IOP Publishing *Phys. Biol.* **00** (2015) 000000

Physical Biology

PAPER

Discovery and information-theoretic characterization of transcription factor binding sites that act cooperatively

RECEIVED
31 December 2014

REVISED
13 May 2015

ACCEPTED FOR PUBLICATION
18 May 2015

PUBLISHED
DD MM 2015

Jacob Clifford[1,3] **and Christoph Adami**[1,2,3]

[1] Department of Physics and Astronomy, Michigan State University, East Lansing, MI, USA
[2] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA
[3] BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI, USA

E-mail: adami@msu.edu

Keywords: DNA binding sites, position weight matrices, dorsal ventral network, information theory

Supplementary material for this article is available online

Abstract

Transcription factor binding to the surface of DNA regulatory regions is one of the primary causes of regulating gene expression levels. A probabilistic approach to model protein–DNA interactions at the sequence level is through position weight matrices (PWMs) that estimate the joint probability of a DNA binding site sequence by assuming positional independence within the DNA sequence. Here we construct conditional PWMs that depend on the motif signatures in the flanking DNA sequence, by conditioning known binding site loci on the presence or absence of additional binding sites in the flanking sequence of each site's locus. Pooling known sites with similar flanking sequence patterns allows for the estimation of the conditional distribution function over the binding site sequences. We apply our model to the dorsal transcription factor binding sites active in patterning the dorsal–ventral axis of *Drosophila* development. We find that those binding sites that cooperate with nearby twist sites on average contain about 0.5 bits of information about the presence of twist transcription factor binding sites in the flanking sequence. We also find that dorsal binding site detectors conditioned on flanking sequence information make better predictions about what is a dorsal site relative to background DNA than detection without information about flanking sequence features.

Q1

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 [1]. 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 [2–4]. 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 [5, 6]. 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.

Position weight matrices (PWMs)

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 PWM [7]. For example, for a length s binding site, the probability that the binding site has the sequence S is:

$$P(S) = \prod_{ij}^{s,3} P_{ij}^{S_{ij}}, \quad (1)$$

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 $i \in \{1, 2, ..., s\}$ 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:

$$E(S) = \sum_{i}^{s} \sum_{j}^{3} E_{ij} S_{ij}, \quad (2)$$

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 (10) below.

In-vitro biophysical PWMs

The energy weight matrix elements used in equation (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 [7–11]. 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

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

$$\approx \sum_{ik}^{s,3} G_{ik} S_{ik}. \quad (4)$$

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.

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 [12], 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 substitutions 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 [12]. 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' [13].

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 [12]. Theoretically, they assumed a binding site was constrained by the binding affinity necessary for binding (i.e., binding that influences gene expression) [14]. 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 (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 (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 [15].

For a simple example, consider a binding site of just one base-pair[4]. The 'Lagrangian' for the constrained minimization problem can be written as (the sum is over the nucleotides that base B can take on)

$$\sum_B P(B)\log\frac{P(B)}{P_0(B)} - \lambda_0\left(\sum_B P(B) - 1\right) - \lambda_1\left(\sum_B P(B)\,G(B) - \langle G\rangle\right). \quad (5)$$

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 (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) [18], 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.

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) [12]; Berg and von Hippel were able to derive a simple relation between their information theoretic logarithmic score $E(S)$ from equation (2), and the known binding energies $G(S)$ of the binding sites to the transcription factor from equation (4). 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 (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) as $E(S) = \log\frac{P(S)}{P(S_0)}$[5]; one then obtains:

$$\log K(S) = \log K(S_0) - \frac{\Delta(E(S))}{\lambda_1}, \quad (6)$$

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 perturbation [9]:

$$\log K(S) \approx \log K(S_0) - \Delta G. \quad (7)$$

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).

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. 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

[4] Binding sites are frequently about 10 bp 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 nm). For an evolutionary argument for why binding sites are about 10 bp in length see [16], and for a diffusion argument see [17].

[5] 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 [19]). 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.

was also analyzed by Berg and von Hippel using one- and three-dimensional diffusion models [20], which has since been reinterpreted several times. In particular, Sela and Lukatsky showed that symmetries in DNA sequences flanking functional binding site loci can dramatically affect binding [17], later verified experimentally [21]. 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 [22].

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 [23]. 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 [24–26]. 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 [27]. 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 [26].

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 [28], 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.

Physical shortcomings of PWMs

Dependencies within transcription factor bindings sites

The linear relation in equation (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 [9, 29]. 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 [30, 31].

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 Nf*κ*B is known to have a specificity that is dependent on co-occurring binding sites [32], and similarly the binding sites of the Glucocorticoid Receptor are specific to their context [33]. The Nf*κ*B homolog dorsal's binding sites have also been shown to encode differences when active in different innate immunity pathways [34], or to signal dorsal's role as an activator or a repressor [35].

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 [12], but now we effectively have a spacer model between binding sites.

Furthermore, Berg and von Hippel in [12] 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.

Materials

Data for known dorsal binding sites in *Drosophila melanogaster* dorsal–ventral (DV) 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 [36–38]. The 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 [39]. The transcription factor dorsal regulates the genes responsible for patterning the DV axis of embryogenesis leading to gastrulation [40–42]. 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 *D. melanogaster* neuroectoderm region of the DV axis that cooperate with a basic helix–loop–helix (bHLH) dimer with twist. These sites are the D_β sites of table S2 of Crocker *et al* [28], the dorsal sites from figure 2 of Crocker *et al* [43], as well as the 'specialized' neurogenic ectoderm enhancers (NEE) and NEE-like dorsal binding sites of Erives *et al* and Crocker *et al* [43, 44]). 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 [28], 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 [45–50] 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 labeled 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 [51] 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.

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* [52]. 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 bp 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 1.1 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.

Methods

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 1.2 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:

$$d(i, j) = z^i - z^j - w^i, \tag{8}$$

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'.

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 (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 (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.

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 (5). The Boltzmann-like distribution in units of the second Lagrange multiplier is:

$$P(B) = \frac{P_0(B) \ \exp \ -E(B)}{Z}, \tag{9}$$

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 (6). Assuming our frequencies from equation (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:

$$E(B) \approx -\log \frac{P(B_0)}{P(B)} = -\log \frac{n_{B_0} + \beta}{n_B + \beta}. \tag{10}$$

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 [53], 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) for a particular position and base plus the interaction term:

$$E(B|C) = E(B) + w(B, C). \tag{11}$$

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:

$$w(B, C) = -\log \frac{P(B, C)}{P(B)P(C)} = -\log \frac{P(B|C)}{P(B)}. \quad (12)$$

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

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:

$$w(S, S') = -\log \frac{P(S, S')}{P(S)P(S')}. \quad (13)$$

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:

$$w(S, C) = \sum_{i}^{s} w(B_i, C), \quad (14)$$

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 (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 [54]. 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 (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:

$$E(S|C) = E(S) + w(S, C) . \quad (15)$$

Here $E(S)$ is from equation (2), where the matrix elements E_{ij} are equal to $E(B(j)_i)$ from equation (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 (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 (9).

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 (15) described above, that is, its output depends on the output of two distinct 'subdetectors', which we call 'dorsal cooperative' (DC) and 'dorsal uncooperative' (DU). 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:

$$y(\mathbf{S}) = E_c - \mathbf{E} \cdot \mathbf{S}, \quad (16)$$

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 (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, $\mathbf{E} \cdot \mathbf{S}$, 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)). 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 described 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), which has a corresponding probability $P(S)$ as in equation (1).

Results

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 (e.g., 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 results section titled 'additional experiment supplement' 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 bp 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 1.6 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

$$I(S;\, C) = \sum_{S}\sum_{C} P(S|C)\,P(C)\log\frac{P(S|C)}{P(S)}, \qquad (17)$$

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 30 bp, and then both parameters are incremented by 30 bp 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 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 were missing flanking parts of the conditional binding sites.

Table 1. Mutual information between functional dorsal binding site sequences and putative twist sites that match 5′-CAYATG using a sliding spacer window scheme.

Spacer	[0, 30] bp	(31, 60] bp	(61, 90] bp
Mutual information, equation (17)	0.49	0.29	0.04

We show the conditional dorsal binding site sequence logos for functional binding sites generated for this first spacer window in figure 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 sections 1.9 and 1.17.

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

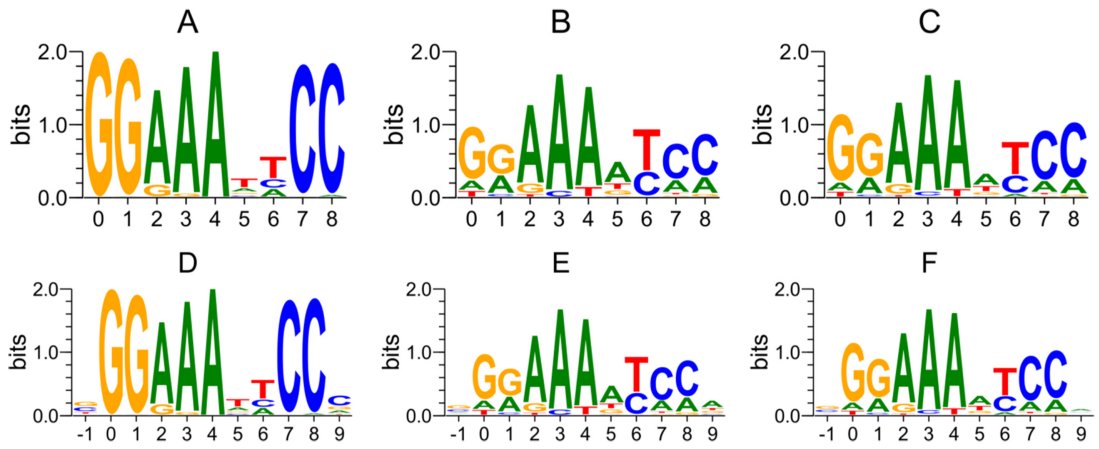

Figure 1. Logos generated for known dorsal sites (the D_{CB} data) tested for adjacency to 5′-CAYATG used as the cooperative class if in the [0, 30] bp distance. Logo A corresponds to the cooperative class, and displays the known 5′-AAATT core, with total information content 13.5 bits. Logo D is the 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' (two out of ten 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.

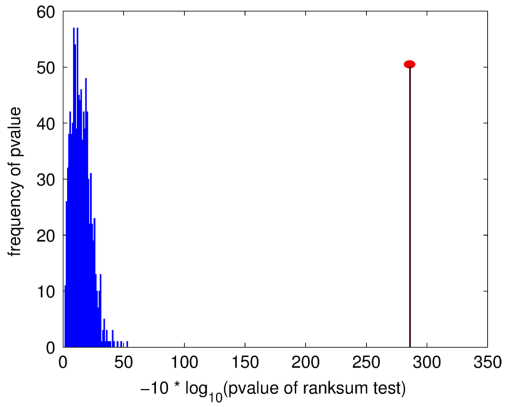

Figure 2. Histogram of p-values of a rank sum test of random partitions of the combined data set $\mathcal{D}_{CB}$. The binning is in units $-10 \times \log_{10}$ of the p-value, rounded to the nearest integer. The p-value of the rank sum test between DC and DU energy data sets based on their energy PWMs was 260 in log base ten units (scaled by 10), which is indicated by the red bar of arbitrary height.

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), 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 1.11.

Performance of optimal classifiers (detectors)

All detectors were built from length 9 alignments (see supplement section 1.6 for 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 $\mathcal{D}_{CB}$ 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 $\mathcal{D}_{CB}$ 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.

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 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

Table 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 $E(S) = E_c = 2.1$.

	Proximal	Distal
DC	61	39
DU	280	345

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 $E_c = 5$.

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 $\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$, where the denominator is the total counts of true positives (TP) and false negatives (FN)). The false positive rate (FPR) is defined as $\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{FP} + \mathrm{TN})}$, 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 1.3, 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 3(A)) 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 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.

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 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

$$I(\mathcal{I}; \mathcal{O}) = H(\mathcal{I}) - H(\mathcal{I}|\mathcal{O}), \tag{18}$$

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 toward 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 FP and TPRs introduced earlier. In particular, the conditional probability $p(i|0)$ is obtained as

$$p(1|1) = p(\mathcal{I} = 1|\mathcal{O} = 1) = \mathrm{TPR}, \tag{19}$$

$$p(1|0) = p(\mathcal{I} = 1|\mathcal{O} = 0) = 1 - \mathrm{TPR}, \tag{20}$$

$$p(0|1) = p(\mathcal{I} = 0|\mathcal{O} = 1) = \mathrm{FPR}, \tag{21}$$

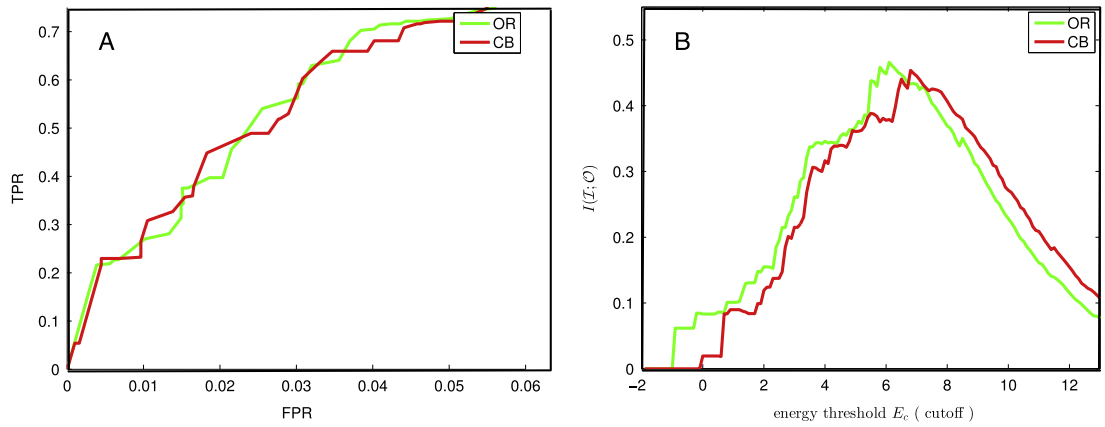

Figure 3. ROC and information. (A) False positive rate (FPR) versus true positive rate (TPR) when varying the energy cutoff E_c. (B) shows the mutual information $I(\mathcal{I};\mathcal{O})$ equation (23) between the input and output of the detectors as a function of the cutoff energy.

$$p(0|0) = p(\mathcal{I} = 0|\mathcal{O} = 0) = 1 - \text{FPR}, \qquad (22)$$

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 (18) as:

$$I(\mathcal{I};\mathcal{O}) = H[p] - pH[\text{TPR}] - (1-p)H[\text{FPR}], \qquad (23)$$

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

$$H[\text{TPR}] = -\frac{\text{FN}}{\text{TP}+\text{FN}}\log\frac{\text{FN}}{\text{TP}+\text{FN}} - \frac{\text{TP}}{\text{TP}+\text{FN}}\log\frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad (24)$$

with a similar expression for $H[\text{FPR}]$. We show the mutual information $I(\mathcal{I};\mathcal{O})$ in figure 3(B) 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_{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 1.13). We show this mutual information $I(\mathcal{C};\mathcal{P})$ in figure 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.

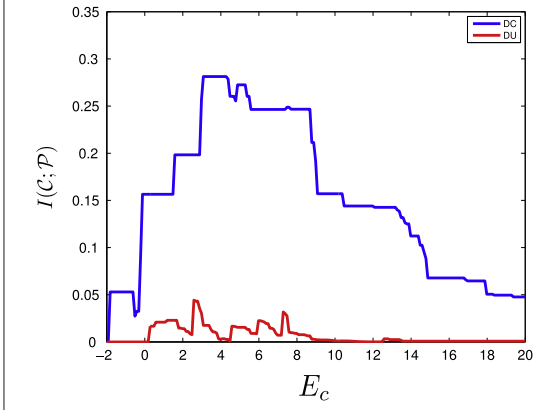

Figure 4. Mutual information $I(\mathcal{C};\mathcal{P})$ between the actual classes $\mathcal{C}$ and the predicted classes $\mathcal{P}$ for detectors DC and DU as a function of the threshold energy E_c that is defined by each detector's conditional energy equation (15).

Discussion

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* [28]. 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 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 [35], 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 [55]. 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 $\boldsymbol{\mathcal{D}}_{\mathbf{DCmel}}$ 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) [28, 44]. However, our DC detector is different than a detector built strictly from the $\boldsymbol{\mathcal{D}}_{\mathbf{DC}}$ 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* [52, 56] showed that binning the spacers between dorsal and twist that there were various optimal bins (namely 14 bp, 20 bp, and 53 bp). 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 and Levine [48] 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* [57] 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.

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 [58, 59]. 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:

$$P(S) = \sum_c \frac{\exp -E(S|C=c)}{Z_c} P(C=c), \qquad (25)$$

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 $\boldsymbol{\mathcal{D}}_{\mathbf{CB}}$ data simultaneously. Recall from equation (1)

$$P(S) = \prod_i P(S_i), \qquad (26)$$

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 (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 1.8.

The mixture model was used by Hannenhali and Wang [59] 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 TPR of the detector is partially cancelled by

the cost of an increased FPR. 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.

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 [60]. 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 [61, 62].

The mutual information between sequences and the OR gate's predictions in figure 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 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 3(B) is not as visible in the ROC curves in figure 3(A), 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 3).

The degree to which the OR gate's ROC does appear shifted relative to CB's ROC in figure 3(A) 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 [12] 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 (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 [56] 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 [56]. It is also consistent with the mathematical definition of 'specialized' sites from Erives and Levine [44] and the D_β sites of Crocker *et al* [28] who defined these sites based on how they were detected (similar to MEME's one occurrence per sequence setting [63], 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).

Conditional detectors

In figure 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 (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:

$$H(\mathcal{C}|\mathcal{P}) = H(\mathcal{C}) - I(\mathcal{C};\mathcal{P}). \quad (27)$$

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 (27) we have

$$H(\mathcal{C}|\mathcal{P}) = 1 - 0.3 = 0.7 \text{ bits}, \quad (28)$$

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.

Conclusion

PWMs 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 [54, 64–72]. In addition, others have looked at the influence of symmetries in the flanking sequence of binding sites [17, 21]. 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 [57, 73].

Acknowledgments

We would like to thank David Arnosti and C Titus Brown for extensive discussions, as well as the members of the Adami Lab. This work was supported in part by NSF's BEACON Center for the Study of Evolution in Action, under Contract No. DBI-0939454.

References

Q2

[1] Landau D and Lifshitz I 1976 *Mechanics* vol 1 (Portsmouth, NH: Heinemann)
[2] Davidson E H and Levine M S 2008 *Proc. Natl Acad. Sci. USA* **105** 20063–6
[3] Arnone M I and Davidson E H 1997 *Development* **124** 1851–64
[4] Lassig M 2007 *BMC Bioinformatics* **8** Suppl 6S7
[5] Carroll S B 2000 *Cell* **101** 577–80
[6] Wray G A, Hahn M W, Abouheif E, Balhoff J P, Pizer M, Rockman M V, Romano L A and Wray G A 2003 *Mol. Biol. Evol.* **20** 1377–419
[7] Stormo G D and Fields D S 1998 *Trends Biochem. Sci.* **23** 109–13
[8] Fields D S and Stormo G D 1994 *Anal. Biochem.* **219** 230–9
[9] Hill T L 1985 *Cooperativity Theory in Biochemistry: Steady-State and Equilibrium Systems* (New York: Springer)
[10] Stormo G D and Zhao Y 2010 *Nat. Rev. Genetics* **11** 751–60
[11] Fields D S, He Y, Al-Uzri A Y and Stormo G D 1997 *J. Mol. Biol.* **271** 178–94
[12] Berg O G and von Hippel P H 1987 *J. Mol. Biol.* **193** 723–50
[13] Sinha S, Blanchette M and Tompa M 2004 *BMC Bioinformatics* **5** 170
[14] Berg O G 1992 *Proc. Natl Acad. Sci.* **89** 7501–5
[15] Hobson A 1971 *Concepts in Statistical Mechanics* (New York: Gordon and Breach)
[16] Stewart A J, Hannenhalli S and Plotkin J B 2012 *Genetics* **192** 973–85

Q3

[17] Sela I and Lukatsky D B 2011 *Biophys. J.* **101** 160–6
[18] von Hippel P H 1979 On the molecular bases of the specificity of interaction of transcriptional proteins with genome DNA *Biological Regulation and Development* vol 1 ed R F Goldberger (New York: Plenum) pp 279–347
[19] Atkins P and de Paula J 2002 *Physical Chemistry* (San Francisco: Freeman)
[20] Berg O G, Winter R B and von Hippel P H 1981 *Biochemistry* **20** 6929–48
[21] Afek A, Schipper J L, Horton J, Gordan R and Lukatsky D B 2014 *Proc. Natl Acad. Sci. USA* **111** 17140–5
[22] Brown C T and Callan C G 2004 *Proc. Natl Acad. Sci. USA* **101** 2404–9
[23] Schneider T D and Stephens R M 1990 *Nucleic Acids Res.* **18** 6097–100
[24] Gehring W J 1998 *Master Control Genes in Developent and Evolution: The Homeobox Story* (New Haven, CT: Yale Univesity Press)
[25] Davidson E H 2001 *Genomic Regulatory Systems: Development and Evolution* (San Diego, CA: Academic)
[26] Davidson E H 2006 *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (San Diego, CA: Academic)
[27] Brown C T 2008 *Methods Cell Biol.* **87** 337–65
[28] Crocker J, Potter N and Erives A 2010 *Nat. Commun.* **1** 99
[29] Bialek W 2012 *Biophysics Searching for Principles* (Princeton, NJ: Princeton University Press)
[30] Siddharthan R 2010 *PLoS One* **5** e9722
[31] Sharon E, Lubliner S and Segal E 2008 *PLoS Comput. Biol.* **4** e1000154
[32] Leung T H, Hoffmann A and Baltimore D 2004 *Cell* **118** 453–64
[33] Meijsing S H, Pufall M A, So A Y, Bates D L, Chen L and Yamamoto K R 2009 *Science* **324** 407–10
[34] Busse M S, Arnold C P, Towb P, Katrivesis J and Wasserman S A 2007 *EMBO J.* **26** 3826–35
[35] Mrinal N, Tomar A and Nagaraju J 2011 *Nucleic Acids Res.* **39** 9574–91
[36] Lawrence P 1992 *The Making of a Fly* 1st edn (New York: Wiley-Blackwell)
[37] Hong J W, Hendrix D A, Papatsenko D and Levine M S 2008 *Proc. Natl Acad. Sci.* **105** 20072–6
[38] Levine M and Davidson E H 2005 *Proc. Natl Acad. Sci. USA* **102** 4936–42
[39] Perry M W, Cande J D, Boettiger A N and Levine M 2009 *Cold Spring Harbor Symp. Quant. Biol.* **74** 275–9
[40] Moussian B and Roth S 2005 *Curr. Biol.* **15** R887–99
[41] Stathopoulos A and Levine M 2005 *Developmental Cell* **9** 449–62
[42] Zeitlinger J, Zinzen R P, Stark A, Kellis M, Zhang H, Young R A and Levine M 2007 *Genes Dev.* **21** 385–90
[43] Crocker J, Tamori Y and Erives A 2008 *PLoS Biol.* **6** e263
[44] Erives A and Levine M 2004 *Proc. Natl Acad. Sci. USA* **101** 3851–6
[45] Markstein M, Zinzen R, Markstein P, Yee K P, Erives A, Stathopoulos A and Levine M 2004 *Development* **131** 2387–94
[46] Kosman D, Ip Y T, Levine M and Arora K 1991 *Science* **254** 118–22
[47] Zinzen R P, Senger K, Levine M and Papatsenko D 2006 *Curr. Biol.* **16** 1358–65
[48] Szymanski P and Levine M 1995 *EMBO J.* **14** 2229–38
[49] Jiang J and Levine M 1993 *Cell* **72** 741–52
[50] Ip Y T, Park R E, Kosman D, Bier E and Levine M 1992 *Genes Dev.* **6** 1728–39
[51] Gallo S M, Gerrard D T, Miner D, Simich M, Des Soye B, Bergman C M and Halfon M S 2010 *Nucleic Acids Res.* **Q4**
[52] Papatsenko D, Goltsev Y and Levine M 2009 *Nucleic Acids Res.* **37** 5665–77
[53] Berg O G 1990 *Biomed. Biochim. Acta* **49** 963–75
[54] Barash Y, Elidan G, Kaplan T and Friedman 2003 Modeling dependencies in protein-DNA binding sites *Proc. 7th Annual Int. Conf. in Computational Molecular Biology (RECOMB)*
[55] Pan D and Courey A J 1992 *EMBO J.* **11** 1837–42
[56] Papatsenko D and Levine M 2005 *Proc. Natl Acad. Sci. USA* **102** 4966–71
[57] Fakhouri W D, Ay A, Sayal R, Dresch J, Dayringer E and Arnosti D N 2010 *Mol. Syst. Biol.* **6** 341
[58] Mustonen V and Lassig M 2005 *Proc. Natl Acad. Sci. USA* **102** 15936–41
[59] Hannenhalli S and Wang L S 2005 *Bioinformatics* **21** i204–12
[60] Adami C 2012 *Ann. New York Acad. Sci.* **1256** 49–65
[61] Adami C 2002 *BioEssays* **24** 1085–94
[62] Carothers J M, Oestreich S C, Davis J H and Szostak J W 2004 *J. Am. Chem. Soc.* **126** 5130–7
[63] Bailey T L and Elkan C 1995 *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3** 21–29
[64] Liu X *et al* 2001 Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes *Pac. Symp. Biocomputing* **6** 127–38
[65] Bais A S, Kaminski N and Benos P V 2011 *Nucleic Acids Res.* **39** e76
[66] Georgi B and Schliep A 2006 *Bioinformatics* **22** e166–73
[67] Bulyk M L, McGuire A M, Masuda N and Church G M 2004 *Genome Res.* **14** 201–8
[68] Hannenhalli S 2008 *Bioinformatics* **24** 1325–31
[69] Pape U J, Klein H and Vingron M 2009 *Bioinformatics* **25** 2103–9
[70] GuhaThakurta D and Stormo G D 2001 *Bioinformatics* **17** 608–21
[71] Li L 2009 *J. Comput. Biol.* **16** 317–29
[72] Moses A M and Eisen M B 2004 *Pac. Symp. Biocomputing* **324** 324–35
[73] He X, Samee M A, Blatti C and Sinha S 2010 *PLoS Comput. Biol.* **6** **Q5**

QUERY FORM

JOURNAL: Physical Biology

AUTHOR: J Clifford and C Adami

TITLE: Discovery and information-theoretic characterization of transcription factor binding sites that act cooperatively

ARTICLE ID: pb515530

The layout of this article has not yet been finalized. Therefore this proof may contain columns that are not fully balanced/matched or overlapping text in inline equations; these issues will be resolved once the final corrections have been incorporated.

Page 1

Q1
We have been provided funding information for this article as below. Please confirm whether this information is correct. Directorate for Biological Sciences DBI-0939454.

Page 14

Q2
Please check the details for any journal references that do not have a link as they may contain some incorrect information.

Page 14

Q3
For this journal, article titles are required for references to journals. Please provide article titles for all journal references.

Page 15

Q4
Please update the volume and page range in reference [51].

Page 15

Q5
Please provide the page range or article number in reference [73].