Supplementary Material

Quantitative Modeling of Gene Expression Using DNA Shape Features of Binding Sites

Pei-Chen Peng[1] and **Saurabh Sinha**[1,2,*]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[2]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

* To whom correspondence should be addressed. Email: sinhas@illinois.edu

Note S1. Main ideas of GEMSTAT, a thermodynamics-based model of transcriptional regulation

Here, we briefly review the main ideas of GEMSTAT and formulate the key modification to its architecture that allows it to utilize DNA shape information. As delineated in Figure S3, transcriptional regulation can be modelled as the interaction of three components: DNA sequence, TFs, and the basal transcriptional machinery (BTM). A TF can bind on any site of the DNA sequence with a site-specific probability or affinity. The BTM can bind on the core promoter and initiate transcription. Presumably, the interactions of TF-DNA, BTM-DNA, and TF-BTM occur in thermodynamic equilibrium. Following Shea and Ackers (1), GEMSTAT assumes the gene expression level is proportional to the fractional BTM occupancy at the promoter.

GEMSTAT computes the fractional occupancy of the BTM by considering an ensemble of molecular configurations, each of which is denoted by σ and specifies which sites are bound and which are free. All configurations assume two states: one where the BTM is bound or another where the BTM is unbound. The statistical weights of the two states are $\mathrm{W}(\sigma)\mathrm{Q}(\sigma)$ and $\mathrm{W}(\sigma)$ respectively. $\mathrm{W}(\sigma)$ represents the contribution of TF-DNA interactions, calculated based on TF concentrations and binding affinities of bound sites; $\mathrm{Q}(\sigma)$, conversely, represents the contribution of TF-BTM interactions, modelled as a α, a vector of free parameters with one scalar for each TF, as indicated in Figure S3. Given this, the relative probability of bound BTM is the following, where the gene expression level is proportional to E:

$$\mathrm{E} = \frac{\sum_{\sigma} \mathrm{W}(\sigma)\mathrm{Q}(\sigma)}{\sum_{\sigma} \mathrm{W}(\sigma)\mathrm{Q}(\sigma) + \sum_{\sigma} \mathrm{W}(\sigma)}$$

In this next section, we detail the derivation of the statistical weight $\mathrm{W}(\sigma)$. The sub components of the statistical weight are the contributions of each binding site in a configuration σ. As shown in Figure S3, $q(S)$ represents the contribution of a binding site S to $\mathrm{W}(\sigma)$ and is given by the following equation:

$$q(S) = K(S_{\max}) v [\mathrm{TF}]_{\mathrm{rel}} \exp[\mathrm{LLR}(S)\text{-}\mathrm{LLR}(S_{\max})]$$

In this formulation, $[\mathrm{TF}]_{\mathrm{rel}}$ represents the relative TF concentration to some constant v. $\mathrm{LLR}(S)$-$\mathrm{LLR}(S_{\max})$ represents the difference in the log likelihood ratio between the site S and the consensus binding site $S_{\max}$, and $K(S_{\max})$ represents the association constant of TF-DNA binding. Since both $K(S_{\max})$ and v are unknown constants, GEMSTAT treats the product of the two as a free parameter. The statistical weight $\mathrm{W}(\sigma)$ is then given by the following equation, in the absence of cooperativity:

$$\mathrm{W}(\sigma) = \prod_{\mathrm{i}} q(S_i)^{\sigma_i}$$

, where σ_i is an indicating variable (values 0 or 1) for site S_i being bound by its TF in configuration σ.

Note S2. Biophysical view of TF-DNA binding

Consider a bimolecular reversible reaction of the TF binding to a short piece of DNA to be represented as

$$\mathrm{TF} + \mathrm{DNA} \overset{K}{\leftrightarrow} \mathrm{TF} \cdot \mathrm{DNA}$$

where K is relative binding affinity based on DNA sequence S and can be calculated from the concentration of TF and the concentration of bound complex $\mathrm{TF} \cdot \mathrm{DNA}$

$$K = \frac{[\mathrm{TF} \cdot \mathrm{DNA}]}{[\mathrm{TF}][\mathrm{DNA}]}$$

Note that the equilibrium probability of a site S being bound is

$$\Pr(S \text{ bound}) = \frac{[\mathrm{TF} \cdot \mathrm{DNA}]}{[\mathrm{TF} \cdot \mathrm{DNA}] + [\mathrm{DNA}]} = \frac{K[\mathrm{TF}]}{K[\mathrm{TF}] + 1}$$

Let $\mathrm{Shape}(S)$ be the score assigned by the Random Forest classifier to binding site S. Assume that the score is normalized to be in the range 0 (minimum) to 1 (maximum). We have tested the following two approaches in combining the shape score into sequence to expression models.

Approach 1.

Assume that $\frac{\mathrm{Shape}(S)}{2}$ is the probability of site S being bound at conditions where $[\mathrm{TF}] = \frac{1}{K(S_{\max})}$, where $S_{\max}$ is the consensus binding site. The relative binding affinity can be represented S as

$$K = K(S_{\max})e^{-\Delta E(S)}$$

where $\Delta E(S)$ is $E(S) - E(S_{\max})$, with $E(S) \geq E(S_{\max})$ and $E(S)$ is the binding energy of the TF to binding site. Therefore, the equilibrium probability of a site S being bound is

$$\Pr(S \text{ bound}) = \frac{[\mathrm{TF}]K(S_{\max})e^{-\Delta E(S)}}{[\mathrm{TF}]K(S_{\max})e^{-\Delta E(S)} + 1}$$

Note that since $[\mathrm{TF}] = \frac{1}{K(S_{\max})}$ at the condition assumed above, we have

$$\Pr(S \text{ bound}) = \frac{e^{-\Delta E(S)}}{e^{-\Delta E(S)} + 1}$$

and therefore

$$\frac{\mathrm{Shape}(S)}{2} = \frac{1}{1 + e^{\Delta E(S)}}$$

Note that for $S = S_{\max}$ we have $\Delta E(S) = 0$. Therefore, $\frac{\text{Shape}(S_{\max})}{2} = \frac{1}{1+1} = \frac{1}{2}$, i.e. $\text{Shape}(S_{\max}) = 1$, as it should be.

In general,

$$\Delta E(S) = \ln(\frac{2 - \text{Shape}(S)}{\text{Shape}(S)})$$

Use the above formula of $\Delta E(S)$ in calculating the statistical weight of a site as

$$q(S) = K(S_{\max})[\text{TF}]e^{-\Delta E(S)} = K(S_{\max})[TF]\frac{\text{Shape}(S)}{2 - \text{Shape}(S)}$$

Approach 2.

Following Pujato et al. (2), the relative binding affinity is defined as

$$K = e^{-\frac{A}{K_B T}(1-\text{Shape}(S))}$$

where A is a proportionality constant in units of Kcal/mol, K_B is the Boltzman constant in Kcal/(mol• K) and T is the temperature in Kelvin. In Pujato et al., the best results were observed when $A = 4.74$ Kcal/mol at 298 K, we therefore treated $\frac{A}{K_B T}$ as one parameter k and set $k = 8.0$ as the default starting value when training the sequence to expression model.

The equilibrium probability of a site S being bound becomes

$$\Pr(S \text{ bound}) = \frac{[\text{TF}]e^{-k(1-\text{Shape}(S))}}{[\text{TF}]e^{-k(1-\text{Shape}(S))} + 1}$$

and the statistical weight of a site is

$$q(S) = [\text{TF}]e^{-k(1-\text{Shape}(S))}$$

Note S3. DNA shape model outperform PWM model under the same sequence length

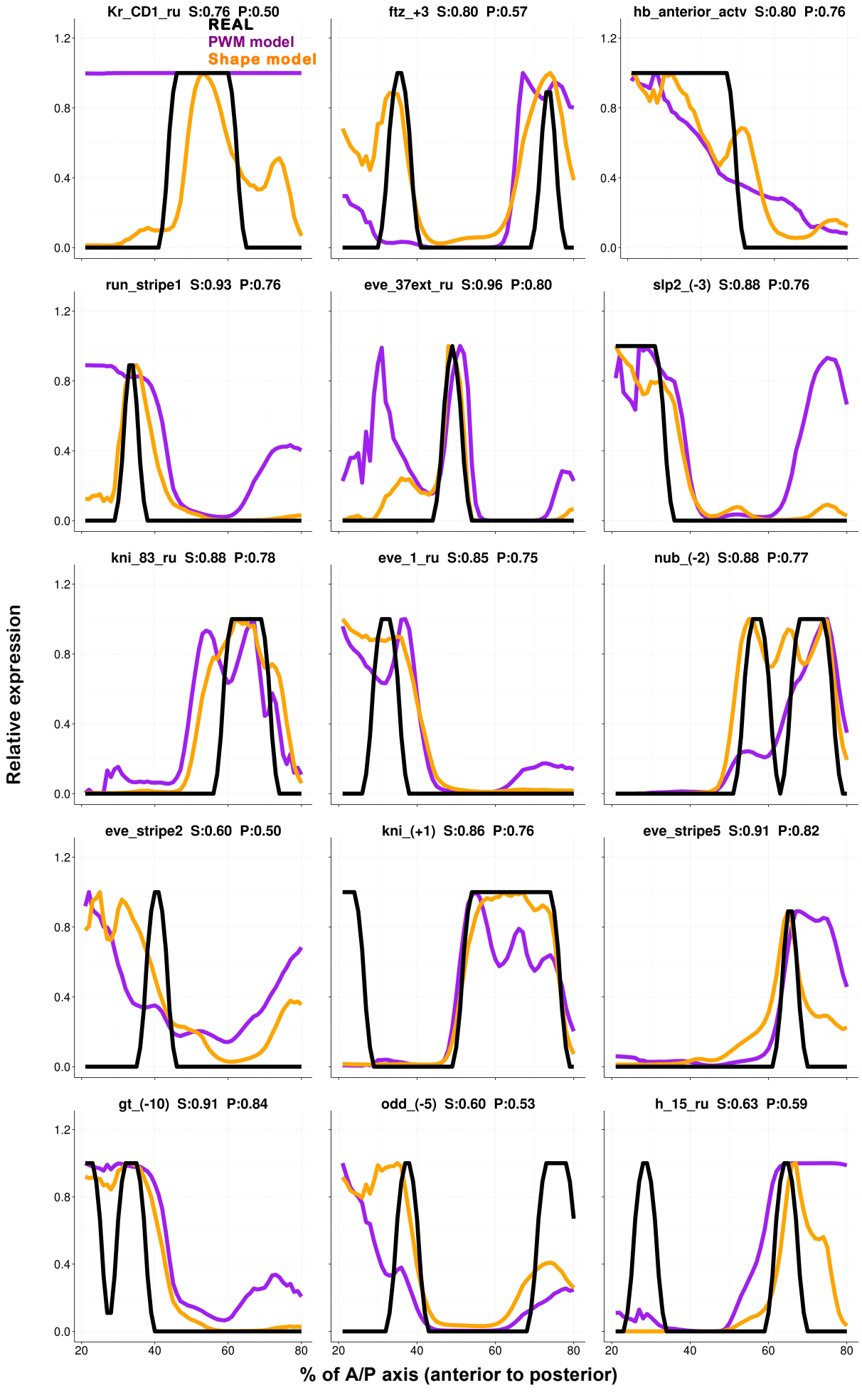
To make a fair comparison and investigate the extent DNA shape would improve sequence to expression modeling, we used trimmed PWMs in all the aforementioned PWM-based models. We first applied MEME (3) to discover motifs and trimmed off less informative positions on either ends. In general, for each of the nine TFs, about zero to three positions were removed from the recommend PWMs. Trimming out less informative positions was ideally acceptable because we resisted to deteriorate the performance of PWM-based model so that the DNA shape-based model would look better. However, one may claim that DNA shape obtained more information from positions where its PWM counterpart ignored and therefore fit enhancers more accurately. Here, we applied a thorough analysis on the original untrimmed PWMs whose lengths were identical to the DNA shape-based putative binding sites.

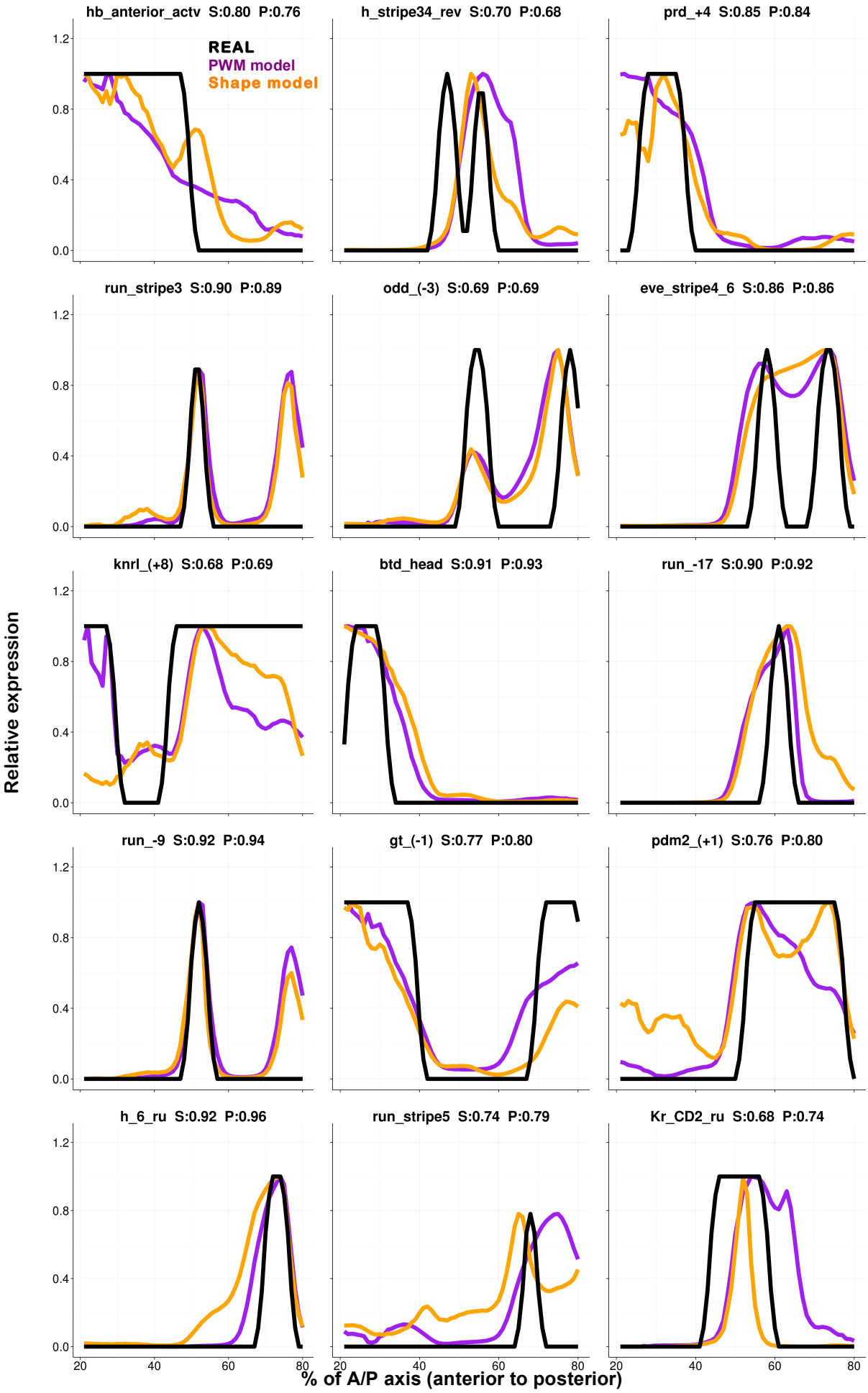
Our intension was first to see how the length of PWMs would affect the modeling. Generally speaking, trimmed PWMs were more suitable for modeling. Figure S2A plots the wPGP scores for each enhancer in models using either trimmed or original long PWMs. The average wPGP score was 0.755 for trimmed PWMs model, outperforming the regular PWMs model whose score was 0.734. Detailed fitting of each enhancer can be seen in Table S4. At this point, we were confident that trimmed PWMs played a better role in the PWM-based model.

On the other direction, we tried to answer the question: given the same binding site length as DNA shape did, would the PWM-based model be able to gather more information and thus make better predictions? Figure S2B and Table S4 reports the comparison of the DNA shape-based model and untrimmed PWM-based model over 37 enhancers. In the majority of cases, DNA shape-based model had considerably better fits than untrimmed PWM-based model. There were 15 out of 37 enhancers having measurable improvements in DNA shape-based model while only three declined. The average wPGP score was 0.784 for the DNA shape-based model compared to 0.734 for untrimmed PWMs model.

Note S4. Artificially perturbing the LLR scores of binding sites show DNA shape carry information complementary to LLR scores

We considered the possibility that the improvement of the shape-based model over the PWM-based model (average wPGP of 0.784 for the shape-based model compared to 0.755 for the PWM-based model) is an artifact of our procedure. Specifically, it was possible that our modeling is fundamentally incapable of discerning an accurate TF-DNA binding model from a noisy version thereof, either due to noise in the data or over-parameterization, or for an unknown reason. To test this possibility, we repeated the PWM-based model-fitting exercise after artificially perturbing the LLR scores of binding sites, and found the PWM model to perform worse with these slightly perturbed LLR scores of sites, ruling out the concern raised above. For each binding site in each enhancer, an artificial LLR score was assigned at random, sampling from a normal distribution with mean equal to the site's true LLR score and a fixed variance. This added 'noise' was tuned to be such that the Pearson correlation between true and perturbed LLR scores was ~0.5, which we noted above to be the overall correlation between shape scores and LLR scores (Figure 4C, 'All'). As shown in Table 1, this PWM-based model performed substantially worse than with true LLR scores: the average wPGP score over 37 enhancers dramatically decreased to 0.643 (compared to 0.755) and the 10-fold cross-validation wPGP score (averaged over ten repeats) dropped from 0.677 to 0.603. This exercise strongly suggested to us that the better fits predicted by the DNA shape-based model compared to the PWM-based model cannot be reproduced merely by a good approximation to LLR scores of sites, and that the shape scores carry information that is complementary to LLR scores and useful for sequence-to-expression modeling.


Kr_CD1_ru S:0.76 P:0.50
ftz_+3 S:0.80 P:0.57
hb_anterior_actv S:0.80 P:0.76
REAL
PWM model
Shape model
run_stripe1 S:0.93 P:0.76
eve_37ext_ru S:0.96 P:0.80
slp2_(-3) S:0.88 P:0.76
kni_83_ru S:0.88 P:0.78
eve_1_ru S:0.85 P:0.75
nub_(-2) S:0.88 P:0.77
eve_stripe2 S:0.60 P:0.50
kni_(+1) S:0.86 P:0.76
eve_stripe5 S:0.91 P:0.82
gt_(-10) S:0.91 P:0.84
odd_(-5) S:0.60 P:0.53
h_15_ru S:0.63 P:0.59
Relative expression
0.0
0.4
0.8
1.2
20
40
60
80
% of A/P axis (anterior to posterior)


hb_anterior_actv S:0.80 P:0.76
h_stripe34_rev S:0.70 P:0.68
prd_+4 S:0.85 P:0.84
REAL
PWM model
Shape model
run_stripe3 S:0.90 P:0.89
odd_(-3) S:0.69 P:0.69
eve_stripe4_6 S:0.86 P:0.86
knrl_(+8) S:0.68 P:0.69
btd_head S:0.91 P:0.93
run_-17 S:0.90 P:0.92
run_-9 S:0.92 P:0.94
gt_(-1) S:0.77 P:0.80
pdm2_(+1) S:0.76 P:0.80
h_6_ru S:0.92 P:0.96
run_stripe5 S:0.74 P:0.79
Kr_CD2_ru S:0.68 P:0.74
Relative expression
0.0
0.4
0.8
1.2
20
40
60
80
% of A/P axis (anterior to posterior)

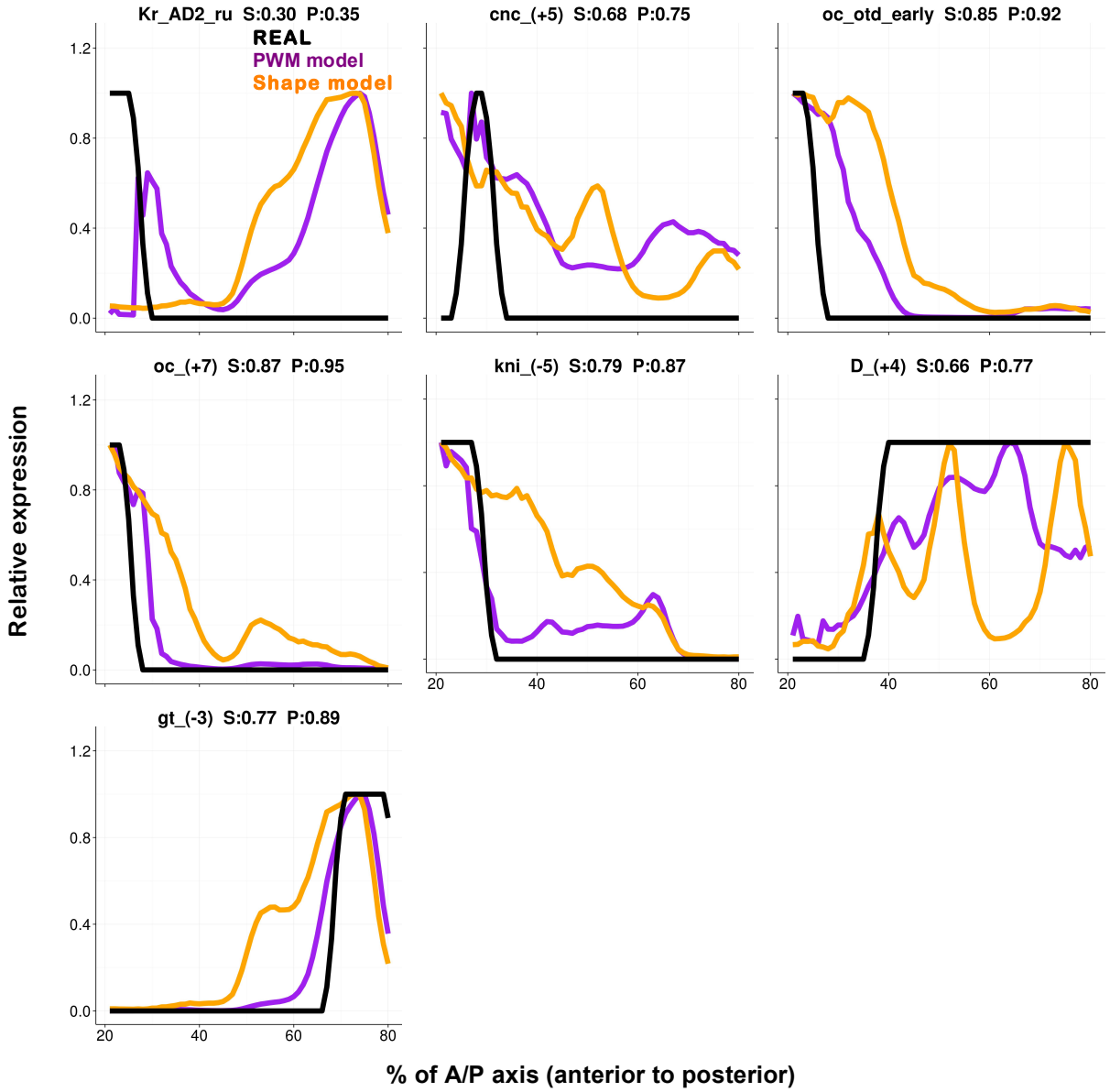


Figure S1. Fits between model and data. Predicted expression profiles of DNA shape-based model (*orange lines*) and PWM-based model (*purple lines*) are compared to experimentally determined expression profiles (*black lines*), for all 37 *Drosophila* enhancers in this study. Each expression profile is on a relative scale of 0 to 1 (*y-axis*), and shown for the regions between 20% and 80% of the A/P axis of the embryo. Title in each panel is in the format of "enhancer name, wPGP by DNA shape-based model ('S'), wPGP by PWM-based model ('P')." The order of enhancers is the same as in Table S1.

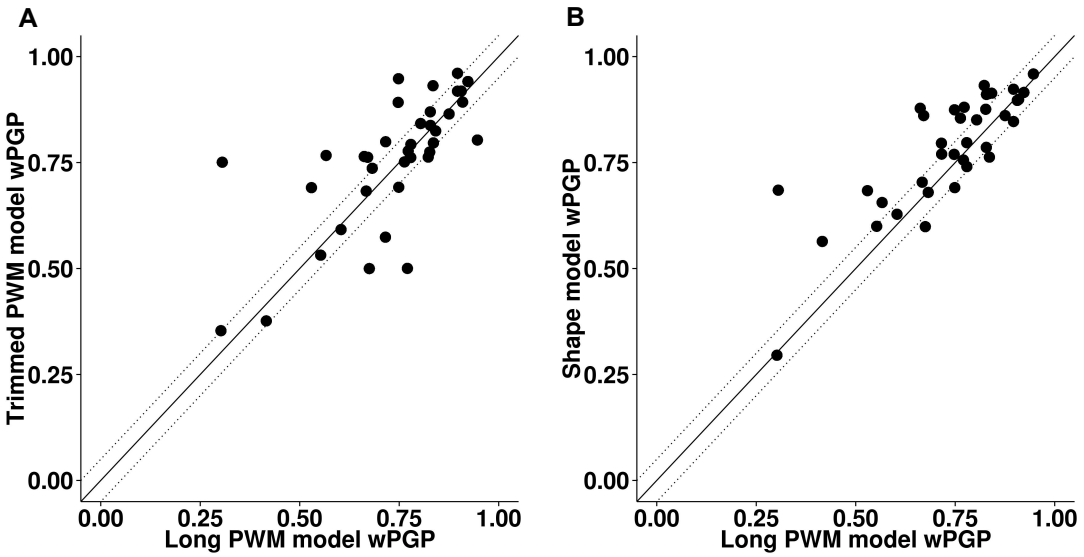


Figure S2. Performance of long PWM models compared to (A) trimmed PWM model and (B) DNA shape model on 37 *Drosophila* enhancers assessed by wPGP scores.

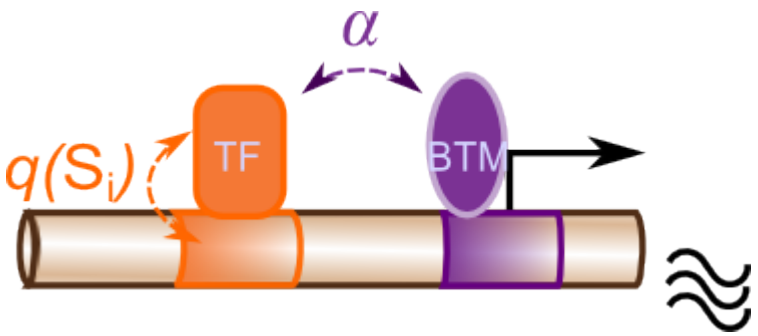


Figure S3. Transcriptional regulation is modelled on major components: TF (*orange*), BTM (*purple*), and DNA (*brown tube*). The interactions of TF-DNA, BTM-DNA, TF-BTM are assumed to occur in thermodynamic equilibrium. Presumably, gene expression level is proportional to the fractional BTM occupancy at the promoter.

Table S1. Lengths of trimmed PWMs and positions being trimmed from MEME predicted PWMs.

TF	MEME PWM length	Positions trimmed	Trimmed PWM length
bcd	6	None	6
cad	8	Last 1 position	7
vfl	11	First 3 positions	8
Dstat	11	None	11
gt	14	First 3 positions	11
hb	10	First 3 positions	7
kni	13	Last 2 positions	11
Kr	10	Last 1 positions	9
slp	11	None	11

Table S2. Evaluations of expression predictions from DNA shape-based model and PWM-based model. The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from DNA shape-based model and PWM-based model over all 37 enhancers are shown, and changes of wPGP scores greater than 0.05 are identified.

Enhancer	DNA shape-based model	PWM-based model	Change > 0.05
Kr_CD1_ru	0.76	0.50	+
ftz_+3	0.80	0.57	+
hb_centr_&_post	0.56	0.38	+
run_stripe1	0.93	0.76	+
eve_37ext_ru	0.96	0.80	+
slp2_(-3)	0.88	0.76	+
kni_83_ru	0.88	0.78	+
eve_1_ru	0.85	0.75	+
nub_(-2)	0.88	0.77	+
eve_stripe2	0.60	0.50	+
kni_(+1)	0.86	0.76	+
eve_stripe5	0.91	0.82	+
gt_(-10)	0.91	0.84	+
odd_(-5)	0.60	0.53	+
h_15_ru	0.63	0.59	
hb_anterior_actv	0.80	0.76	
h_stripe34_rev	0.70	0.68	
prd_+4	0.85	0.84	
run_stripe3	0.90	0.89	
odd_(-3)	0.69	0.69	
eve_stripe4_6	0.86	0.86	
knrl_(+8)	0.68	0.69	
btd_head	0.91	0.93	
run_-17	0.90	0.92	
run_-9	0.92	0.94	
gt_(-1)	0.77	0.80	
pdm2_(+1)	0.76	0.80	
h_6_ru	0.92	0.96	
run_stripe5	0.74	0.79	
Kr_CD2_ru	0.68	0.74	-
Kr_AD2_ru	0.30	0.35	-
cnc_(+5)	0.68	0.75	-
oc_otd_early	0.85	0.92	-
oc_(+7)	0.87	0.95	-
kni_(-5)	0.79	0.87	-
D_(+4)	0.66	0.77	-
gt_(-3)	0.77	0.89	-

Table S3. Evaluations of various models in this study. For each model, shown are the number of free parameters used ("#Pars"), the average wPGP scores from parameter optimization over all 37 enhancers ("Avg. wPGP (Training)"), and the wPGP scores from cross-validation ("Avg. wPGP (CV)"), averaged over ten repeats of cross validation with different (random) definitions of the ten folds. Standard deviations over the ten repeats are also shown.

Model	#Pars	Avg. wPGP (Training)	Avg. wPGP (CV)
Sequence Model			
PWM-based	21	0.755	0.677 ± 0.004
RF-1-mer	22	0.756	0.673 ± 0.014
RF-1-mer+2-mer	22	0.770	0.696 ± 0.012
RF-1-mer+2-mer+3-mer	22	0.765	0.705 ± 0.017
Shape Model			
Shape-based	22	0.784	0.727 ± 0.020
Sequence+Shape Model			
Integrative PWM	22	0.752	0.676 ± 0.011
Integrative Shape	22	0.776	0.727 ± 0.005
RF-Shape+1-mer	22	0.777	0.724 ± 0.013
RF-Shape+1-mer+2-mer	22	0.762	0.696 ± 0.012
RF-Shape+1-mer+2-mer+3-mer	22	0.767	0.708 ± 0.016

Table S4. Comparisons between models in this study. For each pair of models, shown is the p-value of Wilcoxon signed-rank test over ten pairs of average wPGP scores from ten repeats of 10-fold cross-validation. S: shape; m: mer; Inte: integrative.

	1m	1+2m	1+2+3m	Shape	Inte PWM	Inte Shape	S+ 1m	S+ 1+2m	S+ 1+2+3m
PWM	0.421	0.003	0.003	0.003	0.288	0.005	0.003	0.003	0.003
1m		0.003	0.006	0.003	0.341	0.003	0.003	0.003	0.003
1+2m			0.121	0.006	0.003	0.005	0.018	0.003	0.192
1+2+3m				0.008	0.003	0.003	0.006	0.018	0.192
Shape						0.323	0.084	0.003	0.005
Inte PWM						0.003	0.003	0.003	0.003
Inte Shape							0.192	0.003	0.006
S+ 1m								0.003	0.006
S+ 1+2m									0.003

Table S5. Evaluations of expression predictions from long PWM, trimmed PWM, and DNA shape models. The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from PWM-based models and DNA shape-based model over all 37 enhancers are shown.

Enhancer	Long PWM model	Trimmed PWM model	DNA shape model
btd_head	0.83	0.93	0.91
cnc_(+5)	0.31	0.75	0.68
D_(+4)	0.57	0.77	0.66
eve_1_ru	0.76	0.75	0.85
eve_37ext_ru	0.95	0.80	0.96
eve_stripe2	0.67	0.50	0.60
eve_stripe4_6	0.88	0.86	0.86
eve_stripe5	0.84	0.82	0.91
ftz_+3	0.72	0.57	0.80
gt_(-10)	0.83	0.84	0.91
gt_(-1)	0.72	0.80	0.77
gt_(-3)	0.75	0.89	0.77
h_15_ru	0.60	0.59	0.63
h_6_ru	0.90	0.96	0.92
hb_anterior_actv	0.78	0.76	0.80
hb_centr_&_post	0.42	0.38	0.56
h_stripe34_rev	0.67	0.68	0.70
kni_(+1)	0.67	0.76	0.86
kni_(-5)	0.83	0.87	0.79
kni_83_ru	0.77	0.78	0.88
knrl_(+8)	0.53	0.69	0.68
Kr_AD2_ru	0.30	0.35	0.30
Kr_CD1_ru	0.77	0.50	0.76
Kr_CD2_ru	0.68	0.74	0.68
nub_(-2)	0.83	0.77	0.88
oc_(+7)	0.75	0.95	0.87
oc_otd_early	0.90	0.92	0.85
odd_(-3)	0.75	0.69	0.69
odd_(-5)	0.55	0.53	0.60
pdm2_(+1)	0.84	0.80	0.76
prd_+4	0.80	0.84	0.85
run_-17	0.91	0.92	0.90
run_-9	0.92	0.94	0.92
run_stripe1	0.82	0.76	0.93
run_stripe3	0.91	0.89	0.90
run_stripe5	0.78	0.79	0.74
slp2_(-3)	0.66	0.76	0.88

Table S6. Evaluations of expression predictions from higher order k-mer models. The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from integrative PWM-based model and integrative DNA shape-based model over all 37 enhancers are shown.

Enhancer	RF-1-mer	RF-1-mer+ 2-mer	RF-1-mer+ 2-mer+3-mer
btd_head	0.89	0.85	0.87
cnc_(+5)	0.37	0.66	0.64
D_(+4)	0.74	0.66	0.71
eve_1_ru	0.77	0.78	0.86
eve_37ext_ru	0.83	0.90	0.94
eve_stripe2	0.64	0.67	0.73
eve_stripe4_6	0.86	0.83	0.80
eve_stripe5	0.85	0.91	0.82
ftz_+3	0.76	0.78	0.69
gt_(-10)	0.80	0.85	0.82
gt_(-1)	0.75	0.74	0.66
gt_(-3)	0.65	0.79	0.76
h_15_ru	0.67	0.72	0.65
h_6_ru	0.93	0.94	0.85
hb_anterior_actv	0.65	0.72	0.75
hb_centr_&_post	0.43	0.33	0.40
h_stripe34_rev	0.66	0.69	0.65
kni_(+1)	0.78	0.80	0.62
kni_(-5)	0.84	0.85	0.95
kni_83_ru	0.72	0.78	0.81
knrl_(+8)	0.65	0.57	0.59
Kr_AD2_ru	0.34	0.34	0.35
Kr_CD1_ru	0.82	0.74	0.75
Kr_CD2_ru	0.88	0.75	0.77
nub_(-2)	0.81	0.83	0.87
oc_(+7)	0.82	0.84	0.87
oc_otd_early	0.91	0.85	0.90
odd_(-3)	0.61	0.74	0.80
odd_(-5)	0.81	0.73	0.71
pdm2_(+1)	0.66	0.68	0.77
prd_+4	0.85	0.87	0.87
run_-17	0.93	0.94	0.92
run_-9	0.91	0.94	0.88
run_stripe1	0.86	0.83	0.83
run_stripe3	0.87	0.88	0.91
run_stripe5	0.82	0.89	0.73
slp2_(-3)	0.84	0.85	0.83

Table S7. Evaluations of expression predictions from integrative models. The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from integrative PWM-based model and integrative DNA shape-based model over all 37 enhancers are shown.

Enhancer	Integrative PWM-based	Integrative shape-based
btd_head	0.91	0.90
cnc_(+5)	0.29	0.71
D_(+4)	0.66	0.73
eve_1_ru	0.83	0.83
eve_37ext_ru	0.89	0.94
eve_stripe2	0.59	0.82
eve_stripe4_6	0.87	0.87
eve_stripe5	0.86	0.71
ftz_+3	0.60	0.46
gt_(-10)	0.81	0.92
gt_(-1)	0.74	0.79
gt_(-3)	0.87	0.76
h_15_ru	0.68	0.67
h_6_ru	0.96	0.93
hb_anterior_actv	0.73	0.80
hb_centr_&_post	0.41	0.38
h_stripe34_rev	0.70	0.71
kni_(+1)	0.69	0.79
kni_(-5)	0.88	0.84
kni_83_ru	0.74	0.84
knrl_(+8)	0.61	0.72
Kr_AD2_ru	0.35	0.34
Kr_CD1_ru	0.50	0.79
Kr_CD2_ru	0.73	0.71
nub_(-2)	0.79	0.83
oc_(+7)	0.91	0.89
oc_otd_early	0.91	0.91
odd_(-3)	0.65	0.74
odd_(-5)	0.72	0.46
pdm2_(+1)	0.89	0.75
prd_+4	0.85	0.79
run_-17	0.92	0.95
run_-9	0.95	0.90
run_stripe1	0.81	0.91
run_stripe3	0.91	0.93
run_stripe5	0.80	0.87
slp2_(-3)	0.81	0.85

Enhancer	Integrative shape+1-mer	Integrative shape+1-mer+ 2-mer	Integrative shape+1-mer+ 2-mer+3-mer
btd_head	0.85	0.80	0.83
cnc_(+5)	0.57	0.64	0.67
D_(+4)	0.60	0.46	0.74
eve_1_ru	0.84	0.79	0.82
eve_37ext_ru	0.95	0.96	0.93
eve_stripe2	0.58	0.67	0.72
eve_stripe4_6	0.85	0.82	0.84
eve_stripe5	0.72	0.82	0.89
ftz_+3	0.78	0.74	0.64
gt_(-10)	0.85	0.89	0.85
gt_(-1)	0.72	0.71	0.65
gt_(-3)	0.80	0.76	0.75
h_15_ru	0.69	0.69	0.69
h_6_ru	0.95	0.90	0.89
hb_anterior_actv	0.80	0.72	0.80
hb_centr_&_post	0.47	0.55	0.37
h_stripe34_rev	0.69	0.75	0.66
kni_(+1)	0.86	0.76	0.62
kni_(-5)	0.83	0.82	0.89
kni_83_ru	0.86	0.79	0.80
knrl_(+8)	0.78	0.53	0.58
Kr_AD2_ru	0.31	0.66	0.34
Kr_CD1_ru	0.78	0.80	0.74
Kr_CD2_ru	0.73	0.77	0.72
nub_(-2)	0.84	0.77	0.85
oc_(+7)	0.86	0.85	0.87
oc_otd_early	0.92	0.91	0.89
odd_(-3)	0.77	0.68	0.77
odd_(-5)	0.67	0.61	0.72
pdm2_(+1)	0.70	0.54	0.77
prd_+4	0.83	0.79	0.84
run_-17	0.87	0.92	0.94
run_-9	0.92	0.90	0.90
run_stripe1	0.90	0.80	0.84
run_stripe3	0.95	0.91	0.90
run_stripe5	0.81	0.86	0.88
slp2_(-3)	0.80	0.86	0.80

REFERENCES

1. Shea, M.A. and Ackers, G.K. (1985) The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation. *Journal of molecular biology*, **181**, 211-230.
2. Pujato, M., Kieken, F., Skiles, A.A., Tapinos, N. and Fiser, A. (2014) Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic acids research*, gku1228.
3. Bailey, T.L. and Elkan, C. (1994), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Vol. 2, pp. 28-36.