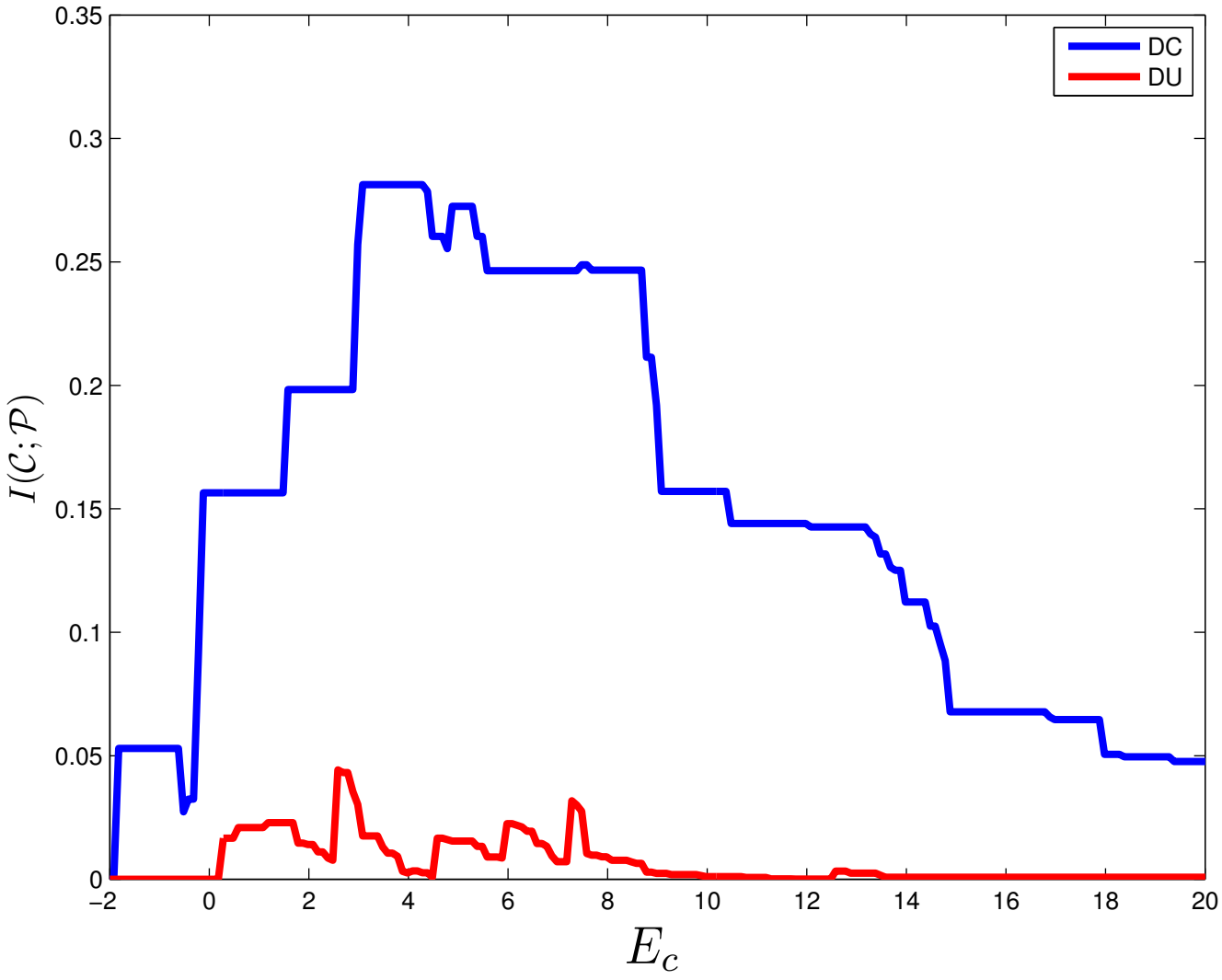

Figure 2.4: Mutual information $I(\mathcal{C};\mathcal{P})$ between the actual classes $\mathcal{C}$ and the predicted classes $\mathcal{P}$ for Detectors DC and DU as a function of the threshold energy E_c that is defined by each detector's conditional energy Equation (2.15).

a C at the end of the site. Similarly we can see that our DU also conforms roughly to A-tract Dorsal binding sites, which are Dorsal binding sites that have four or more contiguous Adenines. Mrinal pointed out that A-tract binding sites have certain physical chemical properties not seen in 5'-AAATT core Dorsal sites [78], namely that A-tract Dorsal binding sites encode a mechanism (like an extra hydrogen bond between the protein and DNA) for Dorsal to switch roles from an activator of gene expression to a repressor of expression based on the binding site Dorsal was occupying. Of course, as mentioned by Mrinal, these sites are still context dependent, namely the context of a site may override any preference a binding site sequence has for causing activator or repressor roles[82]. Inspection of our DU detector's data set shows that it is more than 50% enriched with Dorsal sites that are known to be from repression *cis*-regulatory elements (*zen*, *tld*, *dpp*), hence the DU logo with a 5'-AAAAT core is not surprising.

Our known binding sites, to a degree, come with the class labels already attached. The $\mathcal{D}_{\mathbf{DC}mel}$ data is the known Dorsal binding site data set based on the definition of D_β or 'specialized' sites, or NEE-like Dorsal binding sites (neuroectoderm Dorsal sites that were linked to Twist sites, but were not linked to the canonical 5'-CACATGT Twist sites) [25, 35]. However, our DC detector is different than a detector built strictly from the $\mathcal{D}_{\mathbf{DC}}$ data set (the set of all 12 orthologs for each *melanogaster* locus), since we included additional ortholog CRMs of the NEEs.

Furthermore, within the NEEs one could imagine that the spacer has diverged in species that we analyzed that were not analyzed previously, and our choice of the spacer window is an interval not the same as previous choices. For example, Papatsenko et al. [84, 83] showed that binning the spacers between Dorsal and Twist that there were various optimal bins (namely 14bp, 20bp, and 53bp). It is also possible that the spacer defining the distance of

Dorsal and Twist in *D. melanogaster* has further diverged in its ortholog species, in particular those not previously analyzed and annotated.

Szymanski *et al.*[102] used DU-like Dorsal sites in his systematic study of the role spacing has between Dorsal and Twist site, suggesting that Dorsal Twist still cooperate if one uses a DU-like binding site, which is further corroborated by systematic studies from Fakhouri *et al.*[36] that also used A-track Dorsal sites for the primary Dorsal sites. These studies suggest evolution could have fixed either a DC or a DU type site at an NEE locus utilizing Dorsal Twist linked sites for synergy, which would deteriorate our claim that DC and DU are really different types of Dorsal sites. However, it is highly unlikely that all these sites would have fixed with the same sequence unless they were functional or else if the CRMs containing them were duplications.

2.7.2 The OR gate and the CB detector

The OR gate scores any input k-mer with both conditional detectors DC and DU, and then outputs simply the lowest energy score. Similar detectors have been represented in the literature as a Hidden Markov Model or as a mixture model [79, 49]. Each component of the mixture is simply a conditional PWM, where the mixing frequencies are estimated as the fraction of training data that is associated with a particular component (or class) of the mixture. The mixture is defined as:

$$P(S) = \sum_{c} \frac{\exp -E(S|C=c)}{Z_c} P(C=c), \tag{2.25}$$

where $E(S|C)$ is in units of λ_1 (which is further assumed to have been calibrated to thermal energy units), and $Z_c = \sum_{S \in \mathcal{S}} \exp E(S|C=c)$, where $\mathcal{S}$ is the set all possible k-mers, $|\mathcal{S}| =$

4^k.

The CB detector is the traditional position independent probability model (PWM) of binding sites, where the PWM is constructed by aligning all of the sites in the $\mathcal{D}_{\mathbf{CB}}$ data simultaneously. Recall from Equation (2.1)

$$P(S) = \prod_i P(S_i). \tag{2.26}$$

where, as a consequence of Bayes' Theorem $P(S_i) = \sum_c P(S_i|C = c)P(C = c)$. However, for a sequence of bases, $\prod_i P(S_i) = \prod_i \sum_c P(S_i|C = c)P(C = c) \neq \sum_c \prod_i P(S_i|C = c)P(C = c)$, where the last expression is the mixture of Equation (2.25), and is equivalent to a marginalization of the sequence over the classes. The mixture distribution of the sequence over classes can only be factorized as a product of position distributions *given* the class. We justify our approximation of the marginal sequence distribution over classes as a PWM (the CB PWM) in the Supplement section 2.10.3.

The mixture model was used by Hannenhali et al. [49] in a similar form as the OR gate, where a given transcription factor's binding preference was described by two PWMs. There the authors scanned a given CRM or promoter with both PWMs and selected the highest scoring sites as hits, where the threshold for a hit was determined by the mixing frequencies– the proportion of known sites that are used in constructing each PWM. Upon scoring all the sites within their promoters, the scores were ranked for a given PWM, and then the fraction of sites equal to the mixing frequency were considered positives. This method is different than the OR gate presented here in that we do not use the mixing frequencies in discriminating Dorsal binding sites from background DNA. The OR gate discriminates sites from non-sites by checking if the minimum (i.e., best) score of the component detectors is

below the energy threshold. By always choosing the lowest energy score among the given components as the detector's overall energy score, the benefit of an increased True Positive Rate of the detector is partially cancelled by the cost of an increased False Positive Rate. However, this cost is only in effect at high energies (non-specific sites), where it is unlikely that evolution or physical binding is having any functional effect on the organism. Hence, the OR gate is a useful model for increased sensitivity in the low energy regime.

2.7.3 The CB data set, merging and dividing clusters of binding sites

The mixture distribution of equation of Equation (2.25) implies the data set $\mathcal{D}_{\mathbf{CB}}$ over all loci can not be aligned simultaneously to form an estimate of $P(S)$, as that only makes sense if one is constructing the CB PWM, which assumes no mixture. However, a priori, one may not know whether their set of binding site loci is a mixture of different types. To determine if there is a mixture in the data one must decide on how one will align the mixture model, and whether that alignment should be related to the case that one combines all the data indiscriminately to form an alignment for the CB PWM. Will all the training data over all classes be lumped together to estimate $P(S)$ for CB, and then the conditional probabilities estimated by partitioning their respective set of aligned sites? This technique is commonly used in the case that one is given a set of *aligned* data, and one wishes to find mixtures within the aligned data. Alternatively, will the training data be partitioned into the classes, and then each class aligned individually, and then these class-specific alignments simply be 'merged' to form an estimate of $P(S)$ for CB?

This additional complexity is analogous to the decision made in clustering motifs as to

whether one wants a top-down approach (start from the root and *partition*), or a bottom-up approach (start from the leaves of the tree and *merge* (i.e. combine)), see for example [47][30][57]. We presented results for a bottom-up approach that aligns the training data $\mathcal{D}_{DC}$ and $\mathcal{D}_{DU}$ separately to estimate each conditional PWM DC and DU, then CB was based on merging the count matrices of the DC and DU data sets.

The top-down approach aligns all the sites together, then partitions out the classes, and from those partitions builds (without further alignment) the conditional probability PWMs, $P(S|C)$. The bottom-up method is guaranteed to achieve higher Mutual Information than the top-down approach. This is because the DC alignment will not necessarily be 'in register' with the independent DU alignment, for example the DU alignment may tend to have the first character with more than 0.5 bits of information shifted relative to the DC alignment (i.e. the binding start site of these two set are shifted). This in turn causes their marginalization to have an increased entropy due to mixing alignments out of register, which in turn causes the mutual information to be boosted, since the conditionals are both now substantially different than the marginal due to registration of the start sites. Based on results not presented, we found that the top-down approach still preserves the overall trend in Mutual Information verse spacer window, although the signal in the [0,30]bp window for distances of known Dorsal sites from the 5'-CAYAGT motif is reduced by about 40%.

From a model comparison viewpoint, one may assume the strategy should be to align DC, DU and CB all separately, neither taking a top-down or bottom-up approach. This strategy does not bias either the mixture model (OR gate) or the CB PWM model. However, from results not presented, we found this has little effect on the logos for a length 9bp alignment. Albeit, some alignments (in an ensemble of alignments derived from gibbs sampling) do choose a 'T' rich motif, as opposed to an 'A' rich motif. For our logos, in the case that

a 'T' rich motif was found, we presented the logo derived from the reverse complement of each *aligned* sequence within the training data set so that all logos would be easily visually comparable by inspecting the logos. For longer than 9bp length alignments of Dorsal binding sites, we found that the Conditional Dorsal motifs frequently were not in 'register', where registration of the start sites of motifs is based on a motif-motif alignment program like STAMP[72], but can also be seen by inspection of the logos (sometimes). For example, by inspection, the DC motif for a length 11bp alignment may have had the first position in the alignment with greater than 0.25 bits of information content at position one (in a zero based coordinate system), while DU would have its first position in the logo with more than 0.25 bits at position zero (the start of the logo).

From a physical standpoint, it may be that the conditional binding sites do tend to have a shift in their binding start site positions. For example, Dorsal may bind to : 5'-AAGGAAATTCC in a DC favored environment, while in the DU favored environment it binds: 5'-GGAAATTCCAA. If the flanking A's really are a signal, then one must conclude the best representation for CB would be: 5'-AAAGGAAATTCCAAA, which is a motif that is 3 nucleotides longer than either conditional. The bottom up approach to clustering would miss this signal, since it would merge the two classes such that the CB PWM would contain a fraction of 'A's at the first position based on the proportion of the DC data relative to the CB data, and another fraction of 'G's at the first position due to the DU motif; thereby not only missing some of the signal, but also interfering with the captured portions of the signal. This particular case shows that trying to compare motifs based on having the starting position of the motif being in register, will potentially truncate a signal for the CB model. The logos in Figure 2.1 are in register (the start of the binding sites are the same), which is based on our findings that length 9 alignments are the most reproducible in terms of registration. Once

we had aligned the length 9 binding sites, we then appended the flanking sequence to each already aligned locus, thereby having greater assurance that the CB PWM would not have this type of interference effects.

However, if one starts with an alignment that has flanking sequence to begin with, (such as an alignment of length 15bp;) then one could try and discover if the *aligned* sites do contain a mixture of motifs without having to worry about the problems associated with choosing a strand (such as the 'A' rich strand), or whether the start sites are in register. Such an approach was used in Figure 2 of Barasch et. al. [10]. However, we found that setting the Gibbs sampler alignment's length parameter highly influences the alignment. For example, for a length 9bp alignment, the starting position of the alignment may contain maximum information (2 bits), however if one creates a length 15 alignment this signal (the information) is at least conserved, but may spread into the flanking sequence. This spreading changes the DNA makeup of the logo of various positions. This is partly due to there being so many more ways to spread out information among the positions of the alignment when one is using an objective function that runs over 15 positions as opposed to 9bp. For example, a subsequence of 5'-AAA that is completely conserved in a length 3bp alignment, when allowed to be length 5bp alignment may converge on 5'-AAAAA with the same information content as the 3bp alignment (or slightly larger information content than the length 3bp alignment, while having a smaller per position information content). This is due to a mixture over loci of the form: 5'-NAAAN, 5'-AAANN, 5'-NNAAA.

2.7.4 Mixtures of asymmetric PWMs

In the case that one's data set of binding sites is primarily asymmetric, yet completely conserved at each position; then one can construct a two component mixture of PWMs

that are significantly different by simply choosing each component PWM to represent each concensus of the two asymmetric motifs. For example if one's data set is simply a collection of loci from a database where the sequences from the database were all identical of the form 5'-GGAAACC, then one could take half the loci to be 5'-GGTTTCC, and the other half 5'-GGAAACC. We found that when training the mixture model where one has no knowledge about what strand to use in the component PWMs (i.e. DC and DU) that DC and DU frequently would converge on motifs that appeared to be reverse complements of one another. To resolve this problem one can symmatrize the PWM by using both strands in the construction of the PWM (see for example the paper by Bailey et.al.[8] and the discussion of MNT operator sites by Fields et.al.[37]).

A hypothetical DC concensus 5'-GGAATTTCC may seem to be a problem for symmetrization if the 'T' at position 4 (where the start site is position zero) would lose some of its signal by symmetrization. However, in a sense, this loss would be compensated by a gain in its complement's 'A' signal if one considers the symmetrical PWM to contain a smaller alphabet of symbols. [3] Even for highly non-palindromic sites such as the 'A-rich' core in the Dorsal DU sites, one could still symmatrize the motif and hence not have to deal with issues of strandedness of sites when forming mixtures of PWMs. However, there are a number of reasons why mixtures of asymmetric PWMs is important, assuming one can deal with the potential artefacts such as a loss of the signal due to choosing opposite strands (such as in the 5'-GGAAACC data set example above which would artificially boost the mutual

[3]This compensation may seem to be violated due to the loss of information content when one symmatrizes a pure 'T' signal at position 4 of the DC motif. However, the source of this apparent violation is due to comparing signals from two different sample spaces. For example, the original alphabet of 'A,C,G,T' has at most 2 bits of information. However, when one uses a symmetric PWM the alphabet size is no longer over 4 characters, but simply over 2 characters, hence the most the information can be is now 1 bit. Hence if one compares normalized information contents (normalized by alphabet size), then the normalized information contents are preserved. From another perspective the choice to symmatrize is an additional bit of information gained that compensates the bit of information lost by the original 2 bit 'T' signal.

information).

Our decision to treat the component PWMs as 'asymmetric' PWMs was based on the assumption that the orientation of a site relative to cooccurring sites may be important, hence by using asymmetric PWMs we would allow ourselves the potential to decode any orientation information if it were encoded in the PWM(see Shutzaberger et.al. for more ideas on this topic of strand[95]). The potential to decode context information based on the strandedness, is equivalent to making predictions that a particular type of cooccurring binding site would be either 5' or in the 3' direction of the annotated binding site. This decodable information would also be physically related to cooperative or antagonistic interactions that are contingent on the binding site sequence. An example of this is the asymmetric extended Twist site 5'-CACATGT (extended from the E-box sequence CACATG, where a 'T' is concatenated to the 5' end of E-box), this site has evidence that the 'T' at the 5' end must cooccur with a Dorsal site *downstream* of it (i.e. where downstream means in the 5' to 3' direction)[108]. For example, one would always find evolution selected 5'-CACATGTNNNNGGAATTTCC and never find 5'-GGAATTTCCNNNNCACATGT. Furthermore, one could imagine that the Dorsal site that cooccurs on the same strand in this case is 5'-GGAATTTCC, as opposed to 5'-GGAAATTCC, where position 4 is strand dependent. In this case, had one symmatrized the DC PWM they would have lost the information at position 4. Hence the higher information content of the asymmetric PWM is not an artefact of the asymmetric PWM [4], it's rather telling us that this is physically due to the preference of the 'A' at position 4 of the DC motif cooccurrring on the same strand as the annotated Twist site (i.e. 5'-CACATGT). Similarly the DU concensus' core '5-AAAAT' may be cooccurring on the

[4]This may be an artefact in that asymmetric PWM's always have greater (or at least equal) information content then symmetric PWMs

same strand as other coregulating trans factors, and hence cooccurring with other motifs on the same strand.

2.7.5 Comparing models with unequal parameters

For optimization and model comparison, models that have more parameters may be guaranteed to have a higher likelihood value for a given data set, just as a Taylor expansion approximations of a function improves when one includes higher order terms. Assuming one has sufficient data to stay clear of fitting parameters to noise, and that the more complicated models are a more accurate description, a more complicated model may be useful. However, it is not true that the OR gate, must perform better than CB due to the increased number of free parameters.

In detection, ensemble classifiers (such as the OR gate) can be used to increase the detection performance[17]. However, it is not true that *ensemble detectors* will always perform better than a single detector. The OR gate is a form of the more general ensemble detector that is a machine that produces a decision based on component detector decisions, where the component detector's decisions are pooled and processed in order to come to one overall decision of the ensemble detector. Examples of processing the individual detector's decisions are 'majority vote', where the most frequent decision of the collection among the component detectors is used as a the final machine's decision, similarly the OR gate is an example of an ensemble detector. Ensemble classifiers have a necessary condition to perform better than a single classifier: the component classifiers (e.g. DC and DU) should be independent of each other[17].

DC and DU are independent if we use a conditional independence model for the joint distribution of an extended sequence S, where we now think of the extended sequence S as a

dyad (which can be decomposed into three variables that denote two bindings sites separated by a variable spacer, like in the framework of BioProspector). In this probability model over the joint we partition an extended sequence S (like a CRM) into two parts, S' and C, where S' is the k-mer sequence to be tested as a Dorsal site or background, and C is the class determined by the *spacer window*, which either contains Twist or it doesn't.

Given that we observe the class C, then we can factorize the joint distribution of the sequence S' being tested as a Dorsal site (i.e. S' is no longer a position dependent distribution). Without the above assumption DC and DU are dependent, hence the OR gate, which does not observe the class C, does not have independent component detectors. A consequence of this dependency is that the component detector's errors are correlated (i.e. misclassification errors are correlated). For example, each component classifier's probability of a base 'G' at a given position in a binding site should be just as likely to be above the true value of the probability of G, P(G), as it is below the true value; while for correlated detectors this does not occur.

Hence, it is not always true that increased complexity of a model will increase its performance on training data. Mixture models of binding sites are more complex than a standard PWM, and are best at capturing broad dependencies among the component classes. To a large extent, the difference of DC and DU is a localized dependency at particular position (which can be seen as the difference of their cores '5'-AAATT' and '5'-AAAAT'), and a differences in the 'strength' of the sites (or the energy level spacing). The OR gate seems to perform better than CB for the low energy regime, where the localalized difference between the cores of DC and DU is significant in distinguishing the energies of the concensus sequence of each conditional PWM. However, for the entire energy spectrum, where AUC can be used as a measure of performance, it seems the OR gate behaves about the same

as CB, suggesting that DC and DU are not independent models. This is not surprising, as nonlinear models (like a mixture) are sensitive to certain regimes over their input, and similarly insensitive to certain values of their input (for example see Section 3.5.2.

2.7.6 Information that detectors have about Dorsal binding sites

In a physical NVE ensemble (fixed particle number N, fixed volume V, fixed energy E) the information content of the distribution of momentum and positions (the distribution function) is conserved. This means the number of bits necessary to store the position and momentum information is conserved in time relative to the maximum storage capacity defined by a lattice over phase space (the space of coordinates). For example, if the distribution function is a uniform distribution over phase space, it has zero information content.

Similarly, evolutionary systems under adaptive maintenance (purifying selection) conserve information stored in their genes [1]. The inheritance of information implies that parents pass a fixed number of bits to their progeny. And just as in the NVE ensemble where coordinates and momentum are not conserved, similarly in evolution sequences are variable, but the sequence's information content is conserved. However, when the fixed energy constraint of the NVE system is relaxed and the system exchanges energy with a much larger environment, the system's original information content may deteriorate until the system equilibrates with its surroundings. Biological systems harness energy from their environment to maintain their information content in the never-ending fight against the second law [2, 23].

The mutual information between sequences and the OR gate's predictions in Figure 2.3 suggests that the conditional distributions of functional Dorsal binding sites have encoded synergistic and antagonistic information about flanking sequence features (presence of Twist)

that causes the likelihood to correctly predict the presence of Dorsal to shift downwards in energy (as observed by the shift of the mutual information of the OR gate relative to CB in Figure 2.3). This shift may have been a necessary adaptation in the way Dorsal regulates its targets. For example it is possible that at the phylum level, possibly before the neuroectoderm evolved, Dorsal only needed to regulate the mesoderm and ectoderm. When the neuroectoderm evolved, Dorsal evolved the ability to recognize two subtypes of binding site ensembles, a function that would help to resolve the neuroectoderm Dorsal targets from the more ancient germ layers (mesoderm and ectoderm). In this sense, Dorsal's adaptation to its local environment is seen as the shifted mutual information relative to the CB detector (which just treats all binding loci identically). Dorsal could then use this information to its advantage, in Dorsal real time so to speak, to make better decisions about binding.

The shift in the mutual information plot in Figure 2.3B is not as visible in the ROC curves in Figure 2.3A, in which we used the same TPR and FPR for the detectors. This is because, in general, energy level spacing is not accounted for in an ROC curve, implying that detectors with similarly ranked sequences may actually have different spacings between their energy levels, and the minimum energies of the scales may be shifted relative to one another. For example, DC's ground state is below CB's ground state, which is why the OR gate contains some information at negative energy (as DC's ground state is at about -0.8 in energy units as seen from the horizontal axis of Figure 2.3).

The degree to which the OR gate's ROC does appear shifted relative to CB's ROC in Figure 2.3A is partly due to the fact that the ranking of sequences of the DC detector and DU detector is very similar; it is the energy level spacing that is dramatically different between the conditional detectors. For example, using a substitution model that penalizes all mismatches from the consensus sequence with the same energy score (see the Appendix

of Ref. [13] for details) leads to the elegant formula that a consensus base occurs with probability $1 - \frac{m}{3k}$, and that an error or substitution occurs with probability $\frac{m}{3k}$, where k is the length of the sequence and m is the number of mismatches from the consensus (the 3 in the denominator is due to the three ways a mismatch from a consensus DNA nucleotide can occur). Weak sites will be seen to have large m, which to a degree can be seen as the DU training data. Similarly, strong sites will have small m, which can be seen as the DC data. Hence in this substitution model, the difference between DC and DU is not in the ordering of their ranked sequences, rather the difference lies in their energy level spacings (which can be seen by changing m which affects the energy spacing formula Equation (2.10)).

This picture of DC functional sequences being a strong version of DU's sequences is consistent with our findings that their median energies differed by almost two units, and with Papatsenko *et al.*'s findings [84] that Dorsal binding sites necessary in limiting concentrations of Dorsal protein (such as in the neuroectoderm) tend to have higher information scores (lower energy scores), than other Dorsal sites such as sites active in the mesoderm [84]. It is also consistent with the mathematical definition of "specialized" sites from Erives *et al.*[35] and the D_β sites of Crocker *et al.* [25] who defined these sites based on how they were detected (similar to MEME's One Occurrence Per Sequence setting (OOPS)[8], the specialized sites were one site per NEE CRM sequence, where each discovered site shared the highest sequence similarity between the selected sites between the CRMs), which in a sense, is the Dorsal site that had the slowest mutation rate (i.e., under the strongest purifying selection).

2.7.7 Conditional detectors

In Figure 2.4 we see that the DC detector can resolve whether a Twist site is in the spacer window or not if the detector fires when $E(S|C) < 3$ (see Equation (2.15)). The resolution is not perfect in this regime: the DC detector still has an error rate, which we define as $1 - 2^{-H(\mathcal{C}|\mathcal{P})}$, where the conditional entropy is defined as:

$$H(\mathcal{C}|\mathcal{P}) = H(C) - I(\mathcal{C};\mathcal{P}). \tag{2.27}$$

The conditional entropy, $H(\mathcal{C}|\mathcal{P})$, is simply the uncertainty of $\mathcal{C}$ given $\mathcal{P}$. But what does this mean for a DC detector? We interpreted this conditional uncertainty as a measure of the detector's uncertainty about the underlying Dorsal binding site sequence given how well it predicted its context. For example, if we assume $H(\mathcal{C}) = 1$ bit while DC's information is $I(\mathcal{C};\mathcal{P}) = 0.3$, then plugging into Equation (2.27) we have

$$H(\mathcal{C}|\mathcal{P}) = 1 - 0.3 = 0.7 \text{ bits}, \tag{2.28}$$

and hence Dorsal has decreased its uncertainty about its context.

If the mutual information $I(\mathcal{C};\mathcal{P})$ was maximal (1 bit), then Dorsal could predict with perfect accuracy whether Twist was proximal or distal. At the opposite extreme where the Dorsal detector does no better than random guessing, we see that it would take about two guesses on average to predict if Twist will be near a binding site sequence. From an evolutionary point of view, the information $I(\mathcal{C};\mathcal{P})$ encoded in Dorsal binding sites can be seen as a message passed from an ancestral population of flies to its descendants. Here, the message instructs Dorsal to interact with Twist, and is encoded in the DNA of Dorsal

binding sites.

2.7.8 Forms Of Conditional Detector's score

We have defined the conditional energy as the CB energy plus a shift depending on the cis-context of the binding site locus. A similar bioinformatic measure is to solely use the conditional probability distribution over sequences to construct a bioinformatic scoring function as: $\log \frac{P(S_{o_C}|C)}{P(S|C)}$, where S_{o_C} is the consensus sequence under condition C. This form of the discriminant function can be transformed to our form $E(S|C) = \log \frac{P(S_O)}{P(S)} - \log \frac{P(S,C)}{P(S)P(C)}$ by a simple additive term to the bioinformatic score $\log \frac{P(S_{o_C}|C)}{P(S_O)}$, where S_o is the concensus sequence of the CB distribution. The advantage of our technique is that mixture of data sets of different sub-types of binding sites that are heavily weighted by the consensus sites of the mixture component distributions can be shifted relative to the consensus site of the lumped data set, such as $\mathcal{D}_{\mathbf{CB}}$. Hence, we are able to resolve quantitatively the amount of energetic shift, while a purely conditional based scoring functions will set each detector's consensus score to zero. Information scores $I(S) = \log \frac{P(S)}{P_O(S)}$, or log likelihood ratios, are also commonly used for bioinformatic detection, which can be transformed to our score by adding an additional conversion term that cancels the logarithm of the background distribution. Information scores have the advantage over purely conditional energy scores in that they allow all bioinformatic scores to be based on a common scale (the background), which is equivalent to an energy score with a common reference point in the case that one uses a uniform distribution as the background. This is because in sequence space[5] one can imagine

[5]In this context each position of a sequence can be thought of as a 4 dimensional real vector space. Then the length k sequence is 4k dimensional real vector space, where each point simply determines the scaling along each dimension (such as the counts that a particular nucleotide base is observed in an alignment at a given position, or the energy score at for that base at a given position).

a point that represents a uniform linear combination of all sequences, then the information scores can be seen as hamming distances away from that point; just as energy scores are basically hamming distances measured from a reference point in sequence space. Our aim was more than just discriminating Dorsal sites from random k-mers, we wanted to resolve differences within Dorsal binding sites. Our score has the advantage of allowing one to resolve the most important region of that scale (shifts away from the CB ground state), which can not be done with the pure information score that uses a uniform background.

2.8 Conclusion

Position Weight Matrices represent a linear coarse-grained physical lattice model of DNA-transcription factor binding. At the DNA sequence level and at the level of Darwinian selection PWMs represent one of simplest possible linear models. In the case that each position within a binding site is independently interacting with the protein binding domain, it makes sense to use a simple model for binding since the affinity (the phenotype) is linear, and hence natural selection may behave as if a linear model. However, binding site sequences may be dependent, and hence linear models will miss important information. By conditioning PWMs based on the variables that are causing the dependency structure within binding sites it is possible to resolve the binding sites into independent classes that can then each be modeled as conditionally independent PWMs.

The necessity of introducing nonlinear sequence models into binding site sequence models is known to help improve binding site sequence detection, and to give a more realistic perspective to binding site models. A number of groups have introduced similar models for discovery of co-occuring motifs [71, 9, 42, 21],[10, 48, 85, 46, 70, 76]. In addition, others have

looked at the influence of symmetries in the flanking sequence of binding sites [93, 3]. Here we placed our analysis in the context of Berg and von Hippel's population genetics model that is related to thermodynamics, and hence the interaction term could be placed inside of thermodynamics occupancy models of transcription factors.

Our conditional PWMs account for epistatic interactions between Dorsal binding sites and their *cis*-context. We showed that Dorsal binding sites contain on average around 0.5 bits of information about the presence of Twist in the flanking sequence of each Dorsal site (see Table 1), thereby contributing to disentangling the dependency structure of Dorsal binding sites active in fly development. In the future, our model can be incorporated in the annotation of binding sites of regulatory regions, and could be used for modeling cooperativity and antagonistic interactions directly from the sequence level. Such models could be used by occupancy models of transcription factors that predict gene expression, such as those in Refs. [51, 36].

2.9 Methods Supplement

2.9.1 Alignment of cis-regulatory modules and collection of $\mathcal{D}_{\mathbf{CB}}$

We used the sequence editor SEAVIEW's default MUSCLE multiple sequence alignment settings [40] to align a given gene's 12 orhthologous CRMs[32]. Given the alignment we then manually extracted the blocks that contained the known *D. mel* binding sites, which allowed for flanking sequence to be extracted (these blocks on average spanned about 15bp with no gaps across the 12 species). The data set of all 12 extracted orthologs (sometimes less than 12 if a binding site was not in the block) for all the *D. mel* Dorsal binding site loci is labelled as $\mathcal{D}_{\mathbf{CB}}$. The $\mathcal{D}_{\mathbf{CB}}$ data set is available at: https://github.com/jacobclifford/MIBBS,

and the raw CRM data (the fasta of each of the 12 orthologous CRMs) is available at: https://github.com/jacobclifford/CRMalignment.

2.9.2 CRM alignment using MUSCLE

MUSCLE (multiple sequence comparison by log-expectation) is an alignment tool for protein sequences, which we co-opted for alignment of cis-regulatory modules. MUSCLE uses a similar algorithm to CLUSTALW, both of which use the 'sum of pairs' (SP) objective function to determine the best multiple sequence alignment (MSA). I will briefly outline the SP score, and then discuss how MUSCLE is different than CLUSTALW.

Fundamentally sequence alignment for strings of amino acid symbols or strings of DNA symbols rewards symbol matches and penalizes mismatches. A scoring system is designed to determine the rewards and penalties based on the goals and purpose of the alignment, where the Altshult (the BLAST cofounder) and Karlin system is commonly adopted due to it having an analytic solution for calculation of the p-value for hypothesis testing (where the null hypothesis is that the sequences are evolutionarily independent (random)).

Regardless of whether one has a sequence of amino acids or DNA the match and mismatch score is based on a 'similarity matrices' such as a Dayhoff matrix (a 20x20 matrix for amino acids) or a Henikoff matrix, BLOSUM, which is essentially a look-up table that determines the score for aligning any two symbols (the table treats all positions of the sequences the same, unlike PWMs). For example, Dayhoff estimated her similarity matrix partially based on recently diverged proteins from a number of known protein families (gene families) that were available at the time of construction (around 1970) with sequences that had a known alignment, then counting the frequency of alignment between any two symbols and hence substitutions between symbols, the joint probability, $p(x, y)$, of seeing symbol x at some

arbitrary time and symbol y at some different time could be determined. Hence the score is simply $s(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$, where $p(x)$ is the probability of symbol x in any protein sequence (similarly for $p(y)$).

For pairwise alignment (two sequences being aligned), given a scoring system, there are two exhaustive methods of alignment (i.e. that look at all possible ways to align the two sequences). These two methods are distinguished based on their 'boundary conditions', and are called global and local. Global alignment means all they symbols of the sequences must be aligned and the score across all the aligned symbols must be accounted for in the pairwise alignment score, where the best pairwise alignment (the optimum) is a function over all possible alignments) . While local alignment does not require all the symbols of the sequences be aligned, and the best subsequence pairwise alignment is considered to be the alignment, where now the pairwise alignment score simply accounts for the length of the subsequences, while again the best scoring aligned sequences is taken as the optimum. Both methods, global and local, assume that no transpositions (or crossing over) occur (e.g. one must align consecutive characters of a sequence to consecutive characters of the other sequence or to a so-called indel character (gap or insertion)). In short, global alignment means all characters between the two sequences must be aligned (with the possibility of gaps and insertions), while local alignment means only a subset of the character between the two sequences must be aligned.

The pairwise alignment score of two sequences x and y that are clones of one another is $d(x,y) = \sum_i^L s(x_i, y_i)$ where i is the internal position of each sequence, and L is the length of the sequence. If we now allow gaps between these sequences there are a total of $\binom{2n}{n}$ possible scores (many of which are degenerate). The gap penalty is usually of the form of a constant (independent of nucleotide or amino acid symbol aigned to the gap). For

example, for a simple scoring system where s is 'one' for a match, and 'zero' for a mismatch (the identity matrix), and we only have a constant gap penalty, we see that there is simply $L+1$ possible pairwise alignment scores from the $\binom{2n}{n}$ possible alignments. In general, the computation of the $d(x,y)$ is accomplished through dynamic programming, where one has a progressively computes the global optimal $d(x,y)$ alignment by using the recursion relation $S(x_i,y_i) = max(S(x_{i-1},y_{i-1}) + s(x_i,y_i), S(x_{i-1},y_i) - g, S(x_i,y_{j-1}) - g)$, where $S(x_i,y_i)$ is the partial sum of scores up to position i for sequences x and y, and g is the gap penalty.

Once one has chosen what type (global or local) of pairwise alignment (i.e. $d(x,y)$ score) then to estimate a MSA for n sequences using a SP objective function one must create a n x n distance matrix (same idea as in 2.9.5), where the elements of the matrix are the pairwise alignment scores between any two sequences ($d(x,y)$). Once one has a distance matrix any number of clustering algorithms can be used to begin to group sequences based on similarity, in particular a minimum spanning tree must be built. Upon construction of the minimum spanning tree a pairwise alignment is arbitrarily chosen and sets the foundation of the MSA (a pair of leaves in the tree), then additional sequences are systematically 'added' to this foundation, where upon each addition the MSA grows until all n sequences have been 'added' to the MSA. The addition of sequences to the initial foundation is not arbitrary, the entire point of computing the minimum spanning tree was to guide precisely when each sequence should be added to the growing MSA (see figure 1 of Edgar[33]). Hence MSA grows (is built) by following the branching of the tree, where at each branch point an alignment occurs. This is called progressive alignment, and is the basic idea behind CLUSTALW. For example, given the minimum spanning tree (suggestively called a guide tree) the first alignment is simply the pairwise alignment already calculated; the rest of the alignments are 'profile-sequence'-alignments - where a profile is an existing alignment (it is like a PWM with

the possibility of gaps as a symbol), or a 'profile-profile'-alignment (where the columns of a profile are now regarded as a symbol, and standard global or local alignment is performed on these symbols, similar to STAMP for PWM-PWM alignments, where the 'match' function between two columns is called the profile function (which is called the log expectation score in MUSCLE).

A novelty of MUSCLE, as suggested by the LE in MUSCLE, is the log expectation score used for profile-profile alignment (see Eq. 1 from Edgar[34]). Furthermore, MUSCLE also uses a k-mer counting scheme to determine the matrix elements of the distance matrix (as opposed to the global or local pairwise alignment score), where it is thought that k-mer co-occurrences correlate well with sequence similarity (see Eq. 1 of Edgar[33]). The k-mer counting scheme is similar in spirit to the BLAST algorithm (local alignment between a short and long sequence), which is popular for its speed and high throughput (time and space computational complexity). It appears MUSCLE looks at all possible 6-mers when constructing the similarity measure (and doesn't count overlaps).

The MUSCLE algorithm coarsly is shown in Figure 1 of Edgar[]. The initial step is identical to CLUSTALW with the exception that MUSCLE uses k-mer counting to define it's distance matrix elements. And MUSCLE uses UPGMA clustering to build the guide tree, while CLUSTAL (CLUSTer ALignment) uses neighbor-joining, for reasons discussed by Edgar[]. The initial guide tree is then used to progressively align the data based on the branch ordering of the tree, where MUSCLE uses the LE score while CLUSTAL uses a LA score that is slightly different score (the difference between LE and LA is the same difference as our using a CB PWM to define P(S) and the true mixture to define P(S), where a true mixture can not be 'additive' when the logarithm is used for scoring).

Upon construction of the MSA based on the guide tree, MUSCLE repeats the entire pro-

cedure using the Kimura distance to build a distance matrix (the Kimura distance requires an alignment, which is possible, since the initial iteration in MUSCLE constructs an alignment). Clustering is again performed on the distance matrix (with matrix elements based on the Kimura distance between any two sequences), and again the guide tree (representation of clustering) is used to progressively build the new MSA.

A third type of 'iteration' occurs in the MUSCLE algorithm (after the newly generated MSA from Kimura distance), where the guide tree is randomly partitioned into two subtrees, where a profile (MSA) is built using each subtree as a guide (the profile is progressively built, and hence an MSA). Then profile-profile alignment is performed (similar to STAMP PWM-PWM alignment). The number of random partitions of the guide tree, and hence profile-profile alignments is a parameter the user can adjust, or will automatically halt once the partitions fail to improve the alignment.

2.9.3 MUSCLE is both fast and accurate

There is almost always a tradeoff between speed and accuracy in computation. According to the online MUSCLE documentation at http://www.drive5.com/muscle/manual/accurate.html: "The default settings in MUSCLE were chosen for best accuracy rather than making any compromises for speed, so if you want the best accuracy you should probably stick with the defaults unless you have a new and better parameter tuning method." This suggests that our use of default parameter settings for CRM MSA would result in best possible alignments from the MUSCLE tool. However, MUSCLE was designed for protein alignments, and even for protein alignments, the use of limited availability of high quality benchmark datasets means the MUSCLE parameters were tuned to fit a small subset of protein familys and their peculiarities, hence the generality of the MUSCLE parameter settings beyond these

few protein familys in the benchmark datasets is questionable.

The speed and accuracy of MUSCLE for a given data set is determined by the number of 'iterations'. Faster alignments are achieved by simply reducing the number of iterations, and more accurate alignments are achieved by simply increasing the number of iterations. The 'iterations' that can be adjusted (as a parameter) are the 'refinement' iterations, which randomly deletes a branch of the guide tree, and then does a profile-profile alignment on the two subtrees (i.e. the two data sets of sequences resulting from the partition (branch deletion)).

2.9.4 MUSCLE parameter sensitivity

A possible parameter that our CRM alignments, and hence our CB data set $\mathcal{D}_{textbfCB}$, are sensitive to are the choice of substitution matrix (especially seeing an amino acid 'substitution' doesn't even make any sense for CRMs, albeit using a substitution matrix for aligning regulatory sequences is unlikely to do harm, seeing that in all cases a substitution matrix rewards exact matches, the problem is it will also reward certain cases of mismatches or at least fail to penalize the mismatch.). However, by default, if a DNA sequence is given to MUSCLE the substitution matrix is not used (presumably, because the reading frame is not known, albeit some MSA tools (unlike MUSCLE) will generate all possible reading frames automatically (there's only three of them)). Furthermore, although MUSCLE leaves the option of using BLOSUM or PAM matrices for constructing distance matrices, as I understand, the default of MUSCLE uses k-mer counting as the distance measure, which I see no *apriori* reason why that can not be used for CRMs. An exhaustive list of parameters (options) for MUSCLE are described in the Command Line Reference section at: http://www.drive5.com/muscle/manual/valueopts.html. The Seaview editor allows these

options (parameters) to be varied by setting them in the 'Alignment Options' tab available from the 'Align' pulldown menu available through Seaview's toolbar at the top of Seaview's GUI window.

MUSCLE uses an 'affine' gap score (penalty) that penalizes for opening, extending, and closing a gap. This is the same scoring technique as CLUSTAW, and like in CLUSTALW, by increasing the gap score by a factor of ten will usually have visible effects by causing less 'blocks' to appear in the finished alignment, where the blocks that do appear are commensurately longer. The default gap open score was 10, and this value I found, in some cases could be possibly be increased by an order of magnitude, as some known binding sites that have had indels will likely have a gap inserted in the final alignment (which creates more work for me, as I prefer to have the binding site block to have no gaps, which allows for faster 'sites' creation where we extract the block of putative sites and the known *mel* site. This is done by exact search in the seaview editor for a known *mel* site in the CRMs. Once found the site and everything it had aligned with/to are extracted and placed in their own fasta file; which is done in seaview by using the 'site' pulldown menu, and selecting 'create set' that allows for highlighting blocks (subsegments of an MSA) of aligned sequence, which can then be automatically saved as new fasta files.).

The 12 CRMs that are associated with each target gene along with shadow enhancers (a distinct set of 12 CRMs associated with some target genes, such as *vndV* (*vnd Ventral*) and *sogS* and *brkS* (*brk Shadow*) - where many of the shadows we did not use, such as *vndV* and *brkS*) are available at: https://github.com/jacobclifford/CRMalignment.

2.9.5 GEMSTAT modifications for locus annotation of CRMs

Given the known sites, $\mathcal{D}_{\mathbf{CB}}$, with their flanking sequence from the block alignments we then created an exact search algorithm that allowed us to estimate the coordinate of the known site within the regulatory region for data structures used by GEMSTAT, a platform developed in the Sinha lab [51]. This program has a various inputs, which are irrelevant for our current purposes, the relevant input is a set of PWMs that represent transcription factors, and the set of CRMs regulated by these factor's motifs (PWMs). We extended the GEMSTAT input to allow for a raw Fasta set of variable length binding sites (namely our Fasta file for the data set $\mathcal{D}_{\mathbf{CB}}$). Each of these sites and its reverse complement was transformed to a probability PWM 'singleton' representation (the probability is equal to one for the observed nucleotide and zero for the other nucleotides). The longest binding site of length n in the Fasta file determined the length of all the singletons. Binding sites in the Fasta file that were of length k, where $k < n$, then had $d = n - k$ columns of zeros 'padding' the last d columns of their singleton PWM.

We then constructed a distance matrix from the singletons, where each singleton is represented by a row and column in the matrix. The matrix elements of the distance matrix were defined as a normalized euclidean dot product between the singletons (where corresponding components of each singleton multiple each other). The normalized dot product between the singleton S^x in row x and singleton S^y in column y is defined as sum over all the element-wise multiplications:

$$S^x \bullet S^y = \frac{\sum_{ij}^{nn} S^x_{ij} S^y_{ij}}{\sum_{ij}^{nn} S^x_{ij} S^x_{ij}}. \tag{2.29}$$

The row singletons are used for normalization, and hence for a given row x, by iterating over all columns we can filter out any identical singletons.

For example, if we have 400 positive strand binding sites in our Fasta file, we will have a 800x800 distance matrix (400 of the rows corresponding to positive stranded sites, and another 400 rows for the negative strands); where the matrix elements are normalized Euclidean dot products between any two singletons. By screening the distance matrix for elements that were equal to 'one' (which stands for a duplicate sequence or possible symmetric sites) we are able to determine which singletons are unique. Asymmetric sites contain an instance for each strand (which means they are not unique), which is accounted for in our 'overlapping' sites processing step.

Collecting all the unique singletons, we then annotate (scan) the CRMs with each unique singleton using exact match (*e.g.*, zero 'energy' singleton PWM threshold). This step allows us to map each binding site in our Fasta file to a unique locus within the CRMs, thereby attaining the coordinates of all our known binding sites within the CRM coordinate system, which is an essential step in the spacer calculation of Equation (2.8).

Each unique singleton has a personal factor identification (a name), which we mutate to the name 'Dorsal' for every singleton. Associating the name 'Dorsal' to the annotated sites within the CRMs is a necessary step in order to compute the spacer between each of the annotated 'Dorsal' sites and any predicted motifs sites within each CRM. The coordinate defines the start or end of a binding site depending on what strand of the CRM was annotated as a positive hit. For asymmetric known sites that match the bottom strand, the coordinate defines the end of the binding site, while matches on the positive (or 'top') strand of the CRM indicate the start of the site. The asymmetric sites are further processed to cull out any sites that overlap at a specific locus, where the process is described below.

2.9.6 Overlapping site processing

In order to create independent loci, we wanted to have only one hit per binding site, so we culled all overlapping sites and overlapping footprints. In the singleton construction of known binding sites our set of unique singletons frequently contains asymmetric binding sequences which means both the top and bottom strand sequence at a particular locus will have an associated singleton (overlapping binding sites). We are able to choose just one representative for a given locus by an algorithm explained in the 'Best Predictions' section below.

2.9.7 Error In estimating the spacer length between known Dorsal loci and Twist sites

The spacer between two annotated binding sites is determined by the coordinates of the start site of each binding site using the CRM coordinate system. However, 'known' Dorsal sites were annotated in the CRMs using Euclidean dot product search algorithm (discussed above in 'GEMSTAT Modifications For Locus Annotation Of CRMs'), which has an uncertainty in start site of a binding site due to the searching singleton PWM being longer than the actual binding site (due to the flanking sequence). We have about a 6 base-pair error in the exact spacer length for our length 9 binding sites of Dorsal (assuming the Twist site or other potential cooperating site annotation uses a correct length PWM). This is due to the actual Dorsal binding site not being 15 bps (which is the typical length of a site used in our exact search algorithm). However, in practice, many of the loci had smaller appended flanking sequence, which would reduce the error in the spacer length calculation. Furthermore, for binding sites that were centered (which they usually are) in the extracted block from MSA,

the spacer length error would be reduced in half. Here we call this spacer length error and not bias, because we simply don't know exactly where the site resides within the blocks extracted from the MSA. Given the coordinates of the known Dorsal binding sites within the *cis*-regulatory module we then defined a PWM for putative cooperating factor with Dorsal (such as a Twist PWM) and set a threshold on this factor's energy to annotate its predicted sites (where we assume this factor's annotation uses a correct length PWM).

2.9.8 PWM Best *predictions* of binding site loci

By scanning or scoring each possible subsequence of length k within a CRM with the PWM one can filter out all the subsequences that do not match the PWM, where a match is defined as having an energy score below a defined threshold. The coordinates of the subsequences that match the PWM relative to the CRM coordinate system can then be used to determine the locus of the predicted binding site.

Scanning CRMs with a PWM frequently results in multiple overlapping binding sites due to symmetry (positive and negative strand being called a hit) and due to re-occurring patterns in motifs (such as repeated bases like AAAA). In order to have non-overlapping binding sites we processed the set of match sites from the CRM scan to construct a list of non-overlapping sites.

We treated each position within a CRM as the start site of a binding site of length k that was scored by the PWM using Eq. (2.2) (or Eq. (2.15) depending on the model being used for predictions). The reverse complement of each potential binding site was also scored by the PWM. Each length k sequence (potential binding site) was stored in a data structure, a k-mer, which contained attributes of the potential binding site like the coordinate and strand (relative to the CRM) and the energy score. The k-mers below the energy threshold were

selected as a hit, and temporarily stored in a hit list. In order to have no overlapping hits we sorted the k-mer list according to energy scores. The coordinate attribute of the k-mer with the minimum energy, the best site, was used to mask out any overlapping hits. This best k-mer site was passed to a storage vector, which would ultimately contain the annotated k-mer binding sites of the CRM. Upon deleting the minimum energy k-mer site along with the masked out k-mers from the hit list, we iterated the above procedure until the hit list was empty, thereby creating a storage vector of non-overlapping predicted k-mer sites that corresponded to maximum scoring binding sites within the CRM.

2.9.9 Expectation Maximization Alignment

In this paper EM alignment means Expectation Maximization alignment of binding sites. We use a one site per sequence setting that resembles the MEME [8] EM one site per sequence algorithm [7]. A Fasta list of sites is passed to the tool, and for each sequence in the list one internal position of the sequence is defined as the starting position of the inferred binding site. Only one binding site is allowed per sequence from the list passed to the tool, however, for any given sequence both strands are scored by the current value of the PWM, where the highest scoring site's position, regardless of strand, is saved in order to make the alignment. The output of the alignment is a PWM. The tool requires setting the length of the desired PWM and the number of iterations of the Expectation Maximization algorithm and the number of iterations of a sampler. Recall, the MEME EM simplest form of the algorithm, scores each internal position with a current definition of a PWM. Then upon scoring all sequences and all internal positions of each sequence within the Fasta list, the Maximum score for each sequence, and hence a corresponding position, is determined for the new starting positions of sequences to be are extracted and used to construct a new PWM,

this new PWM is the Expected PWM, in the sense that the Maximum Likelihood values of the expected counts are just the counts themselves. This new PWM is then reiterated upon all the sequences and all their internal positions, thereby iterating through the EM algorithm. In addition at each step of the EM iteration, the stored position of the start site of each site in each sequence is shifted by one base pair and then the PWM is recalculated to check for phase shifts. The shifts are check for both forward and backward shifts up to shifts of half the length of the site [65]. The EM is wrapped inside of a sampler, which allows for a naive global optimization by random starting positions within each binding site sequence being used as the initial conditions that are past to the EM program. A global variable stores the best PWM, upon each iteration of the sampler, if the EM output PWM has smaller Kullback-Leibler divergence (i.e. information content) then that PWM is thrown out, otherwise global variable of the best PWM is redefined by the current iteration's PWM, and the sampler continues until the specified number of iterations are exhausted. The Kullback-Leibler divergence (i.e. information content) of the distribution (the probability PWM) was estimated from the uniform distribution or from a distribution set by a GC content value. In addition we implemented an option to weight each homologous sequence in the alignment based on the divergence time estimated from Obbard *et al.*[81]. However, we have not fully explored the effects of this weighting scheme, and no results were presented with this option.

2.9.10 CB was designed to be an approximation to a mixture

By choosing 'A-rich' strands for representations of DC and DU, we were able to create mixture of PWMs that was not an artefact of the strands when it came to calculating the marginal in the mutual information, and when it came to constructing the CB PWM. To

determine just how similar the CB PWM is to the mixture distribution, we can use the fact that the entropy of a mixture distribution has the property:

$$H(P(S)^{\text{mix}}) \geq f_{DC}H(P(S)^{DC}) + f_{DU}H(P(S)^{DU}), \tag{2.30}$$

where $H(P(S)^{mix})$ is the entropy of the mixture model of Eq. 2.25, and f_{DC} is the fraction of loci in the population that were assigned to class DC and $P(S)^{DC}$ is the probability of S calculated from the DC probability PWM, and f_{DU} is the fraction of loci in the population assigned to class DU and $P(S)^{DU}$ is the probability of sequence S calculated from the DU PWM. Now if $H(P(S)^{CB})$ is similar in magnitude to $H(P(S)^{mix})$ then it would be reasonable to suggest that $E(S|C) = E(S) + w(S, C)$, where E(S) is estimated from the CB energy PWM.[6]. We found the entropy of the mixture for the spacer window of [0,30]bp was 8.2 bits while the entropy of CB was 8.4 bits (note the entropy of a probability PWM is $2 * k - IC$, where k is the length of the motif and IC is its Information Content calculated using a uniform background distribution over sequences(e.g. see the first term, IC, in the Lagrangian in the main text)[7]. Given that the entropies of these distributions are within a couple decibits and inspection of the logos from figure 2.1 suggests the ranking of sequences by the PWMs is preserved between DC, DU, and CB (and hence by the mixture distribution)- the preservation of sequece, of course, breaks down for the 5'-AAATT and 5'-AAAAT cores of DC and DU,

[6]For a physical mixture, the correct form of the marginalized energy is:

$$E(S) = \log\left(\frac{P_{ref}}{\sum_c \prod_i P(S_i|C=c)P(C=c)}\right), \tag{2.31}$$

where P_{ref} is the probability of the most probable sequence from the mixture model (i.e. in the mixture model joint distribution $P(S)$ takes the form: $\sum_c \prod_i P(S_i|C=c)P(C=c) = P(S)$).

[7]The information content of any distribution p is: $IC = H_{max} - H(p)$. In a sense, it's a measure of how far the distribution p is from the uniform distribution, which, in the context of detectors, tells us about the predictive ability of a detector, since $IC = 0$ is just random detection, no better than guessing.

notwithstanding, it seems reasonable, for the Dorsal OR gate, to simply use CB as a proxy for the marginal mixture model in the calculation of $E(S)$. Without this approximation one would have to use a more complicated data structure in order to calculate $E(S)$, such as a look-up table that stores the probability of all 4^k sequences.

2.9.11 Conditional Distributions

The above count table provides the basic elements to estimate the marginal probability of the bases over the two classes. The maximum likelihood, ML, estimate of the counts from the table are the counts themselves. Furthermore, any function of a ML estimate is itself the ML estimate. The ML estimate of the marginal counts of B over C is $n_B = n_{B1} + n_{B0}$. The ML estimate of the marginal probability of B is $\frac{n_B}{n}$, where n is sum over all elements of the table ($n = \sum_B n_B$). ML estimates of the counts and the probabilities enjoy the property that the estimates are unbiased. However, the ML estimates of the functionals energy and entropy are biased, where the ML estimate of entropy always underestimates the value of the entropy[80]. The bias in the ML estimate of the energy is more complex to analyze, since the energy (as defined in Eq. 2.10) is equal to the entropy plus an additional extreme value variable (where this extreme value variable is from a Gumbell like distribution). The bias in these estimators could affect our bioinformatic searches based on a cutoff of the energy, and could affect our calculations of information content and mutual information. Hence we chose a Bayesian approach that uses a hyperparameter β to correct for the small sample bias in entropy and energy. This approach leads to an estimate of the discrete marginal probability of B over C with a Dirichlet prior with a symmetric hyperparameter β, defined as $P(B) = \frac{n_B+\beta}{n+4*\beta}$. We used the same β for all positions of a PWM[69]. Similarly, the

conditional distribution of B given C is defined as $P(B|C) = \frac{{}^{n}BC^{+\beta}}{\sum_{Bn_{BC}+4*\beta}}$, where we use the same β for all positions of the conditional probability PWM for our estimates of the conditional distribution of B given C.

To estimate the uncertainty in our count estimates, a frequentist may assume a Poisson counting-like process, which has a well-known property that the expected counts for a set number of trials is equal to the variance of the distribution of counts, which is supported up to the set number of trials. One can then estimate the confidence interval of their estimates of the expected counts and hence the standard error on $P(B)$. However, from a Bayesian perspective, the normalized counts are simply samples from a probability simplex (the distribution of distributions)[17]. Here one doesn't estimate standard errors on $P(B)$, rather the variance of the distribution over the probability simplex is a measure of the expected spread of $P(B)$ (*i.e.* how much do we expect P(B) to vary from one alignment (sample) to another, in other words, how reliable is our estimate of $P(B)$). Thinking of each B as a category, then we can use the Dirichlet as our prior distribution over the categorical distribution $P(B)$ (choosing the Dirichlet as the prior preserves the form of the categorical distribution when new information becomes available that we use to update our estimate of $P(B)$). The Dirichlet has an elegant formula for its variance, which we reproduce here for convenience:

$$\sigma^2_{P(B)} =< P(B)^2 > - < P(B) >^2 = \frac{\alpha_B(\alpha_0 - \alpha_B)}{\alpha_0^2(\alpha_0 + 1)}, \tag{2.32}$$

where α_B are the concentrations (hyperparameters) for $B = A, C, G, T$, and $\alpha_0 = \sum_B \alpha_B$. After we observe the sample (the alignment), the variance changes because we've gained new information. We can (as a consequence of 'conjugacy') simply recycle the formula above with

a change of variables $\alpha'_B = \alpha_B + n_B \ldots \ \alpha'_o = \alpha_o + n$, this leads to the posterior variance:

$$\sigma^2_{P(B)post} = \frac{(\alpha_B + n_B)(\alpha_0 + n - \alpha_B - n_B)}{(\alpha_o + n)^2(\alpha_o + n + n_B + 1)}. \tag{2.33}$$

We chose to use a symmetric hyperparameter, β, where $\alpha_B = \beta \, \forall \, B$, which can be thought of as a 'pseudocount'. Berg and von Hippel used the same analysis with the standard maximum entropy prior, which they detail in their appendix[13].

2.9.12 Detector energy thresholds, E_c

From a bioinformatic perspective, a detector's conditional PWMs or the CB PWM must test each potential 9-mer in a CRM by making a prediction as to whether the 9-mer is a binding site or random background DNA. Hence the prediction is a binary classification that labels each 9-mer as positive or negative. The positive sites indicate the 9-mer's energy is below an energy threshold, E_c (critical energy), while negative sites have 9-mer sequences with energy above the energy threshold.

We define the bioinformatic specificity, ν, as the cardinality of the number n of sequences of length 9 bp that are considered a positive binding site due to their energy being below the critical energy, divided by the cardinality of the total number N of possible sequences of length 9 bp, where $N = 4^9$. Hence the bioinformatic specificity is $\nu = \frac{n}{N} = \sum_{S \in \mathcal{S}} P(S)\theta(E(S) - E_c)$, where $\mathcal{S}$ is the set of all possible sequences (i.e. the set of cardinality N), and θ is the step function, which acts as an indicator variable that has a value of 'one' when E(S) is below the threshold energy of E_c and theta has a value of 'zero' otherwise. Once a bioinformatic specificity is set, we use the estimated cumulative distribution function, cdf, of the CB energy over the 4^9 sequences to calculate the energy

threshold that matches a particular value ν of the cdf, (where we assume these 4^9 sequences occur based on the background probability).

We naively build the cdf by nested iterations, which allows us to iterate over all possible 9-mers, N. At each iteration we determine the unique 9-mer sequence S's CB energy $E(S)$ (position independent model), and increment the bin of the energy histogram that corresponds to $E(S)$, where the bin widths were 0.1 in arbitrary units. To map an energy E to a bin, we map each energy to a bin number (bin identification), where the bin number is $\lceil 10 * E(S) \rceil$ for a 0.1 precision bin width, or simply $\lceil E(S) \rceil$ for a bin width of 1. For example, for $\nu = 10^{-6}$ we expect $n = \nu N$ possible sequences to be below the energy cutoff, we then can rank each sequence in the set of N unique 9-mers based on their energy score, where the nth sequence's energy is E_c.

For a given energy PWM each 9-mer, S, in a crm is scored as $E(S) + w(S, C)$, where the shift is determined by the PWM that was trained from class specific data ($w = 0$ for CB). For example, if the spacer window is set at $[0, 30]bp$ then the DC detector always expects that there is a cooperating site proximal to it, and hence adds w(S,proximal) to the energy E(S) for a given sequence S. All 9-mers that satisfy the constraint $E(S)+w(S, proximal) < E_c$ are considered a positive hit (where overlapping 9-mers that satisfy this constraint are screened so that the best scoring 9-mer is considered the positive site).

2.10 Results Supplement

2.10.1 Description of rank sum sampling distribution construction

It's possible that any random data set of binding sites that are used to build detectors using our technique would produce p-value's similar to those found between DC and DU detectors.

Hence, our original data set of Dorsal sites $\mathcal{D}_{\mathbf{CB}}$ was randomly partitioned such that half of the sites from it are placed in D_1 and the other half of sequences into D_2. Then we build a detector (a model) with PWMs trained from the D_1 sequences, and similarly we build another detector from the D_2 data. Then using our formula for conditional energy we compute the conditional energy for each of the D_1 sequences. For example, let $D_1 = S_1, S_2, ..S_n$, where the cardinality of $\mathcal{D}_{\mathbf{CB}}$ is 2n. Then we compute the energy of these sequences: $D_{1E} = [E(S_1|1), E(S_2|1), ..E(S_n|1)]$. Similarly the data set D_{2E} will be based on the conditional energies for the corresponding sequences in D_2. Once we have the lists of conditional energies, we compute the median energies between D_{1E} and D_{2E}, and use the ranksum test to obtain a p-value.

We repeat this procedure 1000 times and bin the pvalue. Hence we create a distribution of pvalues for the ranksum test, which can be used to test if our detector's DC and DU have a significant ranksum pvalue against the background of pvalues from ranksums of random partitions of the data.

2.10.2 Logodds ratio test of DC and DU positive hits

For a given detector, defined by the class value C=c, each 9-mer, S, in a crm is scored as $E(S)+w(S, C = c)$. For example, the DC detector always expects that there is a cooperating site nearby, and hence adds *w(S,c=1)* to the energy *E(S)* for any given sequence *S*. All 9-mers that satisfy the constraint $E(S)+w(S, c = 1) < E_c$ are considered a positive hit (where overlapping 9-mers that satisfy this constraint are screened so that the best scoring 9-mer is considered the positive site).

All positive hits for a detector are then classified using the same spacer window scheme that was used in constructing the detector itself (*i.e.*[0,30]bp). All positive hits that contain

a Twist site in the spacing window are classified with class tag 'proximal', and the Dorsal sites without a Twist site in the spacing window are classified as 'distal'.

We constructed a 2x2 contingency table with table elements $n_{M,C}$ that represent the number of predicted Dorsal loci from detector M that match the properties of class C (e.g. C=p indicates the predicted Dorsal locus had a cooccurring Twist site in the spacer window 'proximal'). The detector M can be considered a random variable with outcome m=DC and m=DU, and the class, C, another random variable. The table elements for the DC detector are $n_{DC,p}, n_{DC,d}$, for the number of predicted sites n from the DC detector that were proximal p to Twist's motif, and similarly the number distal d from Twist. We calculated the logodds ratio of the proximal sites of the DC predictions verse the proximal sites of the DU predictions.

	proximal	distal
DC	$n_{DC,p}$	$n_{DC,d}$
DU	$n_{DU,p}$	$n_{DU,d}$

Table 2.3: Contingency table of DC and DU detector versus the class type distal and proximal. Elements of the table are the counts from predictions of each detector for a given energy cutoff and spacer cutoff in a given set of CRMs.

The logodds ratio is the ratio of two odds; odds of DC proximal, labelled as ODCp= $\frac{P(np|DC)}{P(n_d|DC}$ and odds of DU proximal labelled as ODUp = $\frac{P(np|DU)}{P(n_d|DU)}$, where the conditional probabilities (such as $P(n_p|DC)$) are estimated from the 2x2 contingency table in Table 2.3 with a pseudocount of value one. The logodds ratio is : $\ln(\frac{ODCp}{ODUp})$. The sampling distribution of the logodds ratio is a normal distribution with mean zero and width equal to the standard error of the mean. We are interested in the one-sided test, hence, our p-value estimate is the integral of the sampling distribution from the given logodds ratio to positive infinity (the chances of seeing the value found from the test or a larger value).

2.10.3 Mutual Information between known class tags and the conditional detector's predictions of class tags

We use the mutual information between the binary class variable $\mathcal{C}$ ('proximal' and 'distal', defined by the spacer window) from our known binding sites and the detector's binary prediction $\mathcal{P}$ of the class to determine the detector's performance at resolving class types of 'proximal' or 'distal'. Here a prediction still means that the detector is testing whether a k-mer is a Dorsal binding site, however, we are additionally checking to see if the binding site locus of the k-mer being tested has the correct flanking sequence feature.

We use the information identity to transform mutual information into an entropic form:

$$I(\mathcal{C};\mathcal{P}) = H(\mathcal{P}) - H(\mathcal{P}|\mathcal{C}). \tag{2.34}$$

However, this quantity can only be calculated given a model M (a conditional detector and it's corresponding energy threshold), hence, in our own notation we will write this mutual information as $I(\mathcal{C};\mathcal{P}, M = m) = H(\mathcal{P}, M = m) - H(\mathcal{P}|\mathcal{C}, M = m)$, where we make explicit that we know the detector M.

The variables $\mathcal{C}, \mathcal{P}, M$ are all bernoulli-like. The value $\mathcal{C} = 0$ indicates class type 'distal' for a given binding site, and $\mathcal{C} = 1$ indicates class type 'proximal' for a given binding site. The value $\mathcal{P} = 0$ indicates the detector predicted the class of a given binding site as 'distal', and the value $\mathcal{P} = 1$ indicates the detector predicted the class of a given binding site as 'proximal'. The variable M's domain is M=DC and M=DU.

The entropy $H(\mathcal{P}, M = m)$ is the entropy of the predicted class distribution, where we estimate the predicted class distribution based on marginalizing the predictions over the

classes. For example, the outcome $\mathcal{P} = 1$ is computed as:

$$\begin{aligned} P(\mathcal{P} = 1, M = m) &= P(\mathcal{P} = 1|\mathcal{C} = 1, M = m)P(\mathcal{C} = 1) \\ &+ P(\mathcal{P} = 1|\mathcal{C} = 0, M = m)P(\mathcal{C} = 0) \ , \end{aligned} \tag{2.35}$$

where $P(\mathcal{C} = 1, M = m) = P(\mathcal{C} = 1)$, and $P(\mathcal{C} = 0, M = m) = P(\mathcal{C} = 0)$, and we estimate the conditional probabilities for a given detector based on the detector's energy cutoff, since its predictions are based on the energy threshold.

The conditional entropy is defined as:

$$H(\mathcal{P}|\mathcal{C}, M = m) = \sum_p P(p) \sum_c P(p|c, M = m) \log P(p|c, M = m), \tag{2.36}$$

where the conditional probability $P(p|c, M = m)$ is a function of the TPR and FPR that is determined by the conditional detector, for example if $M = DC$ we have:

$$\begin{aligned} P(1|1) &= P(\mathcal{P} = 1|\mathcal{C} = 1, M = DC) &&= TPR \\ P(1|0) &= P(\mathcal{P} = 1|\mathcal{C} = 0, M = DC) &&= FPR \\ P(0|1) &= P(\mathcal{P} = 0|\mathcal{C} = 1, M = DC) &&= 1 - TPR \\ P(0|0) &= P(\mathcal{P} = 0|\mathcal{C} = 0, M = DC) &&= 1 - FPR \ . \end{aligned} \tag{2.37}$$

The TPR and FPR are a function of the conditional detector. For the DC detector we defined the positives as the known 'proximal' sites, previously denoted as D_{DC}, and the negatives are the known 'distal' sites, D_{DU}.

Two equations in Equation (2.37) are based on the FPR, and defined as: $P(p = 0|c =$

$0, DC)$ is the fraction of known 'distal' binding sites whose energy is above the energy threshold (true negatives divided by negatives), and $P(p = 1|c = 0, DC)$ is the fraction of known 'distal' binding sites whose energy is below the energy threshold (false positives divided by negatives). Two equations are based on the TPR, $P(p = 1|c = 1, DC)$ is the fraction of known 'proximal' binding sites whose energy is below the energy threshold, and of course $P(p = 0|c = 1, DC) = 1 - P(p = 1|c = 1, DC)$, which is the fraction of known 'proximal' binding sites whose energy is above the energy threshold (false negative divided by positives).

For the DU detector, the DU positives are what DC call a negative, hence DU's positive data are the known 'distal' sites, and its negatives are the known 'proximal' sites. This is because DU should be firing when Twist is 'distal'. For example, the TPR now determines $P(\mathcal{P} = 0|\mathcal{C} = 0, DU)$, which is defined as the fraction of known 'distal' binding sites whose DU's energy score is below the energy threshold (true positives divided by positives), and $P(\mathcal{P} = 0|\mathcal{C} = 1, DU)$ is simply $1 - P(\mathcal{P} = 0|\mathcal{C} = 0, DU)$. The FPR now determines $P(p = 0|c = 1, DC)$, which is the fraction of known 'proximal' binding sites whose energy is below the energy threshold (false positive divided by negatives), and $P(p = 1|c = 0, DC) = 1 - P(p = 0|c = 1, DC)$.

2.11 Additional Experiment Supplement, Rerunning Model on CACATG Twist Motif

We repeated our analysis of the known Dorsal sites with a different Twist motif (5'-CACATG). We used the same sliding window as in the main text of 30 base pair shifts starting at [0,30]bp, and then incrementing to [31,60]bp etc... We also used the same spacer window as in the

main text ([0,30]bp) for the DC detector for making predictions of known Dorsal loci in the CRMs.

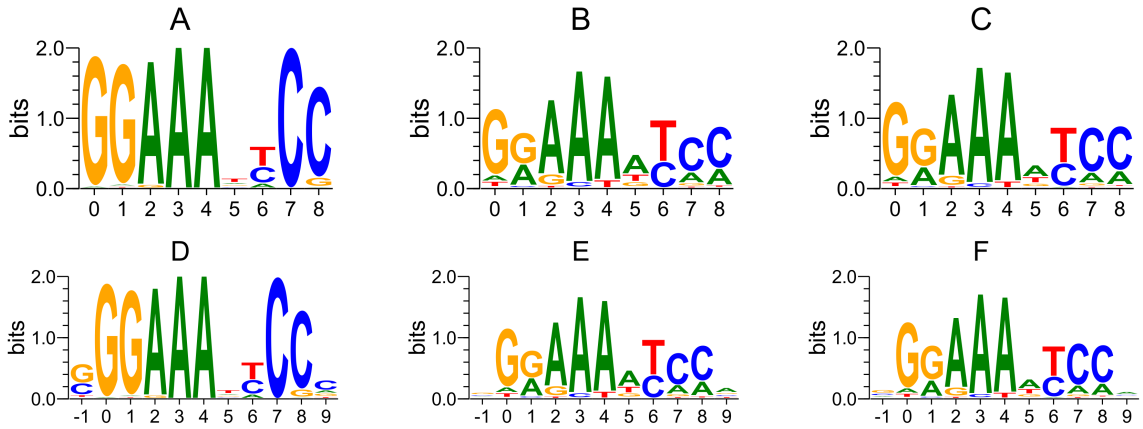

Figure 2.5: Logos generated for known Dorsal sites tested for adjacency to 5'-CACATG used as 'cooperative' class (DC) if in the [0,30]bp distance. Logo A corresponds to the 'Dorsal Cooperative' class, it's total information content we calculated at 13.4bits. Logo D is the exact same logo as A but we've appended one base-pair of flanking sequence onto the start and end of the site (hence, this logo starts at position -1). Position 9 of this logo shows about a couple decibits of information relative to the background sequence and the position -1 contains a half bit of information. Logo B is the 'Dorsal Uncooperative' class for the [0,30]bp window, which we calculated to have 9.4 bits information relative to the background (uniform distribution of bases), and logo E has added the flanking sites to the 'Dorsal Uncooperative' class. Logo C is the CB motif with 9.7 bits of information relative to the background, which looks similar to the 'Dorsal Uncooperative' class at position 6 due to there being many more sites that prefer A to a T at this position amongst all the Dorsal sites in the network. Logo F is the CB motif with the flanking sequence appended.

spacer	[0,30]bp	(31,60]bp	(61,90]bp
Mutual Information Equation (2.17)	0.38	0.28	0.04
Logodds ratio test -log(p value)	4.8	0.17	0.66

The table 2.11 second row corresponds to the log odds ratio test based on CB's specificity set at 10^{-4}, which corresponds to a CB energy of 2.5, as in the main text. The DC detector did show better performance for this spacer window and corresponding energy cutoff. We additionally ananlyzed the mutual information for the case that 5'-CACATGT was used as a motif for Twist, which would correspond to a subset of the sites found for 5'-CACATG.

For the 5'-CACATGT motif we found the mutual information was 0.3, 0.17, 0.0 for the three possible cases of the sliding window.

2.11.1 ROC curve

The ROC curve for the OR gate and the CB detector for 5'-CACATG Twist motif are displayed in Figure 2.11.1. The detectors behave similar to the results presented in the main paper for the 5'-CAYATG motif.

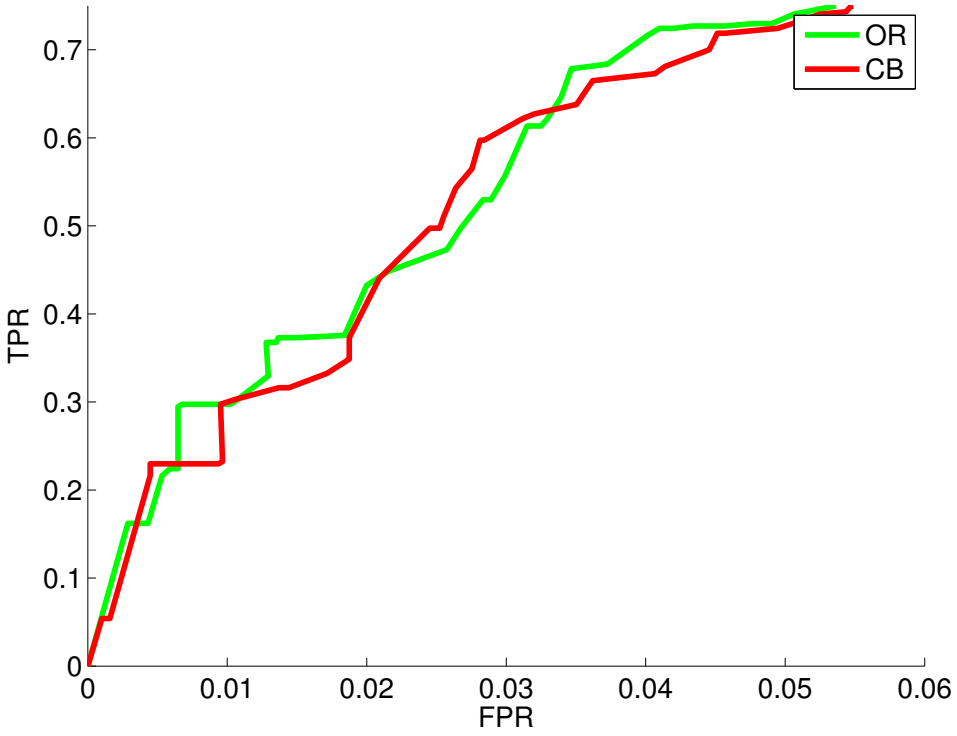

Figure 2.6: ROC curves display the False positive rate (FPR) vs. True Positive Rate (TPR).

2.11.2 Mutual Information between loci classes C and detector predictions of classes P

The mutual information $I(C;\mathcal{P})$ for both conditional detectors based on a 5'-CACATG Twist motifs in Figure 2.11.2 shows similar behavior as the results in the main paper for the 5'-CAYATG Twist motif.

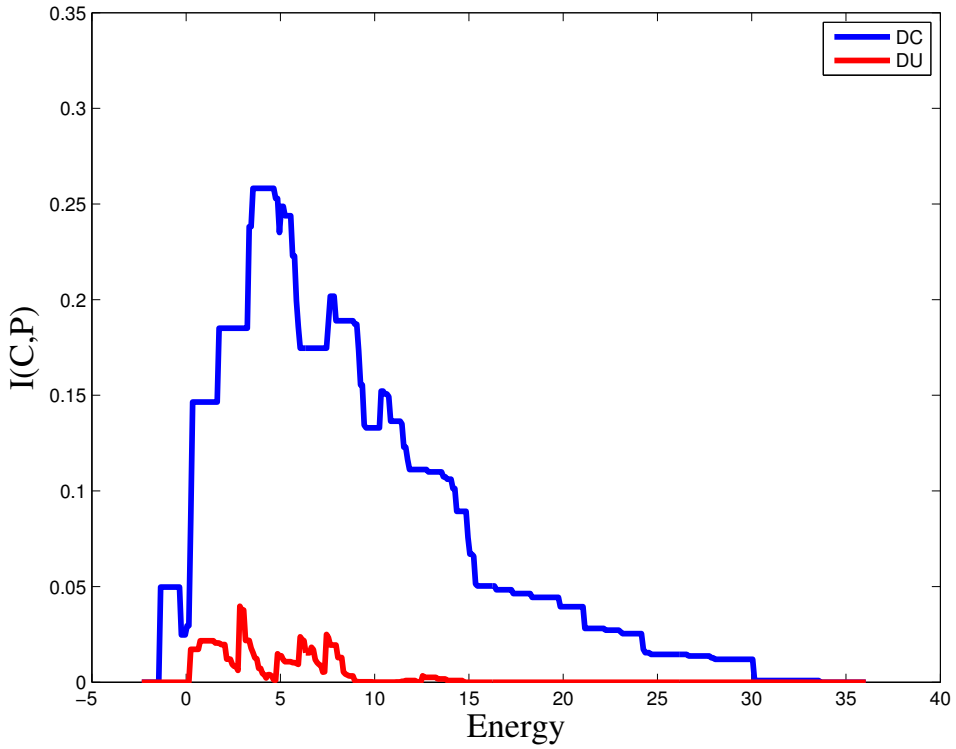

Figure 2.7: The mutual information $I(C;\mathcal{P})$ between the conditional detector's prediction's of class types (distal or proximal) and the known class types, as a function of the detection energy threshold is varied. DC shows about .3 bits of information at about an energy cutoff of 4. DU does performance suggests not much better than random guessing for its predicting class types.

2.11.3 Permutation test using ranksum statistic

The median energy of DC was 0.4, and the median energy of DU was 2.9. The plot of the ranksum sampling distribution generated from random partitions of these data sets of size 66 and 356 respectively is below.

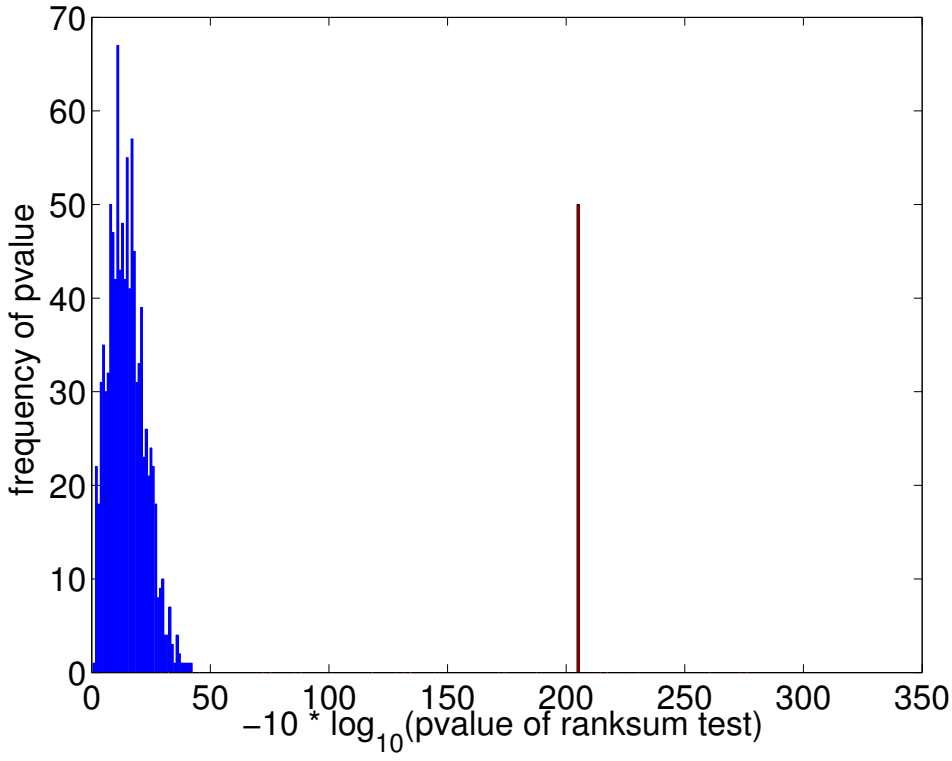

Figure 2.8: Histogram of p-values of random partitions of the combined data set $\mathcal{D}_{\mathbf{CB}}$, where the histogram bins were in units of 10*log(p value). The p-value of the ranksum test for DC and DU median energies was about 205 in the scaled log units, which is the bar at the far left of the sampling distribution.

Propagation of error for estimating error in Entropy estimates

Initially we conflated our notation of the frequency of events i from an alignment with the theoretical probability of the event i. Here we must distinguish these two ideas. Let the frequency of the event i (for example the occurrence of the symbol or nucleotide 'A') be f_i, and the probability be p_i (previously denoted as $P(B_i)$, which we will abandon since it is too cumbersome here). What is the standard deviation of an entropy estimate? One way to estimate this is resampling (bootstrap). Another, much simpler technique is to treat the entropy as a function of a variable, and then simply see how much a variation of the variable (of some specified size) causes variation in the function. Here the idea is to treat the input variation as "error" and see how much this affects the output. Here "error" means uncertainty which is reliably estimated as the statistical standard deviation of the variable (if this is known), or can be estimated by any justifiable means (e.g. the error in a measuring device, is reasonably estimated by the resolution of the device, which is the distance between tick marks on a ruler).

Let the true entropy be $H(p)$, which we estimate by $H(f) = -\sum_i f_i \log(f_i) = \sum_i h(f_i)$, where $h(f_i) = -f_i \log f_i$. We wish to estimate the error in this estimate (the standard deviation of $H(f)$), which can be estimated using the formula:

$$\sigma_{H(f)} = \sum_i \frac{\partial h(f_i)}{\partial f_i} \sigma_{f_i}, \tag{2.38}$$

This formula assumes that our estimate of uncertainty in f_i is ∂f_i, which is equivalent to the standard deviation of f_i, where, for example[8], the expected standard deviation of f_i for

[8]Recall for Poisson processes the variance is equal to the mean for the counts. Hence, $< (n_i - Np_i) >=< n_i >= Np_i$. Also recall that the variance of a random variable scaled by a constant, for example $1/N$, is $var(X/N) = var(X)/N^2$. Therefore the variance in the frequencies f_i is $\sigma^2_{f_i} = \frac{p_i}{N}$

a Poisson process is $\sqrt{p_i/N}$.

$$\frac{\partial h(f_i)}{\partial f_i} = -(\log f_i + 1), \tag{2.39}$$

where we know the derivative of $log(x)$ with respect to x is $-1/x$, and we used the product rule of differentiation on the term $f_i \log f_i$. Using the above computation we can now see how a variation in magnitude equal to a standard deviation of the frequency of event i will contribute to the error (the standard deviation) in the entropy[9].

Now the propagated error to the entropy is:

$$\sigma_{H(f)} = |\sum_i \frac{\partial h(f_i)}{\partial f_i}|\sigma_{f_i}, \tag{2.40}$$

$$\sigma_{H(f)} = |\sum_i -(\log f_i + 1)|\sigma_{f_i}, \tag{2.41}$$

Now we simply must declare the type of measurement process on f_i. If we assume a poisson process, then plugging in to the above formula for the standard deviation σ_{f_i} we have[10]:

$$\sigma_{H(f)} = |\sum_i -(\log f_i + 1)|\sqrt{\frac{p_i}{N}}, \tag{2.42}$$

[9] What is the point of doing propagation of errors? Well, it's reasonable to assume a repeated counting experiment will arrive at a different value of f for each experiment; hence, within the statistical estimate of the standard deviation, which is $f(f-f)$ for a Bernoulli process, we would like to know how much this variation will cause variation in the entropy.

[10] Any information theoretic quantity can be converted to an entropy using standard identities. Hence, for example, the information content of a binding site logo is $IC = H_{max} - H(P(S))$, where P(S) is the PWM estimate of the binding site sequence distribution. The error in this estimate is approximately just the additive error in the two entropy terms. The maximum entropy is known with certainty and hence has error zero. Hence, the information content error of a PWM is simply the error estimate of the binding site sequence distribution's entropy.

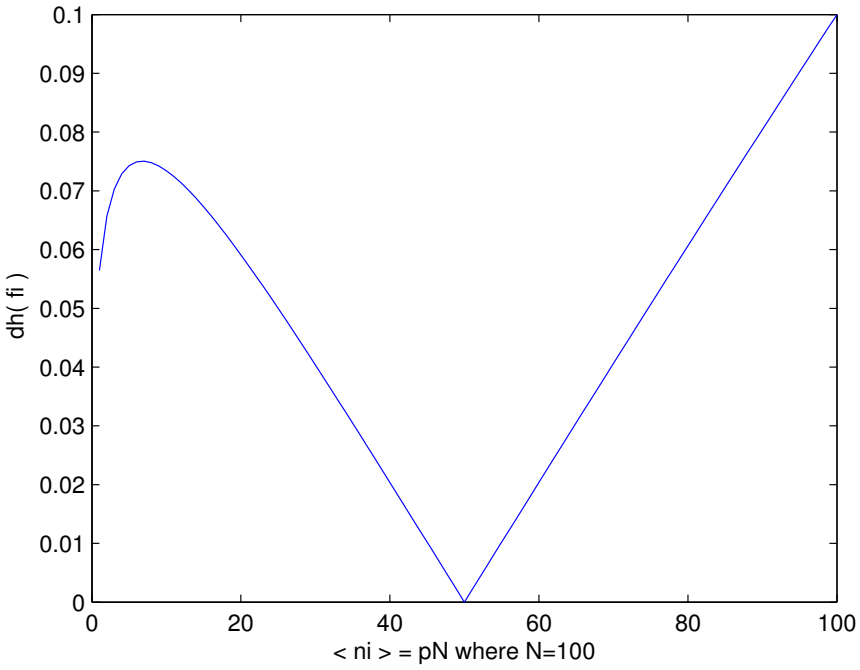

Figure 2.9: The estimate of the error, $dh(f_i) = \sigma_{h(f_i)}$ is plotted on the vertical axis, when the number of counts of event i occur at their expected value, where the horizontal axis is the expected number of counts observed for event i, $< n_i >= pN$.

Determining if differences in Information are significant

From figure 2.11.3 we see that we need, in the worst case scenario, about 4 decibits (about 4*.1 bits, which was the maximum value in the graph) of information content to biologically distinguish two alignments each of 100 binding sites of DNA (i.e. in order to distinguish two columns of a PWM). However, the standard error of the entropy requires dividing the estimated variance by N, in which case, two alignments are statistically resolvable ('significantly different') if they are different by .4 centibits, which may or may not be biologically significant. What is biologically significant? For example, can natural selection resolve binding sites that have differences in information scores of .4 centibits? Depends on the selection pressure, and how big it is. Certainly selection can resolve .4 centibits, given drift doesn't cause the variation in the population to disappear. What is physically different? For example, can Dorsal protein resolve binding sites that are different by .4 centibits? Well, given that Dorsal can sample binding sites, then given $2^{H(P(S))}$ binding events, we would expect a Dorsal to correctly identify a binding site relative to the background of the genome (where $P(S)$) is the PWM probability). So in a statistical sense, yes it seems to be physically significant, since .4 centibits can reduce the amount of sampling required. In fact, any information that Dorsal has about its binding site sequences is physically significant due to this idea[11]

Entropy Bias from Maximum likelihood estimation

The estimate of the error (standard deviation[12]) in the entropy in the above section was based on estimates of the frequencies f. The estimates of the frequencies were based on Maximum Likelihood Principle, which is unbiased for frequency estimates.

[11] This doesn't mean all binding site sequences with information scores greater than zero are functional (adaptations) in Eukaryotes, as Schneider found for prokaryotes[90].

[12] this is not a 'standard error'

Functions of known unbiased estimates (such as the frequencies) are not necessarily unbiased (such as the entropy). It is known that the entropy estimate is biased if one uses ML estimates for their frequencies. To see this we can Taylor expand the above estimator $H(f)$ about the expected value of f, which is p (there's no bias in the expected value of f). Using the same notation as Bialek (see page 547 [16], we will first add and subtract the expected frequencies from each factor:

$$H(f) = -\sum_{i=1}^{4}(p_i + \delta f_i)\log(p_i + \delta f_i) \tag{2.43}$$

where $\delta f_i \approx \sigma_{f_i}$ in my notation above, and since we're expanding about the 'expected value' we have $\delta f_i = (f_i - p_i)$ (plugging that into the equation you see we've just written the definition of the entropy). Now Taylor expanding (where recall, the expansion of $log(1+x) = -x$) we have:

$$\begin{aligned} H(f) \quad &= -\textstyle\sum_i p_i \log p_i - \sum_i (\log p_i + \frac{1}{ln(2)}\delta f_i)) && (2.44)\\ &\quad -\tfrac{1}{2}\textstyle\sum_{i=1}^{4} \frac{1}{ln(2)p_i}(\partial f_i)^2 + \epsilon && (2.45)\\ & && (2.46) \end{aligned}$$

equation here the last term in the expansion is the error term (capturing higher order terms), and expression $(\partial f_i)^2$ acts as the variance $(\partial f_i)^2 = \sigma_{f_i}^2$ in my notation in the section of entropy propagation of errors. The variance term is the average squared fluctuation, the standard deviation squared, which is a positive number (even if any given fluctuation is negative). Now, we can take the expectation value of the Taylor expansion, and of course, the fluctuations are just as likely positive as negative, and they'll cancel (or just recall that

$f_i = p_i$). The first term in the expansion above is a constant (it's the true entropy). Hence we find something a bit odd, the expected entropy is not the true entropy! This is a bias. The expected entropy, to second order is:

$$H(f) \quad = -\sum_i p_i \log p_i - \tfrac{1}{2}\sum_{i=1}^{4} \tfrac{1}{ln(2)p_i}(\partial f_i)^2 + ... \tag{2.47}$$

$$\tag{2.48}$$

Maximum likelihood estimates of the entropy underestimate the true entropy. Why? Well mathematically, the second term on the right side of the equal sign is assumed to be dominant over other higher order terms (where we're assuming this expansion converges). The second term contains the variance (a positive quantity). The true entropy (first term on the right side) is always positive or zero, hence subtracting the *smaller* variance term from the true entropy leads to an underestimate of the true entropy (always).

To estimate the bias in the entropy we must compute $< H(f) > -H(p)$, which as seen above can be estimated by $\frac{1}{2}\sum_{i=1}^{4} \frac{1}{ln(2)p_i}(\partial f_i)^2 = \frac{1}{2}\sum_{i=1}^{4} \frac{1}{ln(2)p_i}\sigma_{f_i}^2$, where we have substituted in the standard variance symbol for the squared fluctuation. Now we know the variance of a poisson process, hence plugging into the term in the last sentence for the variance we have $\frac{1}{2}\sum_{i=1}^{4} \frac{1}{ln(2)p_i}\sqrt{pi/N} = \frac{1}{2}\sum_{i=1}^{4} \frac{1}{ln(2)}\sqrt{1/N}$. Hence, we see for DNA nucleotide counts, the entropy at a given site is biased by $2\frac{ln(2)}{ln(2)N}$. For small sample size (small N) this can have a large effect, even larger than the estimate of the variance itself $\sigma_{H_f}^2$.

2.11.4 Entropy Bias

Changing notation, we now will simply conflate the estimates of frequencies f with the true probability P. The Bayesian estimate of the probability of B with a Dirichlet prior using

symmetric hyperparameter β is $P(B) = \frac{{}^{n}B^{+\beta}}{N+4*\beta}$[69]. Unlike the ML estimate of entropy (which underestimated the entropy) this particular choice of hyperparameter is known to be an overestimate of the true entropy. Hence, in Beyesian estimation, it is possible for the small sample bias to be either over or under the true value. For example, the Baysesian estimate approaches the ML estimate as we let β approach zero, which can be seen in figure **??**. This is expected since if β is zero we then have: $P(B) = \frac{{}^{n}B^{+\beta}}{N+4*\beta} = \frac{{}^{n}B}{N}$.

To minimize the bias in the entropy we selected the value of β that gave the smallest bias for the small sample regime. Of course, we do not know the true entropy a-priori of the binding site distribution. However, for the length 9 bp CB distribution of Dorsal binding site sequences we could initially estimate a 'known' PWM that could then be used to randomly generate synthetic binding sites. Hence, using our best estimate of the CB PWM, we used the PWM to generate data sets of N binding sites. For each synthetic data we estimated a PWM with a predefined value of β, and thereby estimated the entropy as a function of N. For each value of N, we generated n replicate data sets, building a PWM for each set, thereby obtaining n estimates of the entropy for each value of N. We estimated the entropy based on a data sets of size N=[1,50], and $n = 20$. By repeating this experiment for values of β in the domain $[10^{-5}, 0.1, 0.2, 0.25, 0.5, 1]$ we found an empirical value of β that best estimated the 'known' entropy and energy. We similarly repeated this for energy estimates, and found the least biased vale of $\beta = 0.1$ for entropy and energy estimates.

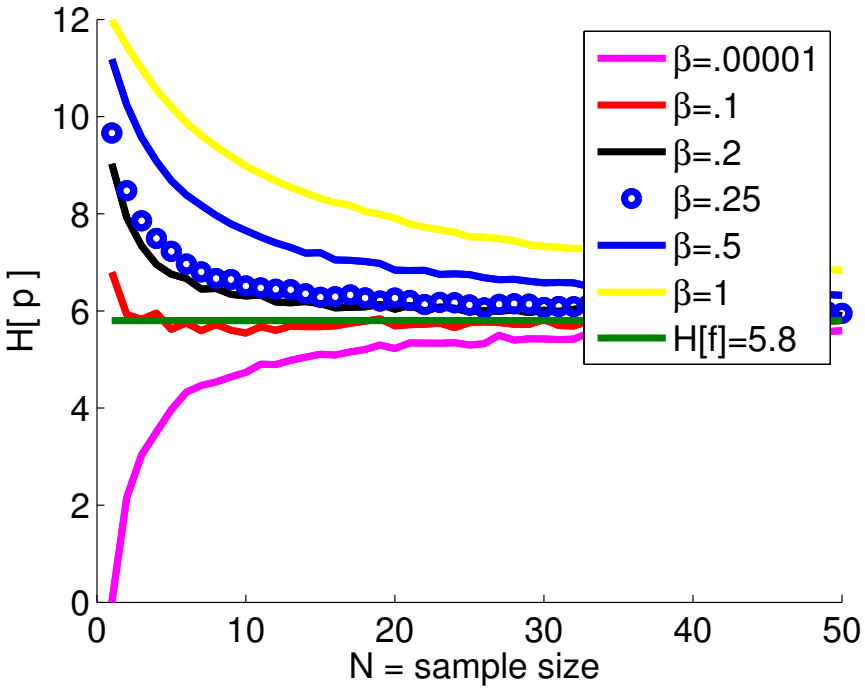

Figure 2.10: The probability distribution p was estimated from N random deviates of a 'known' length 9 PWM that was built from $D_{\mathbf{CB}}$ data. The entropy of p, $H[p] = -\prod_{i=1}^{9} \sum_B p(i,B) \log p(i,B)$, where i runs over the nine positions of the aligned N sequence deviates, and B runs over the bases, was computed for twenty replicates for each value of N and plotted the average entropy over the twenty replicates as a function of N. We computed the functional H as a function of N for values of β in the domain $[10^{-5}, 0.1, 0.2, 0.25, 0.5, 1]$. The 'known' CB PWM had an entropy of 5.6 bits as observed by the green horizontal line, and found an empirical value of β that best estimated this 'known' entropy to be $\beta = .1$ as shown by the red plot of the functional H as a function of N. We similarly repeated this for the functional energy estimates, and found the least biased vale of $\beta = 0.1$ for entropy and energy estimates.

Chapter 3

3.1 Model Background

3.1.1 Fractional Occupancy of Morphogen Binding to DNA binding Site

The binding process is modeled using a first order rate law, where M is morphogen and B is the Binding site of DNA, and MB is the complex:

$$\frac{d[MB)]}{dt} = k_{on}[M][B] - k_{off}[MB], \tag{3.1}$$

where the brackets [ ] denote concentrations, and k_{on} is diffusion limited on rate, and k_{off} is the off rate that is determined by the electrostatic interactions and will vary between morphogens. Using chemical reaction notation, we can write the above rate law as:

$$[M] + [B] \Leftrightarrow [MB]. \tag{3.2}$$

In equilibrium we have:

$$K_a = \frac{[MB]}{[M][B]} = \frac{k_{on}}{k_{off}}, \tag{3.3}$$

where the K_a is the association constant (binding constant). The Binding site is either occupied (o) or unoccupied (u) hence the total (t) amount of B is conserved :

$$B_t = B_u + B_o \tag{3.4}$$

From this equation one is able to construct a Binomial probability space[2] [3]. The fraction of occupied binding site B determines the distribution's parameter P:

$$P = \frac{B_o}{B_t} \tag{3.5}$$

which can be rearranged in terms of the concentrations (i.e. [MB] and [B]):

$$\frac{B_o}{B_t} = \frac{[MB]}{[MB] + [B]} \tag{3.6}$$

$$P = \frac{[MB]}{[MB] + [B]} = \frac{K_a[M]}{1 + K_a[M]}. \tag{3.7}$$

Again, K_a is the association constant of the morphogen, M, to the binding site B. Previously we have denoted this constant as $K(S)$, where S is a DNA sequence that functions as a binding site for the morphogen.

[2] If we encode the bound state and unbound state in the Bernoulli Random variable (i.e. 1,0) we have: $E(X) = \mu = \sum_i X_i P_i = 1P + 0(1 - P)$

[3] $\sigma^2(X) = \sum_i (X_i - \mu)^2 P_i = (1 - P)^2 P + (0 - P)^2 (1 - P) = (1 - P)P$

3.1.2 Fractional occupancy of CRMs containing multiple binding sites

Given that most regulatory sequences have multiple binding sites for multiple different morphogens one has a master equation governing the binding process, a set of coupled differential equations. Here, again, we will assume the binding process has equilibrated, and hence we simply must enumerate the states (configurations of bound and unbound sites) of the many-binding site system. For independent binding this is simply multinomial process (where the 'multi' aspect accounts for the different types of transcription factors binding to the CRM). For example, for a single type of transcription factor binding, to n binding sites within a CRM, one has a binomial process governed by the partition function $(1+q)^n$, as discussed in the introduction of the dissertation. Hence, the probability of k bound factors is simply $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$, where p is a Boltzmann probability for a single bound site, and $p^k(1-p)^{n-k}$ is joint probability of a particular binding configuration (*which* configuration is irrelevant for the case of identical sites binding the same morphogen). However, for the case of dependencies between the sites, we can not factorize the joint distribution over binding sites (*i.e.* $(1+q)^n$ is a factorization of the partition function, and hence factorization of the joint distribution). Furthermore, for the case where the sites are not identical, such as in CRMs that have heterogenous binding sites due to each transcription factor's DNA sequence specificity, we do not simply have one Boltzmann probability p for all the sites. Hence, each site will need a distinguishing notation. Furthermore, the number of sites n is not always known. Hence, in some cases, one must discover n sites, by 'annotating' the CRM which defines the coordinates of each transcription factor's sites and their corresponding energy. In the next three sections we will describe a notation that is a mixture of notations used

for Hidden Markov Models (HMM) and Lattice Gases (Ising Models) that is hybridization of the notations from Segal et.al.[91] (HMM notation) and the notation used by both Xin He et.al.[51][50]. Hereafter these notations will simply be referred to as Segal's notation or Xin's notation. The reason for the hybridization is because our model was originally based on the Segal's paper, which latter was merged with work from Xin He (who collaborated with others in Saurabh Sinha's lab and S. Zhong's lab), by usng Xin's GEMSTAT and STAP C++ programs as libraries for implementation of our model.

3.1.3 Segal's Hidden Markov Model

If one does not know the binding sites *a priori* of a CRM then one could take all possible positions within the sequence as the start of a new binding site as in figure 3.1, where the PWM (see chapter 2) scores each position of the sequence as a potential binding site, thereby creating a list of binding energies ordered according to the position within the CRM. Repeating this for each morphogen's PWM, and by stacking the lists on top of each other, results in a matrix of energy scores that is useful for computation of the partition function through the forward algorithm from Hidden Markov Models (HMMs).

The partition function of the many body system requires the addition of the statistical weight (see the Boltzmann factor in equation 3.9) from each possible 'path' through the matrix. Each 'path' through the matrix represents a 'configuration' as in figure 3.2. The partition function can be calculated by recursively moving through the matrix elements, in a way that is similar in spirit to standard algorithms of multiple sequence alignment[32], where we can think of each configuration as a possible way to 'align' each motif (PWM) to the CRM. The variant of HMM's forward algorithm approach used by Segal for computing the partition function of all paths (configurations) is discussed in his supplementary material.

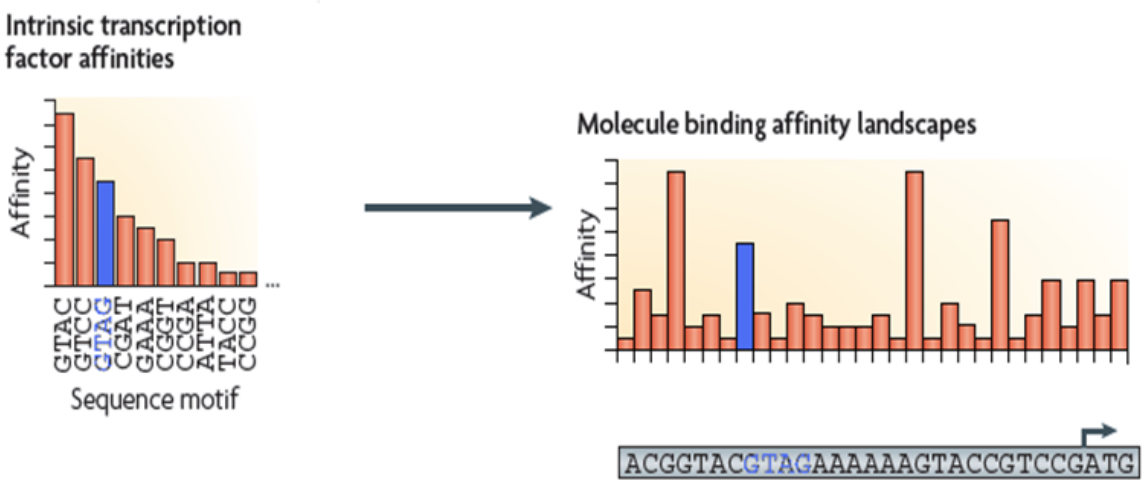

Figure 3.1: Widom, Segal Nature Review Genetics; "motif" denotes path through the PWM[92]

3.1.4 Enumerating the configurations of a CRM sequence

Here we will introduce a notation for the weights of many-binding site systems. First, let $P(c)$, be the probability of a particular configuration (c) occurring

$$P(c) = \frac{W(c)}{\Sigma_c W(c)}, \tag{3.8}$$

where $W(c)$ is the boltzmann factor or weight of the binding configuration, which is a list of occupations for each binding site locus within the CRM, where the occupation is either bound or unbound. Using the notation of Xin, from Sinha lab (which was using Buchler's model as a guide[51][20], a notation used by Terrell Hill that I followed in the Introduction

of the Dissertation), we have:

$$W(c) = [\prod_{i}^{N} q(x)_{tf(i)}][\prod_{j}^{N-1} \omega_{tf(i),tf(j)}(d)] \tag{3.9}$$

Here we have N bound transcription factors, where $q(x)_{tf(i)}$ is the weight of the transcription factor for the ith binding site $(tf(i))$ that is **bound** at position x of the sequence,

$$q(x)_{tf(i)} = K_{s(x)}^{tf(i)}[tf(i)] \tag{3.10}$$

Here $K_{s(x)}^{tfi}$ is the equilibrium constant K_a from equation (3.3) for the binding site sequence s binding by transcription factor i. And $\omega_{i,j}$ is the interaction between the two **bound** factors $tf(i), tf(j)$, where we have one less nearest neighbor interaction than the number of bound factors. And d is the distance separating factor $tf(i)$ from factor $tf(j)$ in units of base pairs ($d = x(tf(i)) - x(tf(j))$ where $x(tf(i))$ is the sequence coordinate of factor $tf(i)$).

3.1.5 The configuration vector nomenclature

Here we have used 'c' to denote the binding configuration of bound and unbound factors on the CRM following the symbol Segal used to denote the configuration, and we have used a map $tf(i)$ to link each binding site i to a transcription factor. This map is similar to notation used by Segal, where he explicitly denotes the type of transcription factor at each bound site.

Xin used a configuration vector, σ, which had indicator variables σ_i for ith componenet of their configuration vector, where the ith component was the ith binding 'site'. Xin's notation is elegant, in that every 'site' in their model is clearly represented in their configuration vector

by the occupancy of the ith component of the vector.

Segal's notation does not display all the possible bound and unbound positions in the configuration (since there is a background occupancy in his model that causes unbound sequence to have a Boltzmann weight of 1.). I have tried to stick to Segal's notation for a configuration by only displaying bound factors in the Boltzmann weight of a configuration, and by denoting explicitly the type of transcription factor at each bound site. However, Segal's notation is based only on the CRM sequence, he does not actually have a 'site' notation, since every possible position in the CRM is considered a binding site for each transcription factor, and he simply looks at all the possible ways that coordinately regulating transcription factors could form monolayers on the CRM, without overlapping one another. Hence, in Segal's notation, there really is no notion of a 'functional' site, as each *possible* position within the CRM is a place for the factor to 'plant' itself (a place for the transcription factor to bind), where *possible* is distinct from *probable* by the use of PWMs.

Seeing that binding sites are 'functional' (adaptations), or at least vestigals or exaptations, the notation of Xin for defining a 'site' we have tried to hybrid with Segal's notation. Xin's notation is fundamentally based on binding 'sites' (adaptations). For example, by either setting a threshold on a PWM to discover sites, or by knowing *a priori* what are the binding 'sites' (regulatory adaptations), Xin starts the configuration problem (the notation) with N binding sites. The indicator variables of the configuration vector are equivalent to occupancy of each site, hence σ_i is either 0 or 1 for each component of the configuration

vector. For example:

$$W(c) = W(\sigma) = W(\sigma_1, ..., \sigma_N, \sigma_{1,1}, .., \sigma_{N,N}) \quad (3.11)$$

$$= \exp\left(-\sum_i^N ln(q_i)\sigma_i - \sum_j^N ln(\omega_{ij})\sigma_{ij}\right) \quad (3.12)$$

Here, $ln(q_i)$, is the free energy of binding to the ith site, and σ_i, denotes the occupancy of the ith 'site', and σ_{ij} denotes the pairwise energetic interaction $ln(\omega_{ij})$ between site i and site j[1]. But what type of transcription factor is binding to the ith site? In Xin's notation, this is not decipherable. Rather, in their notation, each 'site' corresponds to a particular factor. But we simply don't know which factor. Hence in Xin's notation it's possible that two distinct 'sites' occupy the exact same locus. Hence we don't know if two sites are overlapping. Overlapping sites can not both be bound, since all thermodynamic model's for occupancy of transcription factors use 'hard sphere potentials', there's steric hinderance prohibiting two transcription factors to occupy the same space along the CRM.). Hence, in Xin's notation, one is supposed to be cognizant that bound overlapping sites will set that configuration's weight to zero.

We have used a hybrid notation in Eq.3.9, where it's not that our CRM has N fixed binding sites (like in Xin's notation) that are each bound or unbound. Rather, the weight $W(c)$ displays N bound sites from a configuration c. Hence, in our notation N is a variable in the binding configuration space, while in Xin's notation, N is the fixed number of sites. In Xin's notation, advantageously, they can put a bound on configuration space as 2^N, this is an upper bound because overlapping sites will reduce the total number of configurations,

[1]Recall for the canonical ensemble, where H is the Hamiltonian of a configuration we would have $W(c) = \exp -H(c)/kT$ being a simple Boltzmann factor, where kT is the thermal energy. We are working in a grand canonical ensemble, which allows for particle energy as well as energy exchange, hence, the total energy of the configuration is a function of the Hamiltonian and the chemical potential of the factors, which causes concentration of the transcription factors to influence the total energy of a configuration, hence the energy is a thermodynamic free energy

furthermore this overcounts the number of unbound states, since a loci's sequence that match for multiple factors - say m factors- would consider $2^m/2 - 1$ too many configurations, since the loci is actually only unbound in one possible way[2].

3.1.6 An example of the hybrid configuration notation

For an example of the hybrid notation, consider all possible positions of the CRM as a binding site, as Segal would, and multiple morphogens binding to the same site (same position within the sequence, same locus); then the configuration vector (c) can be written in terms of Xin's indicator variables (similar to Ising models).

For a length L CRM and for 3 transcription factor types (like Dorsal,Twist, and Snail) we have in Xin's notation $N = 4 * L$, where the factor of four is due to the four types of transcription factors at each locus - namely: bound by Dorsal, Twist, or Snail or 'background' (background is a type of pseudoparticle that fills the unbound state)).[3]

In this case, the vector of indicator variables is written as a matrix of size 4xL. For example, for the toy CRM sequence 'acggt', we would have a configuration vector of size 20. If we add another 'state' to the configuration vector denoted as 'silent' to represent steric effects, where the type of transcription factor indicator variable now only indicates the 'start' of a binding site, we would then have a indicator matrix of size 25, as indicated in Figure 3.2, where the Dorsal transcription factor occupies positions 2 and 3 of the CRM (in this toy case Dorsal occupies a site just of length 2), and the rest of the CRM is occupied by

[2]Xin's GEMSTAT computes the correct weights and partition function, it's just the estimate on the configuration space that is a bound.

[3]As pointed out by Segal in his supplement, this problem may seem computationally infeasible, since the number of configurations for L=500 is of size 2^{4*500}, but due to the forward algorithm it is possible. Here the number of binding sites, N, neglects edge effects of k-mers requiring length k binding site sequences. Hence, the calculation of the number of configurations is only a bound on the cardinaility of the set of configurations, where the bound is further affected by the neglect of overlapping sites that cause many configurations to be unaccessible (against the rules).

the background factor. Hence the configuration vector (a value of the matrix of indicator variables) is of size N+5 (the 5 due to the 'silent' states) assuming no interactions between bound factors. If there are interactions, for example nearest neighbors, then there are an additional $\frac{16N^2}{2}$ components to the configuration vector, which we will not display.

Now that we have defined our configuration notation, and described the weight of a given configuration, we will explain in the next section our novel form of the pairwise interaction ω between bound factors.

3.1.7 The pairwise interaction ω between bound factors

Transcription factors often interact with each other, causing certain configurations to be more likely by decreasing the total energy of the configurations where interacting factors are jointly bound, thereby increasing the weight of those configurations. This observation, was the physical basis behind A. Hill's famous Hemoglobin Oxygen binding model.

In the work of Xin a number forms of distance dependent pairwise interactions between bound factors was tested, such as sinusoidal functions over space that account for 'phasing' where one bound factor interacts with the nearest neighbor that is in phase (by using the major groove distance) with the factor of interest. Similarly, gaussian decays from the center of the planted factor, and square functions were attemted and implemented. Hence, GEMSTAT has a small database of pariwise interaction forms. All the forms, as ours too, only allow for interactions between nearest neighbor bound proteins.

For the the DV network interactions between Dorsal and Twist have been experimentally shown to be distance dependent, hence we use an interaction that is a function of the DNA basepair distance separating the nearest neighbors. For the network that we are modeling this distance dependent interaction has been shown to be dominant force driving the Dorsal

S	a	c	g	g	t
dorsal	0	0	0	1	0
twist	0	0	0	0	0
snail	0	0	0	0	0
background	1	1	1	0	0
silent	0	0	0	0	1

$$c = 0001000000000011100000001$$

$$\sigma_i$$

Figure 3.2: The toy CRM sequence acggt is annotated at each of its loci (positions) to denote the configuration vector in row major ordering (1st five bits are dorsal's row, second five bits are twist's row etc...). In the language of HMMs, the value of a configuration vector reveals the hidden state of the sequence. Here there are 5 states, where the 'silent' state indicates a transcription factor is bound to an upstream position of the sequence, causing the loci to be covered by an internal position of the planted factor.

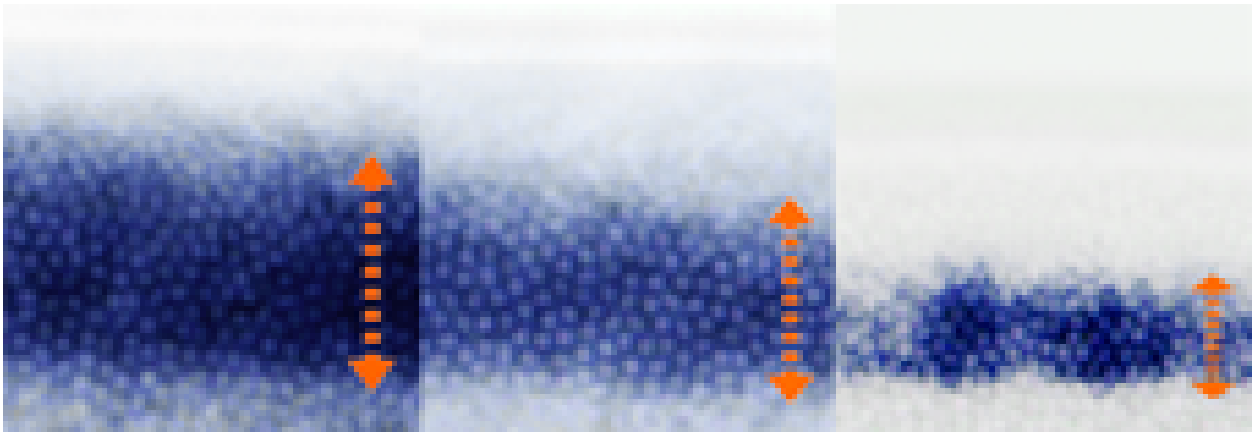

Figure 3.3: Changing the spacing between motifs in modules change the span of cells that are expressed. In this Figure the spacing between two sites was adjusted by Natural Selection in orthologs of the *rhomboid* gene's CRM. When the ortholog CRMs were transgenically inserted into *mel* species the width of the expression pattern changes relative to the endogenous pattern width, suggesting that coopertivity between the sites is a function of the distance between sites that can be used by evolution to 'fine tune' the epression patterns in development. When each sequence or module is expressed in its respective specie (lineage), then the relative widths (w/L where L is the lengths of the major axis of the specie's embryo, w is the width of the tissue in nanometers that express the gene) of the tissues are the same. Different specie's embryos have different sizes hence there is a scaling law - how a characteristic (such as gene expression) changes with body size. This Figure is from Erives Crocker et.al 2008 "Evolution acts on enhancer organization to fine-tune gradient threshold readouts. PLoS Biology"

border of neuroectoderm expressed genes. For the neuroectoderm network[2] Crocker et.al. and Szymanski et.al. [102],[26]have shown that the spacing between sites plays a dominant role in defining the number of cells that are turned on in this region (width or span of cells).

For both the DV and AP (Anterior Posterior) network short range antagonistic interactions have been show to dominate the action of repressor transcription factors[36]. The Snail transcription factor is a short range repressor acting to regulate the DV axis, which we also model using a distance dependent interaction, commonly called 'Quenching'[36].

Furthermore, since we do not know the exact form of the function, we bin the distance

[2]neuroectoderm network are the set of genes coordinately expressed in the lateral regions of the developing embryo, this developing tissue spans about 10 cells at the time point under consideration

separating the proteins, and for each bin a free parameter is fit. One may look for a coarse binning of the separation distance in bins of 10bp, or as fine as bins of 1bp. If we choose the former option than an example of our protein-protein interaction parameter would be as follows:

$$\overrightarrow{\omega(d)} = [\omega_1, \omega_2, \ldots, \omega_b, \ldots, \omega_n] \tag{3.13}$$

where the subscripts of the components of the ω vector (coopertivity or synergistic protein-protein interactions) represent the corresponding bin ,b, (in this case there are n bins), hence they define a bin vector, $\overrightarrow{B}$, where its components corresspond to the interval of base pair distances.

$$\begin{aligned} \overrightarrow{B} &= [B_1, B_2, \ldots, b, \ldots, B_n] \\ &= [(0-10), (11-20), \ldots, (41-50), \ldots, (x-L)] \end{aligned} \tag{3.14}$$

Here bin b, is all interactions where two bound sites are separated by 41-50 basepairs, and x represents the last bin border, and L represents the Length of the module (sequence). This representation of coopertivity vector, is really only useful for programming purposes, mathematically the coopertivity is simply a piecewise defined function:

$$\omega(d) = \begin{cases} \omega_1 & \text{if } d \in (0, 10) \\ \omega_2 & \text{if } d \in (11, 20) \\ \quad \vdots & \\ \omega_n & \text{if } d \geq x \end{cases}$$

Our aim is to fit the biochemcial parameters that tune the probability of configurations that occur in live embryos. However, we simply don't have such detailed biochemical experiments. Hence, we use the mRNA of the target genes regulated by Dorsal Twist and Snail as a readout of what binding configurations are likely occurring, and wheter certain configurations have strong linkage (pairwise interactions) between certain bound factors. Hence, we must infer from mRNA data, what is occuring at the DNA binding level; an expression to sequence model. In the next section will describe a ubiquitous and obvious assumption, mRNA is caused by PolII, hence PolII binding is a proxy for gene expression.

3.1.8 Relating the number of mRNA transcripts to fractional occupancy of PolII

The amount of transcription that occurs at a gene locus is encoded in two segments of DNA sequence; first, the basal promoter that binds the Basal Transcription Apparatus (BTA), which is a massive complex of many proteins including PolII; second, and most important, the CRM that binds the morphogens or distal transcription factors (such as Dorsal) that modify and remodel the chromatin state and possibly have direct linkage with the BTA. Hence the number of mRNA transcripts can be modeled as simply a linear relationship between the number of mRNA and the fractional occupancy of a promoter sequence (which we assume includes the CRM sequence).

$$\langle N_{mRNA} \rangle \propto f_{BTA}. \tag{3.15}$$

Of course, the occupancy of the BTA, f_{BTA}, is a fraction this is at most one, hence $\langle N_{mRNA} \rangle$ is the average number of mRNA molecules produced per nuclear cycle normalized by the

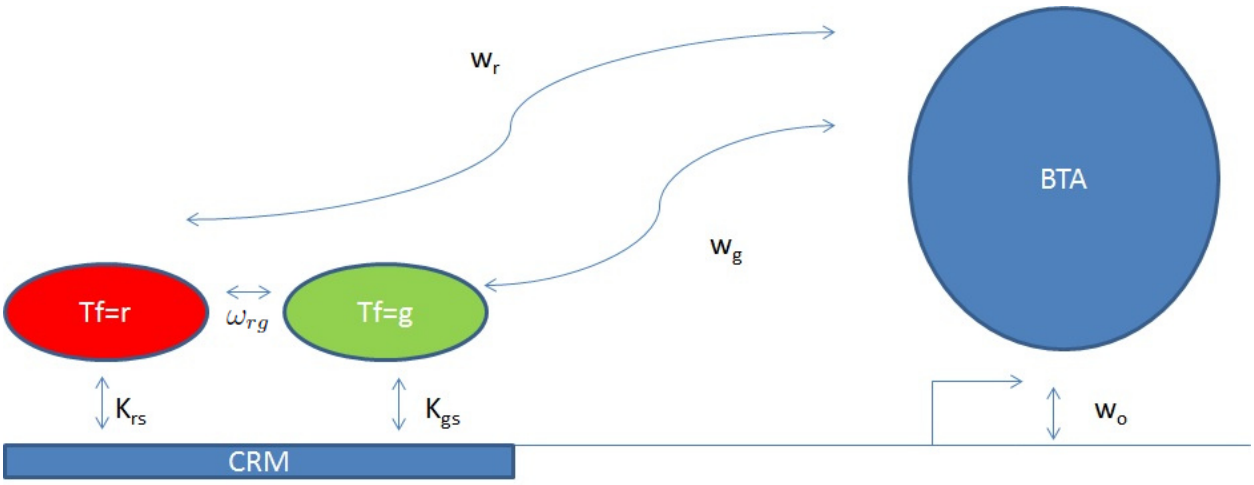

Figure 3.4

maximum production rate over a cycle.

3.1.9 Fractional occupancy of BTA

We can gain more physical insight into the problem by thinking of mechanisms of how BTA's occupancy is a function of the morphogen occupancy. In Figure 3.4 we see that when the morphogen's are bound they have an activation or repression domain, which communicates with the BTA, for our case this communication may be thought of as the complex process of changing the epigenitic state of the chromosome, by coactivators (histone modifiers, nucleosome remodelers) binding to the morphogen. The binding energy of this domain (w_m) on the mophogen can be related to the binding energy of the BTA:

$$\Delta G = w_0 + \sum_m n_m(c) w_m \tag{3.16}$$

Here c is a particular configuration of the morphogens on the promoter, w_0 is the binding energy contribution from the basal promoter, w_m is the binding energy contribution from

each morphogen, and $n_m(c)$ is the number of bound morphogen of species m to state c (for example see [50]). For this configuration we could model the occupancy of the BTA as:

$$f_{BTA} = \frac{1}{1 + e^{-(w_0 + \sum_m n_m(c) w_m)}} \tag{3.17}$$

The fast pace timing of morphogen binding relative to transcription time (e.g. the time for PolII to clear the promoter, and a new (or preloaded on the enhancer) PolII binds), would suggest that the BTA is not sampling or cognizant of each CRM configuration of bound proteins, rather the BTA sees an average occupancy of the morphogens. This can be modeled as:

$$f_{BTA} = 1/(1 + \exp -(w_0 + \sum_m \langle n_m \rangle |_c \quad w_m)), \tag{3.18}$$

where the occupancy of the BTA is related to the theoretical model of average morphogen m occupancy over the CRM configurations $\langle n_m \rangle |_c$[4]. The average morphogen occupancy over configurations is:

$$\langle n_m \rangle |_c = \sum_c n_m(c) \frac{W(c)}{\sum_c' W(c')}, \tag{3.19}$$

where we the normalized weights of each configuration are from Equation3.8, and $n_m(c)$, is the number of morphogen of type m (such as $m = Dorsal$) bound to the configuration c. The expectation value is computed using GEMSTAP software from Sinha's lab, see the following reference for further details[50]. Here after, we simply use $\langle n_m \rangle$ to denote $\langle n_m \rangle |_c$ with the understanding that the expectation value is taken over the binding configurations of the CRM.

[4]This can also be shown, under certain conditions, to be an approximation of the much more computationally intense model of Segal et.al[91] where they compute $\langle f_{BTA} \rangle |_c = \left\langle \frac{1}{1+e^{-(w_0 + \sum_m n_m(c) w_m)}} \right\rangle |_c$. However, as stated above, this is not an approximation of Segal's model, it's a different model of the interaction of the BTA with the CRM.

3.1.10 Fractional occupancy of BTA from a binding reaction perspective

Let P = PolII concentration(this is the best known protein in the Basal Transcription Apparatus, but technically P is the concentration of the BTA), D = DNA (or basal promoter, i.e. TATA box, and binding sites for TF2B etc..), C = complex (PD). We assume the binding process is in equilibrium (different timescale than morphogen binding).

$$P + D \Leftrightarrow C \tag{3.20}$$

Using fractional occupancy we have:

$$\frac{C}{C+D} = \frac{1}{\frac{D}{C}+1} \tag{3.21}$$

$$C = P * D * K_a \tag{3.22}$$

$$\frac{C}{C+D} = \frac{1}{\frac{D}{PDK_a}+1} = \frac{1}{\frac{1}{PK_a}+1} \tag{3.23}$$

now assuming concentrations of unbound PolII is unaffected by our DNA D,we can say free PolII is a constant like 1000, and absorb the constant into the K_a,

$$\frac{1}{\frac{1}{PK_a}+1} = \frac{1}{\frac{1}{K'_a}+1} = \frac{1}{e^{-\frac{\Delta G}{k_b T}}+1} \tag{3.24}$$

Now ΔG is the free energy that is released during morphogen binding, so we can equate ΔG to equation (3.16), however we will not say PolII 'sees' a particular configuration, rather we

will assume PolII sees the average configuration (i.e. $< n_m >$ the average number of bound morphogen, m)

$$\Delta G = w_0 + \sum_m < n_m > w_m \tag{3.25}$$

$$< n_m >= \sum_c n_m(c) P(c) \tag{3.26}$$

plugging in the energy in units of $k_b T$ from equation (3.25) into equation (3.24) we arrive again arrive at equation (3.18):

$$\frac{1}{\frac{1}{PK_a} + 1} = \frac{1}{e^{-(w_0 + \sum m < n_m > w_m)} + 1} \tag{3.27}$$

This yields a range of $N_{mRNA} \in [0, 1]$.

3.1.11 Fractional occupancy of BTA in Cooperative Binding (CB) model in Xin He's GEMSTAT

The fractional occupancy in Xin He's Cooperative Binding model (CB) is closely analogous to our form of the model[5]. This can be seen by taking the simple system of one morphogen binding site with one basal promoter site for the BTA. Hence, the BTA is treated as if it were simply another morphogen. In this case the partition function of the two site system (Ξ) for the case that the two sites are independent is: $\Xi = (1 + q)(1 + q_{BTA})$, where q is the canonical partition function of the morphogen bound to its site and q_{BTA} is the canonical partition function of the BTA bound to its promoter. Now if we assume the sites

[5]This CB is not our related to our CB PWM model from chapter 2.

are dependent, then we can not factorize the joint partition function[6] and we then have $\Xi = (1 + q + q_{BTA} + q_{BTA}\ q\ w_{tf}) = Z_{off} + Z_{on}$, where we have collected similar terms in the expansion, where $Z_{off} = 1 + q$ is the term that does not include BTA binding, and $Z_off f$ is the collection of the remaining terms (where BTA is bound)[7]. In their CB model they have $f_{BTA} = \frac{Z_{on}}{Z_{off}+Z_{on}} = \frac{1}{Z_{off}/Z_{on}+1}$. By equating this to our form of the occupancy (Eq.3.27) we have:

$$Z_{off}/Z_{on} = \exp -w_o + \sum_{tf} w_{tf} < n >_{tf}. \tag{3.28}$$

The above equation is the odds *for* BTA binding, hence the log odds is:

$$\ln Z_{off}/Z_{on} = -w_o + \sum_{tf} w_{tf} < n >_{tf} . \tag{3.29}$$

Now, if we assume there is only one morphogen binding site, then the denominator of $< n >_{tf}$ is Z_{off}, and its numerator is q. Hence we have:

$$\ln \frac{(1+q)}{(q_{BTA} + q_{BTA}\ q\ w'_{tf})} = -w_o + w_{tf}\frac{q}{1+q}, \tag{3.30}$$

where we have replaced the Z's with their original q's and the w'_{tf} denotes Xin's form of the cooperativity between BTA and morphogen (in order to distinguish from our cooperativity's symbol w_{tf}. Using the properties of logarithms we can isolate $(1 + qw'_{tf})$ from the left side

[6] As always, we can *organize* the states over our many-body system by systematically building the configurations, even in the case of dependencies, through a polynomial expansion over binding sites, which is presented in Xin He's Supplementary Material and T.Hill's text[52].

[7] In Xin's Supplement, w_{tf} is denoted as α (which is a different parameter than our α that approximates the morphogen binding constant, where we choose the symbol α to mimic Segal's choice of parameters) w_{tf} is the cooperative binding between the morphogen and the BTA

of the equation, which leads to:

$$\ln(1+q) - \ln q_{BTA} - \ln(1+qw'_{tf}) = -w_o + w_{tf}\frac{q}{1+q}, \tag{3.31}$$

where w_o is $\ln q_{BTA}$ leaving the result:

$$\ln(1+q) - \ln(1+qw'_{tf}) = w_{tf}\frac{q}{1+q} \tag{3.32}$$

rearranging we have:

$$(1+q)\ln(1+q) - (1+q)\ln(1+qw'_{tf}) = w_{tf}q \tag{3.33}$$

if q is less than one (morphogen has low concentration, for example), then the first term on the left is approximately zero, and if w'_{tf} is not too large then to first order the Taylor expansion of $\ln(1+qw'_{tf})$ is -qw'_{tf}. Hence it appears under this regime that Xin's CB model, with free parameters α are equivalent to our w factors.

3.1.12 Fractional occupancy of BTA in Ay's model

The fractional occupancy model in Fakhouri et.al.[36], which we'll call Ay's model is similar to Segal's model, and hence similar to our form of the BTA occupancy[8]. In Ay's model

[8] A distinguishing characteristic between Segal and Ay's work was that Ay's model contained high quality expression data (with error bars)[], while the input expression data to Segal's model was a boolean based 'on' 'off' data, which was smoothed (for example by using cubic splines). Furthermore Segal's CRM input was just the sequences and PWMs of the morphogens regulating the CRMs (where the binding sites were to be 'discovered' using a PWM annotation model), while Ay's model knew exactly where the binding sites were within the CRMs, thereby not being hampered by false postive binding sites from PWM prediction of sites such as in Segal's arpproach. Once Segal's model had annotated a CRM with the morphogen binding sites, his model and Ay's model, we aim to show in this section, were identical (assuming the annotation had no false negatives and false positives).

the probability of each configuration is an element in a vector of probabilities, where the vector is labeled $\mathbf{F}$ (where $\sum_c F_c = 1$, where c encodes the binding configurations as the component index to the vector.). Similar to Segal's model, each configuration causes a particular occupancy of the BTA, where the occupancy of the BTA for each configuration is a component of a vector $\mathbf{T}$. Hence $\mathrm{T} \bullet \mathrm{F} = \sum_{\mathrm{c}} \mathrm{F_c T_c}$ (which given one knows all the configurations, this is then very similar in form to Segal's model $< \frac{1}{1+e^{f(c)}} > \approx < T_c >$, where the expectation is taken over the binding configurations and $f(c)$ is Segal's function of each configuration.

A departure from Ay's model and our model is the quenching function that he denotes as $q(d)$, where d is the spacer distance in base pairs between the repressor and the activator. Their model, like ours, bins the spacer distances to form the queching function (q is a piecewise defined function over the different intervals of spacing between the repressor and activator)[9]. In his model each interval has a free parameter to be trained, which is analogous to our form of the pairwise potential for repression, for example $\omega_{Sn,Dl}(d)$ (the repression pair-wise potential between Snail and Dorsal), which is also a binned pair-wise potential in the same form. However, the parameters depart in that our ω's occur in the partition function just as in Segal's model, while Ay's quenching parameters do not. His quenching parameter goes back to a model form from John Reinitz([88]), where quenching modulates the probability of configurations bound by the repressing transcription factors. In this model modulate means that the $p = 1$ norm of their probability vector $||F|| = \sum_c F_c^p = \sum_c F_c$ is a function of the amount of repression[10].

[9](In the case of multiple repressor types, one would have a notation such as: $q(d)_{tf}$, where each repressor type has its own quenching function, or possibly just one universal quenching function, if all repressors ended up behaving the same. What a discovery that would be!)

[10]The norm of the vector here is not a standard vector norm, rather it is a $p = 1$ norm in mathematics, where the p norm is defined as $||x| = |\sum_i x_i^p$.

For example, a module responding to activators under high concentrations, but with very low concentrations of repressor (making the Boltzmann weights of repression configurations zero) has the usual norm $\sum_c T_c = 1$, while if the conditions are favorable for repression, (such as both high concentrations of repressors and activators AND a very strong repression pairwise potential), then if the q(d) function is large (its largest value is 'one') for the spacer distances d that occur between the activators and the repressors, then configurations with bound repressors have their probabilities modulated by factors of $(1 - q(d))$ for each bound repressor, where d is the spacer between each repressor and its nearest neighbor (while there were other forms, or schemes, tested for this function rather than just nearest neighbors).

For example, for a module with 5 activator binding sites and 5 repressor binding sites (with no sites overlapping), then for the configuration with all binding sites bound, and each activator had a nearest neighbor bound repressor with a spacer at d=10 bp, then the Boltzmann probability of the all bound configuration is modulated by a factor $(1 - q(d = 10))^5$, and the result of this is: $F'_c(1 - q(d = 10))^5$, where c' is the configuration with all 10 sites bound[11]

The difference between Ay's model of quenching and ours is due to the quenching parameters (function) not occurring in the partition function. Here we see if it is possible to *equate* Ay's model form $\sum_c F_c T_c$ to Segal's model. First let the occupancy function of the BTA, which is a function of the configuration, be denoted as:

$$T(x) = \frac{1 - q}{1 + exp(-x)} \tag{3.34}$$

[11]In Ay's model, the configuration vector c takes a similar form to our configuration vector, for example, Ay's representation of the example configuratin with 10 bound sites is $c' = [ARARARARAR]$, where A is for activator and R is for repressor. This is a compact representation of our configuration vector in Figure3.2, compact in the sense that our configuration vector also displays information about the spacers and internal positions of a binding site (a 'silent' state)).

where q is the quenching efficiency from Ay's model, and x is a function of the binding configuration. Ay's exact form was:

$$T'(x) = \frac{1}{1 + \exp(5 - x)}, \tag{3.35}$$

which we wish to adorn with the factors (1-q) (the probability modulation factors that contain the quenching parameter q), which will then allow the BTA occupancy to be modulated, thereby keeping the Boltzmann distribution over the binding configurations unperturbed by repressor morphogen binding. Now, make a change of variables, let $(1 - q) = e^{+w}$, where the symbol w has no particular meaning other than to capture the change of form in the expression. Now we have:

$$T(x) = \frac{\exp + w}{1 + \exp(-x)}, \tag{3.36}$$

whereupon bringing the numerator's factor into the denominator we have:

$$T(x) = \frac{1}{\exp(-w) + \exp(-x)\exp(-w)}, \tag{3.37}$$

Now we simply want this denominator to be of the form $1 + \exp(-x')$, which would allow Ay's model to be expressed exactly in the same form as Segals. Hence we again make the change of variables:

$$\exp(-w) + \exp(-x)\exp(-w) = 1 + \exp(-x'), \tag{3.38}$$

which leads to:

$$-x' = ln(\exp(-x) + q) - \ln(1 - q) \tag{3.39}$$

Hence we have:

$$T(x) = \frac{1}{1 + \frac{\exp(-x)+q}{1-q}} \tag{3.40}$$

Now, it appears that the odds for BTA occupancy from Ay's model: $\frac{\exp(-x)+q}{(1-q)}$ must be equal to the odds for BTA occupancy from Segal's model: $e^{f(c)}$. Setting the log odds equal to each other may allow for some deduction on what certain parameters 'mean' in terms of the two models, which would then be able to be traced to our model.

3.2 Data set

We would like to know the key biochemical parameters that are utilized by Dorsal, Twist, and Snail as they regulate the genes that pattern the Dorsal Ventral axis of early development by producing expression profiles, a list of mRNA counts for each position along the DV axis. Under standard nonlinear regression, one could imagine an experiment where one measures the response (the dependent variable) to systematic variations of the inputs (independent variables), these measurements can then be used to fit a nonlinear regression model, where the fit parameters are our biochemical parameters (such as the binding constants). Here the response, is the gene expression levels, and the inputs are the morphogen and CRM information (described further in the data section). This is a massive amount of experimentation in order to fit a nonlinear model. However, following the interpretation of Segal et.al.[91], which in a sense, is a reinterpretation of Ed Lewis' model of homeogenes, and further refined by Zinzen et al.[108], we can treat *coordinately regulated* genes as if they were the *same* gene, just under different inputs.

What are the different inputs? The CRMs and positional information of the embryo (in terms of morphogen concentrations).

How is this possible? During development morphogens work together to coordinately regulate a set of genes.

What about replicates, as a well designed experiment has statistical *power*? The number of degrees of freedom, which determines the statistical power, can be assumed infinite, as across a population of embryos, each with thousands of nuclei that each have a genome, we know the genome's are identical (neglecting short term evolution and segregation of SNPs during sex, where, SNPs in inbreed lab lineages is irrelevant.). Each embryo is effectively a clone of one another.

So, for the regulatory regions of genes, a modeler only effectively needs one sample to have complete knowledge about the DNA sequence, but what about the *trans* enviroment, the numbers of each molecules in each embryo, doesn't that vary across embryos? All embryo's are the same within limits of diffusion processes. Each embryo across a population of embryo's is producing mRNA at each coordinately regulated gene with a high degree of precision. Morphogen concentration gradients and the gene's expression response (mRNA levels) *are* highly reproducible across different embryos, within the physical limits set by diffusion processes[45]. So yes, the internal molecular environment from one embryo to the next does vary due to random walks of molecules. However, our error in assuming that a population of embryos are all the same will be no bigger that $\sqrt{n}$, where n is absolute number of a molecule type in each cell of an embryo. Furthermore, $n(t)$, where t is time, varies on a time scale much slower than the processes we will study. Hence the *diffusing* concentraition function across the embryo $n(z,t)$ where z is space (the DV axis), we assume is frozen. For large n, the fractional error that we incur in our model is quite small. Hence, variation in molecule in absolute concentration from one embryo to the next (or even within an embryo from one cell to the next cell - at the same position along the DV axis) is negligible.

Hence, we will treat an entire network of regulatory sequences as if they are each just a different measurement, a different input to our model. This is a reasonable interpretation, however, nonlinear models are notorious to being sensitive in certain intervals of the input variables, and hence if the network has not just happened to evolve to coordiantely expressed gene's in the input regions where our model is sensitive, the model will fail, and one must resort to the tedious work such as Fukouri et.al.[36] to assure that all necessary data points are being collected (e.g. if one wants to measure distance dependent quenching, then one, under Segal's interpretation, should hope that evolution has selected a range of different spacers that coordinate quenching within the CRMs, otherwise the model fitting will be insensitive to this parameter. Why? Because there simply is no input data that varies the spacing, which is a prerequisite for fitting distance dependent interactions.). However, the key biochemical parameters we are interested in was motivated by previous experiments of endogenous CRMs that have pointed to our parameters of interest of being the major contributing factors. The model requires three main pieces of data denoted as $\mathbf{D}$. First, the CRM sequences and PWMs of the morphogen's targeting the sequences. Second, the morphogen concentrations along the DV axis at a particular point in time in development. Third, the response of the CRM's to the morphogens (also in the form of 'expression' concentrations along the DV axis at a particular point in time in development).

$$\mathbf{D} = \{\mathcal{S}_t^{crm}, \mathcal{E}_t, \mathcal{E}_{tf}, \mathbf{PWM}_{tf}\} \tag{3.41}$$

$$\mathcal{S}_t^{crm} = \begin{pmatrix} s_{rho}(1) & \dots & s_{rho}(L_{rho}) \\ s_{vnd}(1) & \dots & s_{vnd}(L_{vnd}) \\ \vdots & \ddots & \vdots \\ s_n(1) & \dots & s_n(L_n) \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{rho} \\ \mathbf{S}_{vnd} \\ \vdots \\ \mathbf{S}_n \end{pmatrix}$$

$$\mathcal{E}_t = \begin{pmatrix} E_{rho}(1) & \dots & E_{rho}(m) \\ E_{vnd}(1) & \dots & E_{vnd}(m) \\ \vdots & \ddots & \vdots \\ E_n(1) & \dots & E_n(m) \end{pmatrix} = \begin{pmatrix} \mathbf{E}_{rho} \\ \mathbf{E}_{vnd} \\ \vdots \\ \mathbf{E}_n \end{pmatrix}$$

$$\mathcal{E}_{tf} = \begin{pmatrix} E_{Dorsal}(1) & \dots & E_{Dorsal}(m) \\ E_{Twist}(1) & \dots & E_{Twist}(m) \\ E_{Snail}(1) & \dots & E_{Snail}(m) \end{pmatrix} = \begin{pmatrix} \mathbf{E}_{Dorsal} \\ \mathbf{E}_{Twist} \\ \mathbf{E}_{Snail} \end{pmatrix}$$

$$\mathbf{PWM}_{tf} = \begin{pmatrix} PWM_{Dorsal}{}_{DC}, PWM_{Dorsal}{}_{DU} \\ PWM_{Twist} \\ PWM_{Snail} \end{pmatrix}$$

Sequence Part of the data

The most important input of the model are just the the CRM sequences. The model is the response of the CRM across the entire DV axis, and hence, each CRM will will have response values across all the positions of the DV axis. Each row in $\mathcal{S}_t^{crm}$ is a DNA sequence of the module that drives the *target* expression (concentration) of the same row in $\mathcal{E}_t$ as a function of concentrations of the morphogens and the morphogens' PWMs' annotations on the crm

$\mathcal{S}_t^{crm}$ (through binding site discovery). The CRMs are usually about 500bp that control a given gene in the DV network, or the CRM was an engineered construct that was tested *in vivo*. The columns of $\mathcal{S}_t^{crm}$ are the ordered nucleotides of the DNA of length L, where each base of the sequence $\mathbf{S}_t$ is represented as s(i), and the subcripts on the CRM indicate the label for the target gene it controls.

Positional-dependent target gene response data $\mathbf{E}_t$

The response of CRM to morphogen regulation is the data $\mathcal{E}_t$ which is the CRM's target gene's expression. The target expression is controlled in *cis* by a given CRM, hence each row of $\mathcal{E}_t$, denoted as $\mathbf{E}_t$, corresponds to the same row in $\mathcal{S}_t^{crm}$. The target expression is controlled in *trans* by the morphogen conentrations at a position, z, along the embryo. Each column of $\mathcal{E}_t$ represents the position z along the embryo.

Positional-dependent morphogen data

The transcription factors Dorsal, Twist and Snail also each have an expression profile along the DV axis, which is stored in the table $\mathcal{E}_{tf}$, where tf is the particular factor, and where each factor's profile must be the same length vector as the target gene profiles. There are about forty cells (nuclei) along the DV axis at the time point in development under study (in Foe's time table the time of development is 'stage 4'), hence each nuclei has just one locus containing the CRM of interest (technically, Drosophila is 'diploid' (containing two genomes in each nucleus, but we don't model this), hence it is natural to demaracate the positions along the DV axes into about forty bins.

Hence the columns of $\mathcal{E}_{tf}$ and $\mathcal{E}_t$ represent the concentrations at a given position along

the DV axis of the input morphogens (tf), and output target response t. However, the independent variable data $\mathcal{E}_{tf}$ is not necessarily collected jointly with the dependent variable, $\mathcal{E}_t$, much of the expression data is coming from different embryos, and different labs, so we actually do not have the exact known amount of input morphogen and output gene product for a given position along the axis. However, tedious experiments along the DV axis by a number of labs have already shown that the explanatory variable causing variation in E_t are the morphogen's we use as inputs, furthermore, fly embryos are believed to be very 'reproducible', in their patterning expression profiles, hence it is reasonable to collect the input and output profiles from different embryos.

3.2.1 Collection of data from DV network of Dorsal, Twist, and Snail targets in Neuroectoderm and Mesoderm, and PWMs

The target gene profiles and corresponding regulatory sequences were collected from the following (references)[60][59][58]. The profiles were based on the number of cells that span the mesoderm and neuroectoderm (about 40) at the time point under consideration. The positional information was extracted from the corresponding references results section. For example, the results may say "twistPE enhancer border at cells 12-14", where the mesoderm proper is known to be 18-20 cells wide at the time point under study (at this point an entire cross sectional slice of the Dorsal Ventral axis is 100 cells wide (reference) at the minor axis of the ellipsoid). Furthermore the ventral neuroectoderm is known to span 6-10 cells past the mesoderm boundary, given a bound on dorsal-most border of neuroectoderm genes, and the dorsal border is cited by the author with the uncertainty in its cellular position. The amplitudes were classified according to the comparative analysis in the results section of the

references, each class was then assigned a numeric range, for example the class that was regarded as the strongest staining was arbitrarily assigned the range [.9,1]. It should be noted that the amplitudes for Segal's 2008 paper were binary, hence his border classification ,where to assign the 0,1 transition, has uncertainty.

The expression profiles of Dorsal was based on Robert Zinzen's confocal microscopy results in the Devex database (no longer available online). The Twist expression profile was also from this database. The Snail profile, is known to be uniform in the mesoderm, and 'off' in regions of the embryo dorsal of the mesoderm, the profile is a step function. Hence, Snail is simply 'on' in the mesoderm, and 'off' in the regions dorsal of the mesoderm.

The Dorsal, Twist, and Snail profiles were coordinated with the target gene profiles by the standard assumption that the sharp Snail border demarcates that mesoderm neuroectoderm border, which I will call the Snail step. Hence, all NEE genes, or synthetic constructs of functional NEEs were registered with the position of the step in the Snail profile. Similarly, all mesoderm target genes were scaled such that they were less than or equal to the step position of the Snail profile.

The Dorsal PWM is the DC and DU form chapter 2 of this dissertation. The Twist motif used was 5'-CAYATG, and the Snail motif used was 5'-CACCTG. The Twist and Snail motifs were transformed to energy PWMs too. Hence, the energy levels of the TWist and Snail are not accurate, but, due to the threshold free algorithm for annotation, the accuracy is not as central a question as it would be for a threshold based approach.

3.3 Nonlinear regression model

3.3.1 Putting the data parts and free parameters together to form the nonlinear model of BTA occupancy

Our nonlinear regression model is simply the fractional occupancy of the BTA, which is a function over the high dimensional space of input CRMS and position along the DV axis of the embryo:

$$f(S,z) = \frac{1}{1+\exp -(w_0 + \sum_m \ \langle n_m(z)\rangle \quad w_m)}, \tag{3.42}$$

where the morphogen m's occupancy on the CRM $\langle n_m(z)\rangle$ is now a function of the position, z, along the DV axis of measurement. Hence, z is the index of our expression and morphogen profiles.

3.3.2 Free parameters to be fit

The nonlinear model has set of biochemical constants that we optimize. Hence these parameters are unknown, and trained based on the data. The parameters:

$$\beta = \{\overrightarrow{\alpha}, \overrightarrow{\omega_d}, \overrightarrow{w}\} \tag{3.43}$$

The alpha vector, $\overrightarrow{\alpha}$, is related to the protein DNA binding interaction. Recall $q_i = K(S)[tf(i)] = K_0 \exp\left(-(E(S))[tf(i)]\right.$, where K_0 is the maximum binding constant in k-mer space, and $E(S)$ is the the energy score from the PWM, and $[tf(i)]$ is the concentration of the transcription factor that corresponds to the PWM. We do not know the absolute concentration of the factor $tf(i)$, rather we have normalized laser intensities of florescent

markers for the protein in 'stained' embryos (which we assume is proportional to the absolute affinity). We have denoted this intensity data as E_{tf}. Hence, for example, for the kth bin (or kth cell) along the z axis (DV axis) we have $q_i = K(S)[tf(i)] = \exp(-E(S))E_{tf}(k)\alpha_{tf}$.

An interaction energy, $\overrightarrow{\omega(d)}$, is defined for nearest neighbor interactions for each type of interaction between species, homotypic and heterotypic (this represents a form of 'cooperativity' and a form of 'quenching'). These constants depend on binning the separation distance between interacting factors (which factors are interacting must be prespecified). For example if we assume detectable variations on the scale of 10bps for protein-protein interactions, the function would be defined as follows:

$$\overrightarrow{\omega(d)} = [\omega_1, \omega_2, \ldots, \omega_b, \ldots, \omega_n] \tag{3.44}$$

where the subscripts of the components of the ω vector represent the corresponding bin ,b, (in this case there are n bins), hence they define a bin vector, $\overrightarrow{B}$, where its components correspond to the interval of base pair distances.

$$\begin{aligned} \overrightarrow{B} &= [B_1, B_2, \ldots, b, \ldots, B_n] \\ &= [(0-10), (11-20), \ldots, (41-50), \ldots, (x-L)] \end{aligned} \tag{3.45}$$

Here bin b, is all interactions where two bound sites are separated by 41-50 basepairs, and x represents the last bin border, and L represents the Length of the enhancer (sequence).

For repressors (like Snail) that have 'quenching' interaction, ω is in the range $[.01, 1]$, while for binding cooperativity ω is in the range $[1, 100]$.

Lastly, $\overrightarrow{w}$, contains a post binding constant for each transcription factor. These parameters represent the interaction of the bound transcription factor with the BTA. Each transcription factor has an associated w_{tf} factor, for example, see figure. These factors, in a sense, represent the domain of the transcripion factor that interacts with the BTA. However, this interpretation is a controversial; as transcription factors recruit histone modifiers and remodelers, which change the epigenetic state of the chromatin, and in this way they affect transcription, and hence BTA binding, so the interaction isn't necessarily due to the protein 'touching' or binding to the BTA.

3.4 Annotation model of binding sites

3.4.1 Discovering the binding sites within the CRM

The CRM sequence space of all possible CRMs of length 500 is of dimension 4^{500}. Segal only had about 40 sequences available to train his model, which seems small with respect to a possibly *ideal* data set that would contain all 4^{500} CRM sequences **S** of length 500 along with each CRM's response **E** at each position along the DV axis. Of course, most of those **S** sequences do not respond to Dorsal Twist and Snail morphogens, and hence, a MSE (Mean Square Error) objective function under an ideal data set would be overwhelmed by Negative data, visible by the 'reduced Chi square', which is the SE (Squared Error) divided by the number of degrees of freedom $4^{500} * m - p$ (where m is the number of positions along the axis, and p is the number of parameters to be fit). In reality, what I have called an *ideal* data set is NOT ideal at all. Nor is a random sampling of the 4^{500} sequences, since the prevalence of functional (CRMS that form a usable pattern by the organism) is very rare. Hence, an approach that concentrates on known Positive CRMs seems to make sense (endogenous or

human designed).

Experts in CRMs have spent decades trying to decipher what sites are functional (adaptations). This has lead to additional functional sites and functional CRMs that are not naturally evolved but arguably just as useful for fitting the biochemical parameters (human designed CRMs or binding sites that work to recapitulate the work of evolution- such as recapitulating the expression pattern etc..). The result of this tedious work, is that the Segal idea of using all possible positions within a CRM as a possible site to plant the protein is simply not biological. Transcription factors have evolved to recognize a few specific sites within CRMs. Hence, a 'site' based approach is more biological. Even Segal's approach, really did not use all possible positions as a site, as mathematically one can set a threshold on PWM energy scores due to high energy sites having negligible effects on the model[12].

Hence, before we fit the parameters of the nonlinear model, we will explain a novel annotation model, an algorithm, to transform from the CRM sequence to a binding site space (much like Xin's configuration space) of much smaller dimension than the sequence space. This algorithm will actually use the model parameters, however, a full analysis of model fitting and parameter uncertainties, requires refitting the parameters, which we describe after our annotation algorithm.

Traditionally 'annotation' of binding sites is accomplished by 'scanning' PWMs over the CRM, setting a threshold, and calling a positive site any site below the threshold. Of course, this algorithm begs the question of how to set the threshold on the PWM. Cutting a true functional site from the list of sites to be used for modeling (the sites used in the configuration vector) would cause severe overfitting, or cutting a site that is a known Positive (at least

[12]Segal noted in his Supplement that he did not use all possible positions as a site for all possible morphogens. Furthermore, in his Supplement he pointed out that his computations of the configurations ultimately relied on MCMC (Markov Chain Monte Carlo) sampling of configurations more than the HMM recursion algorithm to compute the weight of a configuration.

'known' by evolution as a Positive). Furthermore, PWMs are notorious for false positives. Very weak sites are not a concern for our nonlinear model (since the boltzmann factor of a very weak site will will exclude that configuration from making a detectable contribution to the partition function). However, 'weak' sites that are at the border of being functional or non-functional can have enormous effects. A weak site that cooperates with a strong site may have functional interaction, functional cooperativity or quenching, but a conservative threshold may cut the weak site, which in turn will cause the cooperativity to be missed in the configuration space, which in turn will cause the fit to compensate for the cooperativity by tuning other parameters. Similar observations were made through sensitivity analysis of Dresch [31].

3.4.2 Annotation Model of Binding Sites without a PWM threshold

Standard bioinformatic annotation of the binding sites uses a threshold on the PWWs energy, here we define an alternative aproach for identifying binding sites within the CRMs. Instead of using a threshold for the PWM energies of all the possible sites within a CRM we use a threshold of the CRM's response, its gene's expression, to set a minimal constraint on the occupancy of Dorsal transcription factor at the position along the DV axis where the gene is first turned 'on'. Hence, the annotation model assumes Dorsal occupancy must reach a critical value before a gene switches 'on'. This assumption is analogous to the assumption that the mRNA counts are proportional to the BTA occupancy, here we are just pushing the assumption closer to an occupation that we can explicitly compute, in the sense that the BTA occupancy is caused by Dorsal occupancy, but we don't use basal promoters (like TATA

boxes) in our sequence analysis and hence can't compute their occupancy. Furthermore, BTA occupancy, in terms of differet occupancy across the genome is a function primarily of CRMs, as all genes have TATA boxes (TATA boxes do not differentially regulate BTA occupancy).

The occupancy of Dorsal is entangled with the parameters that we aim to fit, hence this annotation step requires setting these parameters to a specific value (all parameters must be set, since they are utilized in this annotation algorithm). For example, if a Dorsal site is adjacent to a Twist site, and Twist is bound at this site, and if the Dorsal site is within the range of influence of the cooperativity parameter between Twist and Dorsal$\omega_{Dl,Tw}(d)$, where d is the distance separating Dorsal and Twist, then the odds of Dorsal being bound will increase. In fact, as pointed out by A. Hill, if the cooperativity is strong enough, then the only configuration vectors that will contribute to the partition function are those that contain these bound Dorsal and Twist sites.

The annotation model assumes a conserved quantity governs all the enhancers at the point along the DV axis where their target gene becomes activated. Based on previous literature it seems Dorsal alone is sufficient and necessary, while Twist is neither. Hence, the assumption is that Dorsal's occupancy within the enhancer must reach a critical value, a conserved value, after which the gene expresses.

Only two points along the DV axis are used for the annotation model. The first point of activation when veiwed in the direction from Dorsal to Ventral represents one point. For example neurectoderm genes, the expression profile E is searched or scanned starting with the dorsal-most position for differential expression that is greater than 0.5, which occurs at the dorsal border of the neurectoderm genes. For genes that are only activated in the mesoderm, this search will similarly find the gene is activated at the mesoderm-neuroectoderm border, or at least the dorsal border of the mesoderm gene's expression. For gene's only expressed

in the mesoderm, this is the only point which is used to estimate Dorsal occupancy within the target CRM (a mesoderm CRM).

The second point of activation, is actually a point of repression of the gene, where Snail turns off the expression of the neurectoderm genes. Hence, the second point, for neuroectoderm genes is the Snail border, which demarcates the ventral border of neurectoderm expressed genes.

I assume Snail is not present in the neuroectoderm. This gives a classification scheme, which is used to divide the enhancers into two categories (mesoderm, neuroectoderm). Hence the first step, is to scan the expression profile, starting from the dorsal ectoderm and find the position where the gene is activated. If activation is first detected in mesoderm, then an algorithm to calculate Dorsal occupancy is used which allows for snail binding.

However, neuroectoderm classified genes are solely annotated with Dorsal and Twist sites. Here annotation means sufficient Dorsal and Twist sites are annotated such that the Dorsal occupancy reaches a value of one unit (which is like one Dorsal bound, but this could be, for example, because there are 5 Dorsal sites along with 2 Twist sites all working together such that the average occupancy of Dorsal in the promoter is 1 unit).

If the neuroectorder expressed genes indicate repression by Snail in the mesoderm, we assume that the mesoderm neuroectoderm border is a zero sum game, that is, enough Snail sites must be found to cancel the gene activation based on the activator sites that were annotated in the neuroectoderm. Hence the two points used for the expression profiles are usually the 'dorsal ectoderm - neuroectoderm border' that is determined by a search of the target gene's profile, and the other point is the 'mesoderm-neuroectoderm' border that is determined by the border of Snail's profile (i.e. the mesoderm is defined by Snail expression).

Once a gene is classified as a neuroectoderm expressor (i.e. it shows expression in the

neuroectoderm above a fixed value I've set at .5), then the occupancy of Dorsal for the gene's CRM must be 'one'. The location that the gene first achieves expression of 0.5 is determined by a profile search starting from the Dorsal ectoderm. Once that position along the profile is found, then we extract the vector of transcription factor concentrations at that exact same point, which are needed for the computation of the occupancy. Hence, given the network parameter's values, and the concentrations of the transcription factors where the gene 'swithches' on (its expression is 0.5), I then 'select' the best Dorsal site from the annotated list of Dorsal's PWM scores for the given CRM. This selected site's occupancy is calculated, and if it is below the value one, I select the next best Dorsal site. If its occupancy is above or equal to one, then I stop searching for Dorsal sites in the CRM UNLESS there happens to be Dorsal sites with the exact same energy, in which case these sites are annotated too. This is because a number of the Jiang and Szymanski constructs (CRMs in the training data) had multiple replicates of Dorsal sites in the same enhancer.

In the case that one Dorsal site is annotated and the CRM's occupancy of Dorsal is below a value of one, then the best Twist site is annotated from the list of *occupancy* scores of Twist in the CRM (note I did not say PWM scores, which are agnostic to the free parameters, while occupancy is highly sensitive to the free parameters such as Dorsal-Twist cooperativity). With the newly annotated Twist site, Dorsal's occupancy is recalculated, and checked to see if it is above a value of one. If the Dorsal occupancy is still below the critical occupancy, then the list of PWM scores is processed such that the original Dorsal annotated site is masked out along with all overlapping sites. Then the new list (i.e. the list post-masking) of Dorsal PWM scores is transformed to a list of occupancy scores, where each occupancy is calculated with the original Dorsal and Twist annotated site along with the Dorsal site from the list of PWM scores (these scores can be thought of as a data structure

that contains the energy and coordinate of the binding site in the CRM). The Dorsal site that achieves the highest occupancy is then selected as the next annotated Dorsal site. If the occupancy achieves a value of one unit then the annotation process halts. If the Dorsal occupancy does not achieve a value of one unit, then these steps are repeated until the the occupancy reaches the critical value, or the CRMs are 'filled' or 'covered' with binding sites (where no overlaps are allowed), the CRM is literally jam packed at this point.

Once the sites have been annotated, then the 'sites' are stored in a vector. This process is repeated for all neuroectoderm expressed genes in the network (all CRMs). Once all the neuroectoder CRMs have been annotated at their specific 'switch' points, the vectors of sites for each CRM is passed to another annotator that builds a list of Snail sites at the mesoderm-neuroectoderm border of the DV axis, where the snail site's are to 'repress' the gene (if the gene shows repression in the mesoderm). This annotator first selects the best scored Snail site from the list of PWM scores of the Snail PWM, and then calculates the target gene's expression. If the expression reaches below 0.5 then the annotator halts. If the expression is above 0.5, then the the annotated Snail site is masked from the list of Snail PWM scores, and the next best Snail site is selected. This annotation process is continued until the critical expression is reached, at which point the annotator halts.

Lastly, for genes not expressed in the neuroectoderm, such as *twist*, the annotation process is similar to the neuroectoderm genes with the exception that Snail is allowed to be annotated for functional sites.

We do not know *a priori* what the 'true' parameter values are, hence we start with a best guess point in parameter space, and then optimize the algorithm by gradient descent

using the objective function:

$$\sum_{i}^{n}\sum_{j(i)}[(\ <n_{Dorsal}^{i,j(i)}>-1\ )\ ]^2 \tag{3.46}$$

Here i is the target gene label and since we are only using 2 of the m points along the profiles, j represents those two points or positions (which are target gene dependent). Hereafter, we call this annotation method or model the Maximum Parsimony Annotator, MPA, since, in a sense, it is aiming to predict the minimal set of binding sites that are consistent with the nonlinear regression model -BTA occupancy and morphogen occupancy- model parameters.

3.5 Model fitting

Once we have annotated each CRM in our data with the Dorsal, Twist and Snail binding sites we wish to fit (refit in some sense) our nonlinear model Equation 3.27 and equivalently Equation 3.42, using standard nonlinear regression.

A simple nonlinear model, is 'logistic regression', which is used frequently for classification and is of the form:

$$Y = f(X_1, X_2, X_3, ..X_N) = \frac{1}{1+\exp{(\beta * \mathbf{X})}}, \tag{3.47}$$

This model's objective function's surface in parameter space has a unique optimum due to the objective function (the cross entropy) being convex. However, logistic regression requires boolean response variables, while our response is of a continuous form, ideally suited for regression.

To fit any regression model, one has a table of data, where the first column (for example) contains the values of the responses y (dependent variable Y) from the measurements, and

the next N columns contain the values of the explanatory variable x (independent variables $X_1, X_2..$). Given this data, one can then adjust the free parameters β to 'fit' the model to the data.

Here we will assume there is random error in our expression profiles (albeit small, such that the **trans** environment between embryos is almost the same). Hence the deviations of our model from each data point $(y_i, X_1(i), X_2(i), ..)$ will take the form of a random variable, called the error:

$$f(\mathbf{X(i)}|\beta) - y_i = \epsilon_i, \tag{3.48}$$

which we assume is normally distributed with zero mean. Hence, the probability of the entire data set is a multivariate normal:

$$P(\mathbf{D}|\beta) = \frac{1}{2\pi|\Sigma^{-1}|} \exp \frac{1}{2}(\mathbf{y} - \mathbf{f}(\beta)^T)\Sigma^{-1}(\mathbf{y} - \mathbf{f}(\beta)) \tag{3.49}$$

Here Σ^{-1} is the data's covariance matrix. We assume the errors are independent with unit variance, hence, the covariance matrix is simply the identity matrix. In this picture, each data point i occurs with probability $\frac{\exp -\epsilon_i^2}{2\pi}$, and the multivariate normal can be factorized as a product of independent univariate gaussians, which can be written as $P(\mathbf{D}|\beta) = exp{-}1/2\chi^2$. Hence maximizing the probability of the data given the parameters (maximum likelihood principle) is equivalent to minimizing the squared errors from the multivariate distribution, which we call 'chi squared', denoted as χ^2 (that's not a squaring operation!). In this picture we see the ith row of the data matrix contains the information needed for the ith element of

χ^2, where the model parameters are fit by minimizing the squared errors, χ^2:

$$\chi^2 = \sum_i (y_i - f(X_1(i), X_2(i), X_3(i), ..X_N(i)|\beta))^2 \tag{3.50}$$

In general, χ^2 is not a convex function, and therefore is susceptible to getting stuck in local minimas. Hence, nonlinear regression is a bit of an art[87], unlike its linear counterpart (linear regression), which has a global optimum at $\beta = (XX^T)^{-1}X^T b$, where X is the 'design matrix' and b is the .

Both Segal and He's method of optimization of their thermodynamic models used gradient descent and simplex, where the best fit parameters of gradient descent were passed to simplex method, where its best fit parameters were passed back to gradient descent, until a fixed set of alternations between the algorithms were exhausted. This was repeated for different starting points in parameter space, (by randomly selecting a point in parameter space), which was an attempt to evade getting stuck in local minima. We have implemented the Levenberg Marquardt algorithm in GSL (Gnu Scientific Library) which is similar to gradient descent, upon completion of the alternations between gradient descent and simplex, the best parameter vector is passed to the Levenberg Marquardt algorithm (this was purely for the advantage that GSL has implemented a number of routines for parameter error estimates using Levenberg Marquardt not found in their general 'optimization' algorithms).

3.5.1 Covariance matrix of fitted parameters

In linear regression, with no estimate on the error in the data, a standard estimate of the error in the fitted parameters is simply the χ^2 divided by the degrees of freedom (the number of data points minus the number of fit parameters), or more conservatively just χ^2. In linear

regression with known errors (the covariance matrix of the data, Σ), one is better able to estimate the error in the fit parameters by $(A^T \Sigma A)^{-1}$, where A is the design matrix.

In nonlinear regression, with no estimate on the error in the data, again an estimate of the error in the fitted parameters is simply the χ^2 at the minima. Another estimate is $(J^T J)^{-1}$, and an even better estimate is to invert 1/2 the Hessian matrix of the χ^2 over the parameters. To better understand our model's fitting behavior we will briefly discuss the derivation of the covariance matrix of fitted parameters.

The estimate of the covariance matrix of the best estimated parameters is based on the following Taylor expansion of the χ^2 about the best fit point in parameter space:

$$1/2\chi^2 = 1/2\chi_0^2 + 1/2(\beta - \beta_0)^T \frac{\partial^2 \chi^2}{\partial\beta\partial\beta}(\beta - \beta_0) + ... \tag{3.51}$$

where the Hessian matrix, denoted as $\frac{\partial^2 \chi^2}{\partial\beta\partial\beta}$ is a 'two form' (*i.e.* The i j element of the Hessian matrix is: $\frac{\partial^2 \chi^2}{\partial\beta_i \partial\beta_j}$). In the expansion the first derivatives of the chi square with respect to each parameter are zero (we're at a minima of the chi square surface). The first term in the expansion is simply the χ^2 value at the minima. As we obtain more data, the higher order terms will go to zero (assuming $(\beta - \beta_0)$ is less than one, then using Taylor's theorem, we know $(\beta - \beta_0)$ approach zero faster than the derivatives). Using a Bayesian argument, we can now show that at the minima of the χ^2, our best estimate of the free parameters are gaussian distributed. In a Bayesian setting, we wish to estimate the distribution of parameters (not a just a point in parameter space). Hence, using Bayes theorem we have:

$$P(\beta|\mathbf{D}) \propto P(\mathbf{D}|\beta)P(\beta). \tag{3.52}$$

In this picture we still wish to optimize the likelihood, however we now have the priors

over the parameters to deal with. Assuming the parameter priors are uniformly distributed (uninformative) then maximizing the likelihood $P(\mathbf{D}|\beta)$ is equivalent to maximizing the posterior of the parameters given the data (the priors simply don't play a role). Hence we see the maximum likelihood estimates (the minimum of the chi^2) is equivalent to inferring the posterior distribution over parameters. Hence, we have:

$$P(\beta|\mathbf{D}) \propto P(\mathbf{D}|\beta) \propto P(1/2\chi^2) \propto P(1/2(\beta - \beta_0)^T \frac{\partial^2 \chi^2}{\partial\beta\partial\beta}(\beta - \beta_0)). \tag{3.53}$$

where in the last expression we have replaced the χ^2 with its Taylor expansion. Upon plugging in the Taylor expansion we see that our posterior distribution has the form of a multivariate Gaussian distribution with mean β_0 and covariance matrix $\frac{\partial^2 \chi^2}{\partial\beta\partial\beta}^{-1}$, which we will denote as Σ_β. Hence the standard errors of the parameters are estimated simply as the square root of the diagonal elements of the inverted matrix of (1/2 the Hessian matrix), where the Hessian was evaluated at the fitted parameters (see for similar analysis the 'Laplace Approximation' on page 213 of Bishop[17], and Press' model fitting section[87], and the text by Beck[11]).

What if Hessian is not full rank? Then the Hessian can not be inverted. This can be traced to either poor experimental design (e.g. too little data, which usually can't be helped), or to poor model design (which can be improved).

Non invertible Hessians and non full column rank Jacobians have influenced my estimation of parameters. Hence, I will briefly discuss these issues, in order to better determine what parameters will be fit.

3.5.2 The overdetermined and underdetermined problem

In linear regression one simply needs to ensure their design matrix is full column rank in order to fit their data. For a two parameter linear model f = Y=mX +b, where one has (X,Y) paired data, for example: (5,10) (10 ,15) (1,5). One has a Design matrix of the following form:

$$A = \begin{pmatrix} 5 & 10 \\ 10 & 15 \\ 1 & 5 \end{pmatrix}$$

Here the design matrix represents two 3 dimensional vectors (two vectors that live in the column space) do these vectors span the column space? If the two vectors do not span the column space, then this design matrix results in an underdetermined problem (not enough *independent* equations to solve for the unknowns). Oddly, this matrix in the context of *systems of linear equations* is also a so-called overdetermined problem (more equations than unknowns). Hence, the problem is simultaneously overdetermined and underdetermined. A quick test for this type of scenario (in statistics) is simply to check if det $A^T A$ is zero, where det is the determinant.

The nonlinear regression follows the linear problem in almost every detail. The design matrix is generalized to the 'Jacobian' matrix J, where the i j element of J is the partial derivative $\partial f(X_i)\beta_j$ where f is our nonlinear model evaluated at data point i and the derivative of f is taken with respect to the jth parameter. A quick test to see if the Jacobian is full column rank is to check if det $J^T J$ is zero.

An example of non full column rank Jacobian, or a singular $J^T J$ matrix is the following model $f = \exp((\beta_1 + \beta_2)X)$[11]. Regardless of how much data one collects for (Y,X), the chi square surface at the minima is a 'trough' over the two dimensional plane of β_1 and β_2.

This trough is parametrized by the equation $\beta_1 + \beta_2 = \beta$, which is an infinite line through the β_1, β_2 plane (it is a one dimensional subspace, the null space of the transpose of our Jacobian matrix). Hence one can at best estimate one parameter (not two). In this example it is clear the problem with a singular Hessian, or not full column rank Jacobian, the problem is that the optimal solution (minimum chi square) is not a point, it's an infinite line or infinite surface or hyper surface in parameter space, depending on the number of dependent columns in the Jacobian matrix. This is the consequence of a Jacobian matrix that is not full column rank. This means some of the parameters are dependent, hence some of the columns of the Jacobian matrix are linear combinations of other columns of the matrix. Hence the column space of the Jacobian spans a subspace of a dimension smaller than the number of desired parameters, which can not be remedied by more data in the case of poor model design.

For example imagine one knows with certainty two independent data points $(y_1, x_1), (y_2, x_2)$ for this model, constructing a linear model for the above equation one arrives at the following system of equations $\begin{pmatrix} x_1 & x_1 \\ x_2 & x_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \ln y_1 \\ \ln y_2 \end{pmatrix}$, which is just a line in the β_1, β_2 plane (the parameter vector space), where a unit vector along this line is: $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, we add another free parameter α to denote any point along this line, hence α parametrizes the null space of the Jacobian transpose). This can be seen analytically since the analytic Jacobian is:

$$\frac{\partial f(x, \beta_1 + \beta_2)}{\partial \beta_1} = x \exp^{(\beta 1 + \beta 2)x} \tag{3.54}$$

$$\frac{\partial f(x, \beta_1 + \beta_2)}{\partial \beta_2} = x \exp^{(\beta 1 + \beta 2)x} \tag{3.55}$$

These two equations *should* span parameter space, but the two equations are not indepen-

dent, they are identical, hence they (or one of them (which ever one you like)) spans a one dimensional line in parameter space.

3.6 Results and Discussion

3.6.1 Best fit of parameters for data from Section3.2.1, Experiment 1

We fit the data from 3.2.1 using the model from Equation 3.42, where the χ^2 was defined as:

$$\chi^2 = \sum_z \sum_t (f(S_t, z) - E_t(z))^2, \tag{3.56}$$

where $f(S_t, z)$ is equation 3.42, where S_t is a CRM from our data set, and z is a position along the DV axis, and $E_t(z)$ is the the target expression driven by the CRM's corresponding gene denoted by t at position z along the DV axis.

The annotation model MPA was originally fit to a best set of parameters, and those parameters were then used as the initial point in parameter space for the nonlinear regression model 3.42, and for minimizing the objective function3.56. *HOWEVER*, due to the χ^2 surface of 3.46 being flat at the minima, we decided not to try and fit parameters using the MPA model. Rather we decided to just set $w_0 = 5$ and set all other model parameters to 'one', the default values of the nonlinear regression model parameters from Equation 3.43 and allowed the annotation model MPA to make predictions of binding sites within the CRMS using these default parameters. *Given*, the MPA annotated CRMs, we then were able to fit the nonlinear regression model3.42, where the fitted profiles are in Figure 3.7.

In the profiles in Figure 3.7 we set the model parameter $w_0 = 5$ (as in Ay's model) and we

estimated the following parameters: $\omega_{Dl,Tw}(d_1) = 5 \pm 2, \omega_{Dl,Tw}(d_2) = 64 \pm 37, \omega_{Dl,Sn}(d_1) = .1 \pm 25, \omega_{Dl,Sn}(d_2) = .7 \pm 12, w_{Dl} = 7 \pm .05, w_{Sn} = 39 \pm 687$, where d_1 represents the spacer bin of $[0, 30]bp$, and d_2 represents the spacer bin of $[30, 60]bp$. The errors were estimated as the square root of the diagonal of $(J^T J)^{-1}$. The $\chi^2 = 41$, where we had 1200 data points (each position along the z axis for each gene). The Hessian, had a high condition number (10^4) where the largest eignenvalue was 1.4, and the two smallest eigenvalues were .0002 and .001, suggesting that χ^2 surface was close to flat along those eigenvector directions of parameter space.

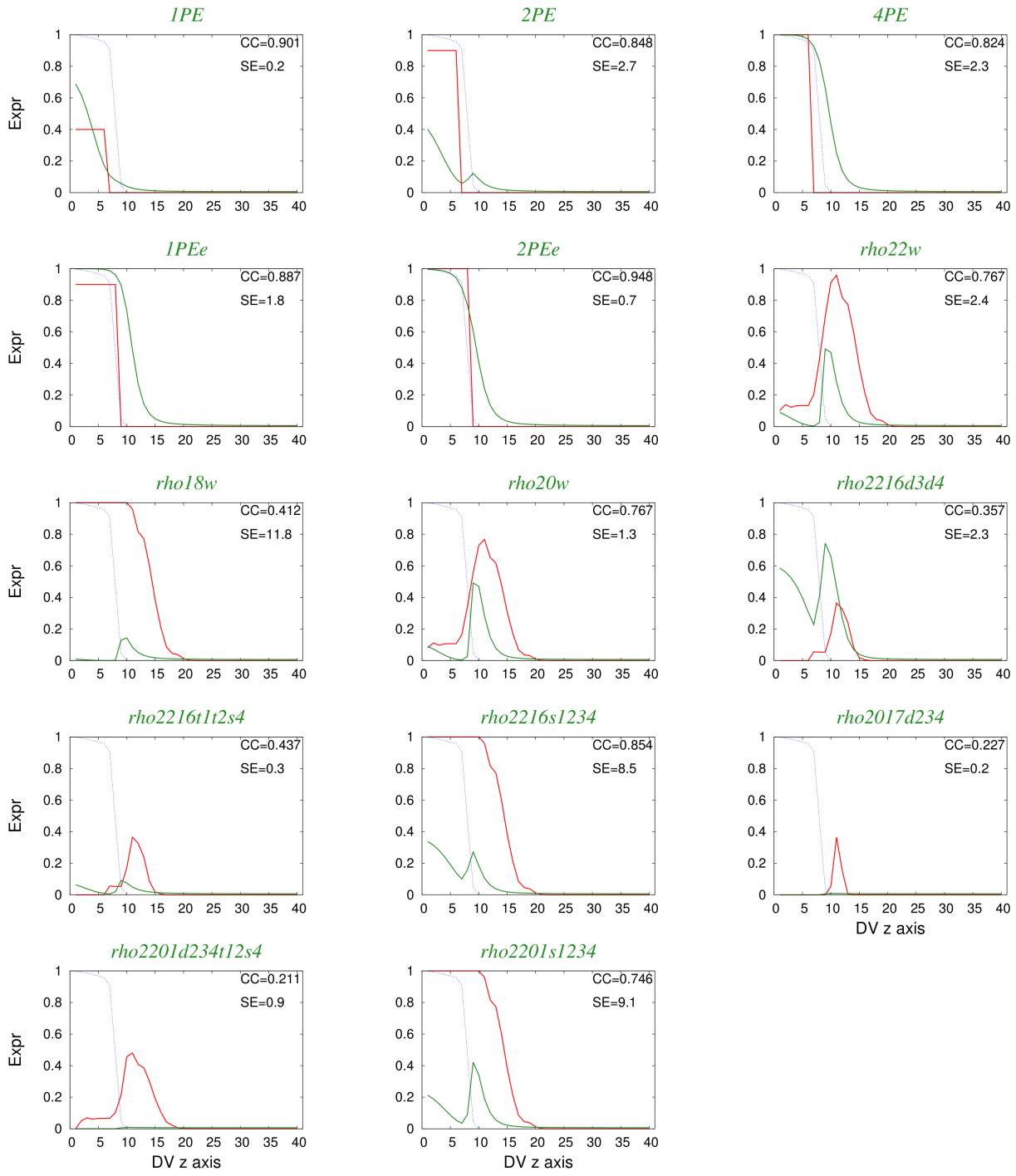

Figure 3.5: the legend is in the upper right corner of the table, denoting the Observed profiles ($E_t(z)$) as red, the model predictions as green along with the header above each figure denoting the CRM (gene target) in green, and the Dorsal morphogen profile ($E_{Dl}(z)$) as dotted blue curve. The correlation coefficient between the observed pattern and the predicted pattern is denoted as CC for each gene (which is at most 'one'), and also the squared error between the observed pattern and predicted pattern is denoted as SE for each gene (where each gene had 40 positions, z, along the DV axis (i.e. SE is at most 40). The Snail profile is uniform from positions 0 to 8, where it is 'on' (at a value of 'one') and Snail is off from positions 9 to 40 along the axis, and the Twist gradient (profile) was replicated as the Dorsal gradient.

3.6.2 Redesigning the parameters to be fit

The results above suggest poor nonlinear regression model design or insufficient data, where the free parameters to be fit from 3.43 were based off Segal's thermodynamic model, which was similarly used by Xin in GEMSTAT and Fukhouri et.al.[36]. In general, the more parameters we add to the nonlinear regression model will either decrease the χ^2 statistic (albeit, the reduced chi-squared that accounts for the degrees of freedom may increase) or it will maintain the χ^2 statistic at a particular value (*i.e.* the statistic becomes insensitive to additional parameters); but the χ^2 will not increase. This assumes one's fitting algorithm is allowed to 'find' the point along the newly added parameter's line in parameter space that improves the fit or at least does not spoil the smaller parameter set's fit. Spoiling a previous fit with a newly added parameter can be avoided, for example, by simply setting the new parameter to a value where the model is insensitive (and hence χ^2 will not be disturbed). For model's with extensive dependencies between parameters this may not be possible. However, by inspection, the parameters of our nonlinear model can all be set to values such that they have no contribution to the χ^2. For example, all the parameters in our nonlinear model, when set to numerically one, have the effect of not influencing the form of the model, in a sense, this is the parameter free form of the model. I will call these values of the parameters the default values (*i.e.* the default values have the odd property that it appears we are not fitting the parameters at all; and, in another sense, one could imagine fitting all the parameters and all the parameters being fit to the default values.). While setting the values to be zero have the effect of 'knocking out' all binding sites of a factor (in the case of α) or 'knocking out' pairs of interacting binding sites (in the case of ω)).

Fitting subsets of the parameters in Eq.3.43 (such as a parameter subset without our

novel pairwise potential ω) leads to good fits, by eye, of the data, such as RMSE's of .05. However, upon implementation of a Hessian matrix using a two point finite difference, we found that the Hessian matrix was not full rank for the parameter set even without the pairwise interactions ω, our novel form of the potential. Without pairwise interactions, one still has six free parameters to fit for the CRM's responses to Dorsal Twist and Snail morphogens. Based on expert knowledge of the DV network, certain choices of two of the parameters were selected to fit the data, which occasionally lead to a Hessian with a condition number of about 100 (the condition number is a numerical technique to determine if a matrix is singular[13]), while many of the choices of parameters lead to Hessians with condition numbers on the order of 10^4. Furthermore, the Hessian should be positive definite for a unique solution, while we found the eigenvalues were frequently mixed in sign (indicating saddle points) - of course the gradient (the derivative of the χ^2 with respect to each parameter, a different object from the Jacobian) was zero (on the order of $10^{-}10$ for each component) at all of our estimates of the minima of the χ^2 surface[14].

In order to better estimate the errors of the parameters for our model we aimed to obtain a full column rank Hessian (regardless of whether the Hessian was positive definite). We analysed the analytic derivative of our nonlinear model with respect to each parameter β_i, and checked if other parameters occur in the result (indicating dependencies between parameters). If possible, parameters that do depend on one another were grouped to form

[13]In a computer, in a numerical representation, a matrix is rarely actually singular, it just has some eigenvalues that are much smaller than the largest eigenvalue of the matrix. The number of eigenvalues of Hessian is the number of parameters to be fit, and if some of these eigenvalues are close to zero, this suggest a zero determinant Hessian.

[14]These results lead to the implementation of the Levenberg Marquardt that has a number of GSL built in functions to help diagnose pathological nonlinear models (unlike the gradient descent and simplex implementations). If Levenberg Marquardt revealed the same types of errors in our parameter estimates then we could be more confident that there was not an error in our implementation of the Hessian. (i.e. a zero determinant Hessian, means our parameter error bars are infinite in extent - unless one decides to use a different method to estimate their error bars).

just one parameter[11]. Furthermore, certain regions of the Data space (such as certain positions in the embryo (like the dorsal ectoderm, where none of our genes are active) may be very insensitive to our parameters (causing small values of the Jacobian for these data points, however this does not appear to affect the Hessian of the χ^2, since the analytic Hessian, $\mathcal{H}$ of the χ^2 is exactly equal to $J^T J + G$, where G is a matrix of second derivatives over the model[15].

3.6.3 Analytic Jacobian

A simple function is $f(z) = 1/(1+\exp(-\beta z))$, where β is a constant, where the domain and range of z and f are: $z \in [-\inf, \inf]$ $f(z) \in [0, 1]$. Now if we imagine that β was a free parameter, then we could ask how sensitive this function is to variations of β as a function of particular positions along the domain of z. Hence the derivative is:

$$\frac{\partial f(\beta, z)}{\partial \beta} = \frac{\beta \exp(-\beta z)}{(1+\exp(-\beta z))} \approx \beta f(z,\beta)(1 - f(z,\beta)) \tag{3.57}$$

The above equation, in a sense, is an analytic representation of the Jacobian matrix elements for the nonlinear model (e.g. BTA occupancy) as a function of data (all possible data, which is represented by the domain of z).

We do not need to analyze our more complicated BTA occupancy function to understand

[15]The analytic Hesssian in exact form is $\mathcal{H} = \frac{\partial^2 \chi^2}{\partial\beta\partial\beta} = J^T J + \frac{\partial^2 f}{\partial\beta\partial\beta} = J^T J + G$, where G is matrix with second derivative elements of the nonlinear model f (the fractional occupancy of the BTA) over the model parameters (for example, see equation A.4c page 482 of Beck[11]). In the case that G's Frobenius norm is nearly zero (the elements of G are nearly zero), then we have $\mathcal{H} = J^T J$, which is why with certain data sets and experiments, it is possible to estimate the covariance matrix of parameters based on Jacobian alone. $J^T J$ for the case of one free parameter is just a dot product (i.e. the Jacobian is just a vector of size n x 1, where n is the number of data points), which will be insensitive to data points that were uninformative for parameter estimation (they're just adding a term zero to the dot product). Hence, it seems reasonable that uninformative data (data where the model is insensitive, such as Dorsal ectoderm regions of the embryo, where none of our CRMs are active) will not cause harm to model fitting.

the behavior of the Jacobian matrix elements, rather we can simply use the above equation as a phenomenological parametrization of our model of CRM response to morphogen gradients (which are encoded in the spatial position z). For example, imagine one wished to analyze phenomenologically the response of *rhomboid* expression pattern just in the positions of the embryo of the neuroectoderm and the dorsal ectoderm (in our binning of the DV axis this would be equivalent to the z interval of [9,40], where the interval [1,8] contains the mesoderm, which we are uninterested in for the moment). We can model the *rhomboid* expression as a function of DV axis using the above equation, where we set β to be negative since we must reflect the function about the 'switch' point where the gene is turned 'on (the point in space where *rho* turns 'on' is at the neuroectoderm-ectoderm border). [16] We must also offset the position where the logistic function reaches 1/2 max (which in the above equation is position $z = 0$). Hence we mush add a constant to the argument of the exponential

$$f(z) = 1/(1 + \exp(\beta z - \beta_0))$$

, where $beta_0$ will define the 1/2 max of the logistic function when $z\beta = \beta_0$.

The analytic Jacobian for these two parameters (as a function of data z):

$$\frac{\partial f}{\partial \beta} = \frac{-z \exp(\beta z - \beta_0)}{(1 + \exp(\beta z - \beta_0))^2} = -zf(1 - f) \tag{3.58}$$

[16] If we modeled the 'switch' point where the gene is turned 'off' (the point in space where *rho* turns 'off' is at the mesoderm-neurectoderm border) we would not have to reflect our graph about the switch.

[17]

$$\frac{\partial f}{\partial \beta_0} = \frac{\exp(\beta z - \beta_0)}{(1+\exp(\beta z - \beta_0))^2} = f(1-f) \tag{3.59}$$

[18]

Now we can see some simple relations. First of all $J^T J$ =0 if one only collected data in the dorsal ectoderm region of the embryo (where the logistic is nearly '1', meaning we can not invert that matrix (so J is not full column rank). Similarly, if one just has Boolean data for the expression pattern of rho along the z axis, say 000000011111111000000000=$\boldsymbol{E}_r ho$, the only possible point where J is nonzero is at the borders where the gene switches from off to on or vice versa. Now imagine for an NEE gene that we have an expression pattern over space like 000000010101010000000000 where the switching behavior (10101010) is in the neurectoderm

[17] The logistic function when evaluated at a particular point z' can be thought of as a Bernoulli Distribution, hence, the last expression is the variance of the Bernoulli distribution, where the function $f(x')$ at each point x' happens to be the 'parameter' that describes a Bernoulli distribution (normally this parameter is denoted as 'p' for probability, where the mean of Bernoulli random variable is also p, and the variance of a Bernoulli random variable is p(1-p). This is effectively how the BTA occupancy is described along the DV axis where x' now denotes the position along the DV axis, and β is a free parameter.

[18] Do these two two equations span parameter space? Recall previously we analyzed $\exp(\beta_1 + \beta_2)z$ for its behavior in the context of nonlinear regression. Similarly for the logistic nonlinear model we would like to analyze $1/(\exp(\beta_0 + \beta z)$. For example, and if one knows with certainty two independent data points $(y_1, x_1), (y_2, x_2)$, then we can transform to a simple linear model for the above equation, where one arrives at the following system of equations $\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} = \begin{pmatrix} \ln \frac{y_1}{1-y_1} \\ \ln \frac{y_2}{1-y_2} \end{pmatrix}$. Does this span the β, β_0 plane (the parameter vector space)? The determinant of the design matrix is $x_2 - x_1$, which suggest as long as the experimental design was such that the data points x_1 and x_2 are not too close together we could actually 'fit' β_0 and β. However, one must be careful here, since data collected in regions of very large or very small x_1 and x_2, will cause the original response y to become close 'one' or 'zero' causing the logarithm's value to be very large, possibly causing numerical issues. This can be seen by symbolically solving for the eigenvalues of the design matrix, where we must solve for the roots of $(1-\lambda)(x_2 - \lambda) - x_1 = \det(X - \lambda I) = 0)$, where X is design matrix, and λ is the eigenvalues of the design matrix. This yields the quadratic $\lambda^2 + x_2\lambda + x_2 - x_1$. If λ is zero (which means we would have a singular design matrix) we see that indeed $x_2 = x_1$, which makes sense, we can't expect repeated measurements of the same position in space to tell us anything about the global behavior of the sigmoid (as defined by the β parameters). Solving the for the roots of the polynomial using the quadratic formula we have:

$$\lambda = -\frac{x_2 \pm \sqrt{x_2^2 - 4x_1}}{2}, \tag{3.60}$$

here we know constraints on the data due to the embryo size, where the DV portion of the embryo constrains $[x_1, x_2]$ to reside in the interval $[0, 40]$, since there are only 40 cells along that axis. Given the constraint on embryo size it seems we can at least say $\sqrt{4x_1} < x_2$.

region of the embryo, while the other two regions are the mesoderm and ectoderm. Using a mathematical metaphor, it is as if one has $J^T J_{meso} + J^T J_{neuro} + J^T J_{ecto} = J^T J = J^T J_{neuro}$, since the regions in data space of *meso*, *ecto* are insensitive to parameter variations, hence these regions in data space (the row space of J), act as labels to the null space of $J^T J$ that is a subspace of parameter space (the column space of J).

Analysis of binding energy effects on occupancy

For one binding site, we can compute the average occupancy, it is just the probability of the bound configuration $P(c = 1)$ where $c = 1$ is the configuration vector in the bound state (c takes on only two values 1 or 0).

$$\frac{\partial P(c)}{\partial E(S)} = P(c)(1 - P(c)) \tag{3.61}$$

This function's maximum is .25, which occurs at $P(c) = .5$. Hence the greatest effect of energy on the occupancy of one site is at half max occupancy. Assuming the differentials are standard deviations we have:

$$\sigma_{P(C)} = |P(c)(1 - P(c))|\sigma_{E(S)}. \tag{3.62}$$

Balanced Data sets of mesoderm and neurectoder enhancers, Experiment 2

The parameters of the thermodynamic model are all contingent on the free parameters that correspond to Snail (such as it α for its protein-DNA interaction, and its quenching strength, and its w_{SN} for Snail's strength of repressing BTA binding). Hence, to disentangle the

cooperativity of Twist and Dorsal (if it exist) from the other free parameters, an experiment was conducted with no Snail protein (no Snail in the network). Hence, the MPA simply annotates for Twist and Dorsal (where again the MPA was not 'fit' rather default parameters were set for annotation). All NEE modules are known to be repressed by Snail in the mesoderm. However, the cooperativity of Dorsal and Twist is known to be important when there are limiting concentrations of Dorsal in the neuroectoderm (where Snail protein does not exist). Furthermore, the modules that are not NEE's (such as mesoderm targets of Dorsal), are know to have no Snail binding sites in their module (at least this will be hypothesis). Here we used four CRMs to fit two parameter, the Dorsal binding strength and Dorsal-Twist cooperativity. We found a $\chi^2 = 8.9$ for 80 data points[19]. We removed the constraint on the parameter ranges, and set $w_0 = 5$, $w_{Dl} = -5$, $w_{Tw} = 0$. We set $w_{Dl} = 5$ because we wanted one unit of Dorsal occupancy to correspond to half max BTA occupancy (which means the target gene is at half max). We set $\alpha_{Tw} = 1$, the default value. We found $\alpha_{Dl} = 0.3 + / - 0.03$, and $\omega_{Dl,Tw} = 3980 + / - 982$, where the standard deviations were estimated from the square root of the corresponding parameter's diagonal element of $(J^T J)^{-1}$. The

These results are as expected from the modules shown. For example, the construct $6xdl$ was a Szymanski construct that was shown that even with 6 Dorsal binding sites, the CRM does not respond to Dorsal in the neuroectoderm (see profile in Figure). Hence, if these were zero energy Dorsal sites, then we know for what was defined as the minimum value of $\alpha_{min} = 1$ that Dorsal occupancy would reach half max at the position z and a concentration

[19]The goodness of fit can be tested here, with 78 degrees of freedom, we can assume our fit is a deviate from the χ^2 distribution. If the value of our squared error is a reasonable deviate from the χ^2 distribution then we can accept the fit. I computed the p-value as nearly one, suggesting the fit is good (the cdf of our the χ^2 distribution to the value we found was 10^{-28}.

$E_{Dl}(z)$, hence we have: $E_{Dl}(z)/(1+E_{Dl}(z)) =< n_{Dl} >= 1/2$, hence $E_{Dl}(z) = 1$. However, the CRM $6xdl$ had 6 Dorsal binding sites, hence one unit of occupancy could be reach at 1/6 this concentration (since the sites are all independent, by assumption).

Another construct that was used was the double knock out of Twist sites by Ip, which reported catastrophic loss of expression in the neuroectoderm, the construct labelled $rho2216t1t2s4a$. Hence the Ip construct along with the $6xdlPLZ$ acted as a control group. The treatment group (in a sense) were two constructs that are known to have Twist sites the rho CRM (from mel specie), and the vn CRM (from vir specie).

The Snail protein was not used in this model, , and the Twist gradient (profile) was replicated as the Dorsal gradient. The $6xtwPLZ$ was also left in the training set, but this has negligible effect (due to model assumptions (i.e. default parameter values of Twist), and the reported expression profile by Szymanski for a this construct was roughly zero (i.e. Twist binding sites alone are not sufficient)).

3.6.4 Robustness analysis, Experiment 3

We increased the Dorsal expression by .15 in each position along the DV axis to see if the annotated binding sites were the same when using the annotation model MPA. The data set was four NEE modules $rhomel, rhovir, vnmel, vnvir$.

The Twist gene's free parameters were all set to 'one' for α, and $\omega_{Tw,Dl}$ was set to 'one' for all bins except $B = [0, 20]bp$, and w_{Tw} parameter was set to zero. Furthermore, the Twist's morphogen concentration gradient $\boldsymbol{E_{Tw}}$ was set equal to Dorsal's gradient to assure their differential gradients were not influencing the result. Snail's quenching was set for $w_{Sn,Dl}$ for only one bin $B = [0, 50]bp$ (quenching means ω is in the range [0,1], while cooperativity means ω is in the range [1,100]. Hence, we only allowed the Dorsal and Snail parameters to

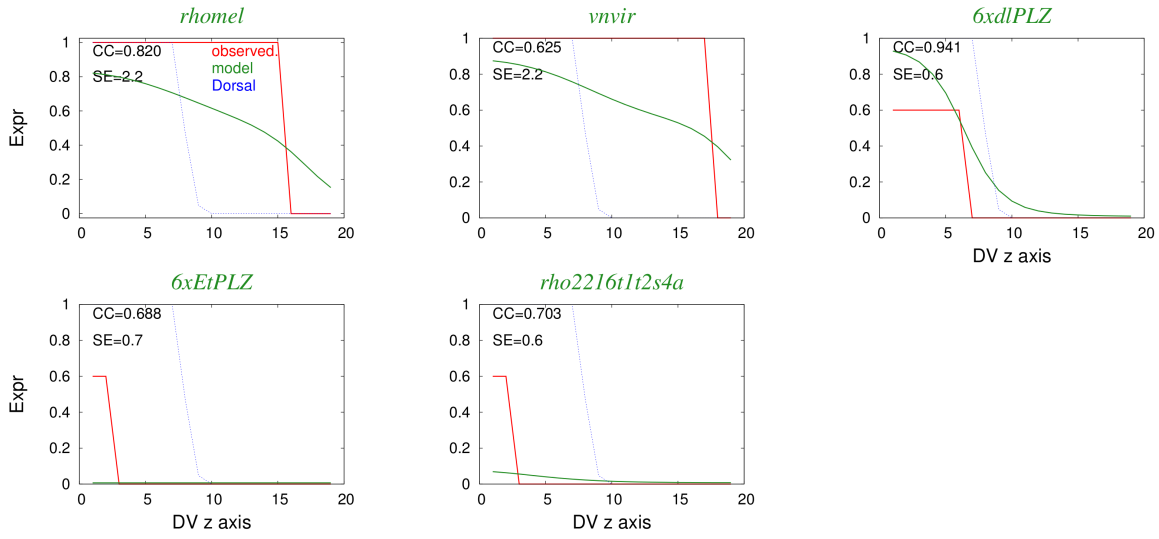

Figure 3.6: the legend is in the upper right corner of the table, denoting the Observed profiles ($E_t(z)$) as red, the model predictions as green along with the header above each figure denoting the CRM (gene target) in green, and the Dorsal morphogen profile ($E_{Dl}(z)$) as dotted blue curve. The correlation coefficient between the observed pattern and the predicted pattern is denoted as CC for each gene (which is at most 'one'), and also the squared error between the observed pattern and predicted pattern is denoted as SE for each gene. Each gene had 20 positions, z, along the DV axis, which as always (in DV literature), is plotted such that ventral is at the zero position.

aatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcg
aaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggcggaatttcctgattcgcgatgccat
gaggcactcgcatatgttgagcacatgttttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtggggaaaagc

ccacgtcctacccacgcccactcggttac

rhomel cell 6
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeteeeeeeeeeeeedeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeedeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeee
gtcaagggctgcacacacaccctacctgcgtgggaaaatgtccattgcgctctcgggctaagctgttttggttggcgggcacaggtaaca
ggcggtagggcaacgttgcaaccggcaaatgtggctgtgccacatgtttgtgccggaaattccccgtgcacgccttaccgcctgcctttt
tatgttgaaatttttcctatcacttggcgggtttttccgggccactggcggaaattccacaatgtcgccactggccacaggcagaaagt

gcaaccaaacgcgcttacaaaaattattataaaatgtgtattattaaatgccaaattgcgtccaatcatccgcagttctctgccggccggc
aagctgaccgtgtgctaaacaaaaatcaaaaaaaaaaaaaaacagccaaaattgccgctggctgctgcacaaacatgcgagcgcgcgcgc
gggctccaactggccactggggcaagctg

vnvir cell 6
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeteeeeeeeeeeeeedeeeeeeteeeeeeeeteeeeeeeteeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeedeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeee
atcctgggaaaacccgagatgatcctgggaaaacccgacctgggaaaacccgagatcctgggaaaacccgagatcctgggaaaacccgag
atcctgggaaaacccga

6xdlPLZ cell 6
eeeeeeedeeeeeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeedeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeeeeedeeeeeeeeee
eeeeeeedeeeeeeeeee
aaaaaaaaaagatccatatgagatccatatgagatccatatgagatccatatgagatccatatgagatccatatga

6xEtPLZ cell 6
eeeeeeeeeeeeeteeeeeeeeeeeteeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
agcttttcctctgctcaaaatcaaaatgattaaaacaacagtttgatacgaattttaattccccttttttgctgcggagtcagttaagtga
gtcgctttcaggactcagggcatcatccagatcgcacgatcccatttgcatctgccttctcagaagctgcttgaaagacgcgcccctgtc
ggatgattagtgctaagatccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaac
ttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcg
caagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgccaagcttgacgcgttgttttgggggaaattcccgggcgacgg
gccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactcggttacctgaattcgagctcgagtgttttgg
gtggctgagattgctttggtacggtggctgaccttgccagtgccagtgggtccatgtcc

rho2216t1t2s4a cell 6
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeedeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeedeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

Figure 3.7: The CRMs and predicting binding sites from MPA for default parameters on all proteins. Dorsal annotated blue, Twist green. The column d+ denotes added noise to Dorsal concentration profile (which was zero in this case, hence d+ should not be there). A bug in the printing code caused one of the *vnvir* sites to not appear in the CRM sequence highlighting.

be tuned by the fitting routine.

The results of annotation with the wild-type Dorsal expression (no perturbation of the profile) are in figure3.8. The results of the perturbed Drosal profile are in figure 3.9. The annotations are the identical. This possibly is an artifact of the way the Dorsal profile was perturbed (by simply adding .15 to each cell). Due to the sigmodal nature of the Dorsal profile this has little effect. A better designed experiment would shift the half max of the Dorsal profile by a fixed number of positions along the z axis.

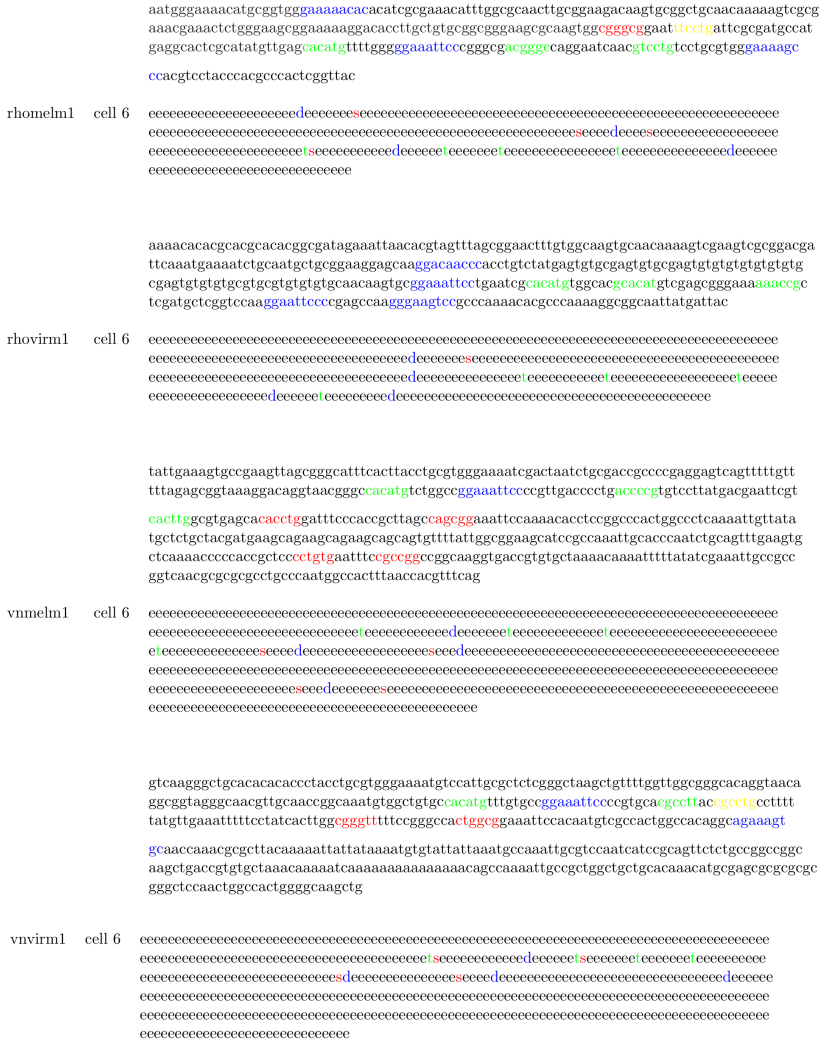

Figure 3.8: Here the target gene is denoted in the left column, and the cell along the DV axis is denoted in the second column. Twist site's are annotated green, Dorsal blue, Snail red, and brown denotes overlaps. The sites annotated at the mesoderm bottom border were used to annotate the sequence. For example, the first gene is *rhomel*, for *rhomboid* in the species *melanogaster*.

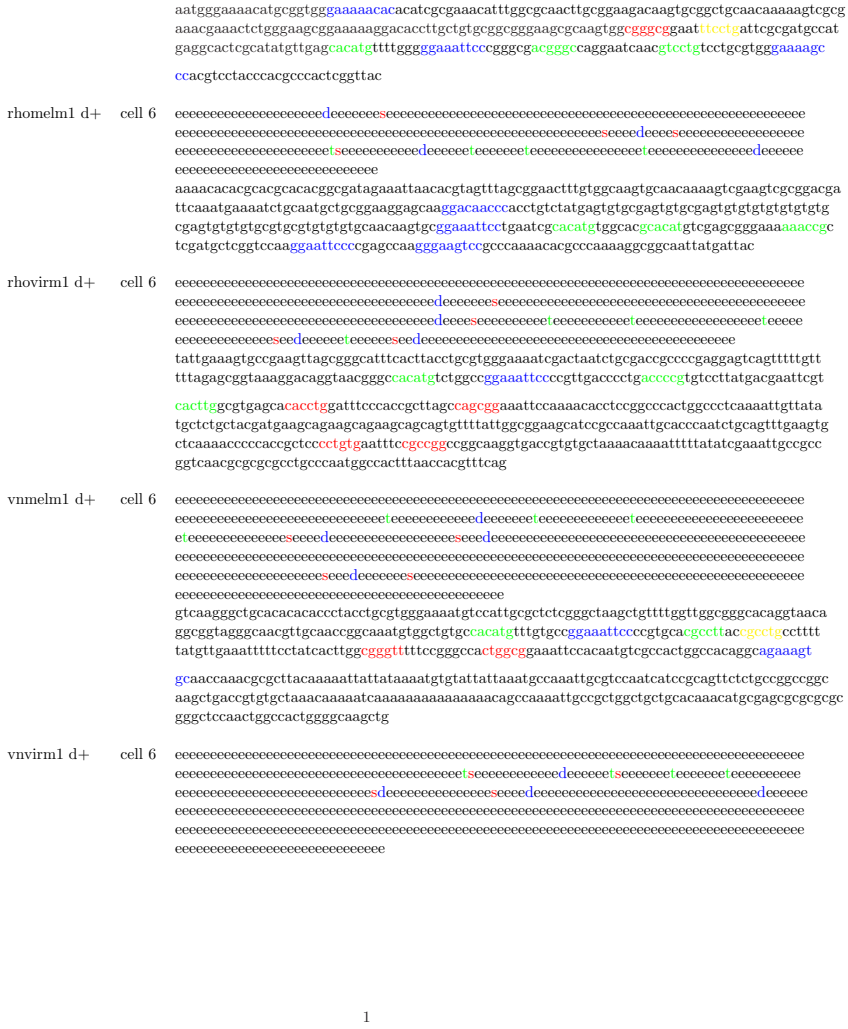

Figure 3.9: Here the target gene is denoted in the left column, where, where the second column contains d+ to denote an in increase (+) in the Dorsal (d) gradient along the DV axis, and the cell along the DV axis used for extracting concentrations for the annotation model MAP is denoted in the third column. Twist site's are annotated green, Dorsal blue, Snail red, and brown denotes overlaps.

3.7 Conclusion

The MPA annotation model developed here was used in conjunction with the OR gate for Dorsal binding sites (DC and DU PWMs). The OR gate was designed to be sensitive (it is more likely to call a positive hit than its component PWMs), while disregarding the pitfuls of specificity (a false positive picked up by just one component detector will be declared a positive for the OR gate, even if the majority of component detectors declare the site a negative.). The high sensitivity of the OR gate used in conjunction with the the MPA annotator (which is designed for high specificity), yields a highly effective bioinformatic tool for discovery of binding sites. Annotation of CRMs at half max occupancy of the Basal Transcription Apparatus (the transition from the boolean 0 to 1 expression), reduces the likelihood of picking up binding sites that are nonfunctional and hence misclassified as a 'hit' in the annotation process, since high occupancy of BTA could result from many binding sites that are not critical to get the initial switch.

The occupancy model generalizes over binding sites, in that the input to the model is an enhancer sequence, not a list of binding sites. So a priori one does not know what binding sites will be selected, rather the physical control parameters (concentration, binding energy, protein-protein interactions) determine what is a binding site. In this sense the MPA annotation model could be thought of as generating a fixed number of binding sites as part of its output. This is a contribution to the field, which i point out Segal said specifically their model could not address at this level of detail, they could only make statements about the numerical values of cooperativity quenching. With this being said one could judiciously choose parameter values to reconstruct the exact engineered enhancers (which in a sense we did for our last experiment), but of course, we wish to find parameters which are consistent

with not just one construct but the entire GRN.

Much of the analysis is about experimental design, as experiments are not only necessary to test theory, but they are necessary to tune parameters for well known theories. Binding of morphogens to CRMs is not a theory under dispute, nor is the theoretical calculation of the occupation of the morphogen's to the CRM, that's been known for many years. What is not known is the strength and importance of the parameters controlling the occupation of morphogens on the promoter, and how much each of these in turn affect BTA binding. This requires model design, to infer what the value of these parameters. However, this data is not always available, hence I have shown an alternative technique, where I have tried to use low quality data from the literature to tune the parameters. As was seen from Experiment 2, it is possible to mathematically arrive at certain conclusions reached by Szymanski (that cooperativity is necessary in the NEEs). However, Szymanski's $6xdl$ construct is not consistent with reported parameter ranges of thermodynamic models collected by Buchler et.al.[20]. Hence, it was necessary to adjust the thermodynamic model parameter ranges. Possibly one could argue that one can not adjust the parameter ranges found by Buchler since they cover a broad spectrum of proteins. Indeed, the values of the cooperativity we found in Experiment 2 are so high that is unlikely Dorsal and Twist would ever unbind once bound; albeit it is possible Dorsal-Twist dimers do occur in the nucleoplasm. Furthermore, Szymanski and others have reported 'uniform' expression profiles in certain constructs (CRMs) that are not fully 'on' (the staining of the target gene was weak). Such reports are not consistent with the themodynamic model for Dorsal occupancy, since Dorsal occupancy is never 'uniform' it is a gradient, like Lewis Wolpert's French Flag model of morphogens. Hence, reconstruction of basic assumptions I have made here may be necessary to accurately represent Dorsal targeted genes in DV.

Lastly, the idea that one should search for the ideal positive definite Hessian at the best estimates of the parameters is incorrect for the DV network. The elusive positive definite Hessian is only for independent parameters. The biochemical parameters that control the logic (gene switches) governing early development have coevolved to work together to jointly regulate the DV axis, hence it is wrong to state that the parameters must be independent.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] C. Adami. The use of information theory in evolutionary biology. *Ann. N. Y. Acad. Sci.*, 1256:49–65, May 2012.

[2] Christoph Adami. What is complexity? *BioEssays*, 24(12):1085–1094, 2002.

[3] A. Afek, J. L. Schipper, J. Horton, R. Gordan, and D. B. Lukatsky. Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. U.S.A.*, 111(48):17140–17145, Dec 2014.

[4] Bruce Alberts. *Molecular biology of the cell*. Garland Science, New York, 4th; 4. edition, 2002.

[5] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, May 1997.

[6] P. Atkins and J de Paula. *Physical Chemistry*. W.H. Freeman and Company, 2002.

[7] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[8] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.

[9] A. S. Bais, N. Kaminski, and P. V. Benos. Finding subtypes of transcription factor motif pairs with distinct regulatory roles. *Nucleic Acids Res.*, 39(11):e76, Jun 2011.

[10] Y. Barash, G. Elidan, T. Kaplan, and Friedman. Modeling dependencies in protein-DNA binding sites. In *Proc of the 7th Ann Int Conf in Comp Mol Bio (RECOMB)*, pages 28–37, 2003.

[11] James Beck and Kenneth Arnold. *Parameter estimation in engineering and science*. Wiley and Sons, 1 edition, 1977.

[12] O. G. Berg. Base-pair specificity of protein-DNA recognition: a statistical-mechanical model. *Biomed. Biochim. Acta*, 49(8-9):963–975, 1990.

[13] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750, Feb 1987.

[14] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24):6929–6948, Nov 1981.

[15] Berg, O G. The evolutionary selection of DNA base pairs in gene-regulatory binding sites. *Proceedings of the National Academy of Sciences*, 89(16):7501–7505, 1992.

[16] W. Bialek. *Biophysics Searching for Principles*. Princeton University Press, 2012.

[17] Bishop. *Pattern Recognition and Machine Learning*, volume 1. Computer Science Press, Rockville, MD, 1985.

[18] C. T. Brown and C. G. Callan. Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 101(8):2404–2409, Feb 2004.

[19] C Titus Brown. Computational approaches to finding and analyzing cis-regulatory elements. *Methods in cell biology*, 87:337–365, 2008.

[20] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. U.S.A.*, 100:5136–5141, Apr 2003.

[21] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. *Genome Res.*, 14(2):201–208, Feb 2004.

[22] Matthew S Busse, Christopher P Arnold, Par Towb, James Katrivesis, and Steven A Wasserman. A sequence code for pathway-specific innate immune responses. *The EMBO Journal*, 26(16):3826–3835, 2007.

[23] J. M. Carothers, S. C. Oestreich, J. H. Davis, and J. W. Szostak. Informational complexity and functional activity of RNA structures. *J. American Chem. Society*, 126:5130–5137, 2004.

[24] S. B. Carroll. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101(6):577–580, Jun 2000.

[25] J. Crocker, N. Potter, and A. Erives. Dynamic evolution of precise regulatory encodings creates the clustered site signature of enhancers. *Nat Commun*, 1:99, 2010.

[26] J. Crocker, Y. Tamori, and A. Erives. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol.*, 6:e263, Nov 2008.

[27] E. H. Davidson and M. S. Levine. Properties of developmental gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, 105(51):20063–20066, Dec 2008.

[28] Eric H. Davidson. *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego, CA, 2001.

[29] Eric H. Davidson. *The Regulatory Genome: Gene Regulatory Networks in Drvelopment and Evolution*. Academic Press, San Diego, CA, 2006.

[30] Thomas A. Down and Tim J. P. Hubbard. Nestedmica: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research*, 33(5):1445–1453, 2005.

[31] Jacqueline Dresch, Xiaozhou Liu, David Arnosti, and Ahmet Ay. Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects. *BMC Systems Biology*, 4(1):142, 2010.

[32] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Bilogical sequence analysis. *Cambridge Press*, 1998.

[33] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004.

[34] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

[35] A. Erives and M. Levine. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.*, 101:3851–3856, Mar 2004.

[36] W. D. Fakhouri, A. Ay, R. Sayal, J. Dresch, E. Dayringer, and D. N. Arnosti. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol. Syst. Biol.*, 6:341, 2010.

[37] D. S. Fields, Y. He, A. Y. Al-Uzri, and G. D. Stormo. Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, 271(2):178–194, Aug 1997.

[38] D. S. Fields and G. D. Stormo. Quantitative DNA sequencing to determine the relative protein-DNA binding constants to multiple DNA sequences. *Anal. Biochem.*, 219(2):230–239, Jun 1994.

[39] Steven M. Gallo, Dave T. Gerrard, David Miner, Michael Simich, Benjamin Des Soye, Casey M. Bergman, and Marc S. Halfon. Redfly v3.0: toward a comprehensive database of transcriptional regulatory elements in drosophila. *Nucleic Acids Research*, 2010.

[40] N. Galtier, M. Gouy, and C. Gautier. SEAVIEW and PHYLO WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, 12(6):543–548, Dec 1996.

[41] Walter J. Gehring. *Master Control Genes in Development and Evolution: The Homeobox Story*. Yale University Press, New Haven, CT, 1998.

[42] B. Georgi and A. Schliep. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, 22(14):e166–173, Jul 2006.

[43] Scott F. Gilbert. *Developmental biology*. Sinauer Associates, Sunderland, Mass, 5th edition, 1997.

[44] Scott F. Gilbert and David Epel. *Ecological developmental biology: integrating epigenetics, medicine, and evolution*. Sinauer Associates, Sunderland, Mass, 2009.

[45] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek. Probing the limits to positional information. *Cell*, 130:153–164, Jul 2007.

[46] Debraj GuhaThakurta and Gary D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.

[47] Naomi Habib, Tommy Kaplan, Hanah Margalit, and Nir Friedman. A novel bayesian dna motif comparison method for clustering and retrieval. *PLoS Comput Biol*, 4(2):e1000010, 2008.

[48] S. Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24(11):1325–1331, Jun 2008.

[49] Sridhar Hannenhalli and Li-San Wang. Enhanced position weight matrices using mixture models. *Bioinformatics*, 21(suppl 1):i204–i212, 2005.

[50] X. He, C. C. Chen, F. Hong, F. Fang, S. Sinha, H. H. Ng, and S. Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS ONE*, 4:e8155, 2009.

[51] X. He, M. A. Samee, C. Blatti, and S. Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput. Biol.*, 6, Sep 2010.

[52] Terrell L Hill. *Cooperativity Theory in Biochemistry: Steady-state and Equilibrium Systems*. New York: Springer-Verlag, 1985.

[53] T.L. Hill. *An Introduction to Statistical Thermodynamics*. Dover Books on Physics and Chemistry. New York : Dover Publications, 1986.

[54] A Hobson. *Concepts in Statistical Mechanics*. Gordon and Breach, New York, 1971.

[55] Joung-Woo Hong, David A. Hendrix, Dmitri Papatsenko, and Michael S. Levine. How the dorsal gradient works: Insights from postgenome technologies. *Proceedings of the National Academy of Sciences*, 105(51):20072–20076, 2008.

[56] Y. T. Ip, R. E. Park, D. Kosman, E. Bier, and M. Levine. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. *Genes Dev.*, 6:1728–1739, Sep 1992.

[57] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[58] J. Jiang, D. Kosman, Y. T. Ip, and M. Levine. The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. *Genes Dev.*, 5:1881–1891, Oct 1991.

[59] J. Jiang and M. Levine. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72:741–752, Mar 1993.

[60] Jin Jiang and Michael Levine. Binding affinities and cooperative interactions with bhlh activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72(5):741 – 752, 1993.

[61] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, Apr 1975.

[62] D. Kosman, Y. T. Ip, M. Levine, and K. Arora. Establishment of the mesoderm-neuroectoderm boundary in the Drosophila embryo. *Science*, 254:118–122, Oct 1991.

[63] D. Landau and I. Lifshitz. *Mechanics*, volume 1. Butterworth Heinemann, 1976.

[64] M. Lassig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8 Suppl 6:S7, 2007.

[65] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, et al. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262:208–208, 1993.

[66] P. Lawrence. *The Making of a Fly*. Wiley-Blackwell, 1 edition, 1992.

[67] T. H. Leung, A. Hoffmann, and D. Baltimore. One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell*, 118(4):453–464, Aug 2004.

[68] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U.S.A.*, 102(14):4936–4942, Apr 2005.

[69] Raphael D. Levine and Myron Tribus, editors. *The Maximum Entropy Formalism*. MIT Press, Cambridge, MA, 1978.

[70] L. Li. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.*, 16(2):317–329, Feb 2009.

[71] X. Liu, D.L. Brutlag, J.S. Liu, et al. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pac Symp Biocomput*, volume 6, pages 127–138, 2001.

[72] S. Mahony and P. V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, 35(Web Server issue):W253–258, Jul 2007.

[73] M. Markstein, R. Zinzen, P. Markstein, K. P. Yee, A. Erives, A. Stathopoulos, and M. Levine. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development*, 131:2387–2394, May 2004.

[74] John Maynard Smith. *Evolutionary genetics*. Oxford University Press, New York; Oxford [Oxfordshire], 1989.

[75] S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, Apr 2009.

[76] A. M. Moses and M. B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput*, 324:324–335, 2004.

[77] B. Moussian and S. Roth. Dorsoventral axis formation in the Drosophila embryo–shaping and transducing a morphogen gradient. *Curr. Biol.*, 15(21):R887–899, Nov 2005.

[78] N. Mrinal, A. Tomar, and J. Nagaraju. Role of sequence encoded DNA geometry in gene regulation by Dorsal. *Nucleic Acids Res.*, 39(22):9574–9591, Dec 2011.

[79] V. Mustonen and M. Lassig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc. Natl. Acad. Sci. U.S.A.*, 102(44):15936–15941, Nov 2005.

[80] Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems 14*, pages 471–478. MIT Press, 2002.

[81] D. J. Obbard, J. Maclennan, K. W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins. Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Mol. Biol. Evol.*, 29(11):3459–3473, Nov 2012.

[82] D. Pan and A. J. Courey. The same dorsal binding site mediates both activation and repression in a context-dependent manner. *EMBO J.*, 11(5):1837–1842, May 1992.

[83] D. Papatsenko, Y. Goltsev, and M. Levine. Organization of developmental enhancers in the Drosophila embryo. *Nucleic Acids Res.*, 37(17):5665–5677, Sep 2009.

[84] D. Papatsenko and M. Levine. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proc. Natl. Acad. Sci. U.S.A.*, 102(14):4966–4971, Apr 2005.

[85] U. J. Pape, H. Klein, and M. Vingron. Statistical detection of cooperative transcription factors with similarity adjustment. *Bioinformatics*, 25(16):2103–2109, Aug 2009.

[86] M. W. Perry, J. D. Cande, A. N. Boettiger, and M. Levine. Evolution of insect dorsoventral patterning mechanisms. *Cold Spring Harb. Symp. Quant. Biol.*, 74:275–279, 2009.

[87] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2 edition, 1992.

[88] J. Reinitz, S. Hou, and D. Sharp. Transcriptional Control in Drosophila. *Complexus*, 1:54–64, 2003.

[89] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[90] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, Apr 1986.

[91] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451:535–540, Jan 2008.

[92] E. Segal and J. Widom. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.*, 10:443–456, Jul 2009.

[93] I. Sela and D. B. Lukatsky. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.*, 101(1):160–166, Jul 2011.

[94] E. Sharon, S. Lubliner, and E. Segal. A feature-based approach to modeling protein-DNA interactions. *PLoS Comput. Biol.*, 4(8):e1000154, 2008.

[95] R. K. Shultzaberger, D. Y. Chiang, A. M. Moses, and M. B. Eisen. Determining physical constraints in transcriptional initiation complexes using DNA sequence analysis. *PLoS ONE*, 2(11):e1199, 2007.

[96] R. Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE*, 5(3):e9722, 2010.

[97] S. Sinha, M. Blanchette, and M. Tompa. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, Oct 2004.

[98] A. Stathopoulos and M. Levine. Genomic regulatory networks and animal development. *Dev. Cell*, 9(4):449–462, Oct 2005.

[99] Alexander J Stewart, Sridhar Hannenhalli, and Joshua B Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–85, Nov 2012.

[100] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 23:109–113, Mar 1998.

[101] G. D. Stormo and Y. Zhao. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, 11(11):751–760, Nov 2010.

[102] P. Szymanski and M. Levine. Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo. *EMBO J.*, 14:2229–2238, May 1995.

[103] N. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 3 edition, 2007.

[104] P. H. von Hippel. On the molecular bases of the specificity of interaction of transcriptional proteins with genome dna. In R.F.Goldberger, editor, *Biological Regulation and Development*, volume 1, pages 279–347. Plenum Publishing, New York, 1979.

[105] L. Wolpert. The evolutionary origin of development: cycles, patterning, privilege and continuity. *Dev. Suppl.*, pages 79–84, 1994.

[106] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, L. A. Romano, and G. A. Wray. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, 20(9):1377–1419, Sep 2003.

[107] J. Zeitlinger, R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young, and M. Levine. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.*, 21:385–390, Feb 2007.

[108] R. P. Zinzen, K. Senger, M. Levine, and D. Papatsenko. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr. Biol.*, 16:1358–1365, Jul 2006.