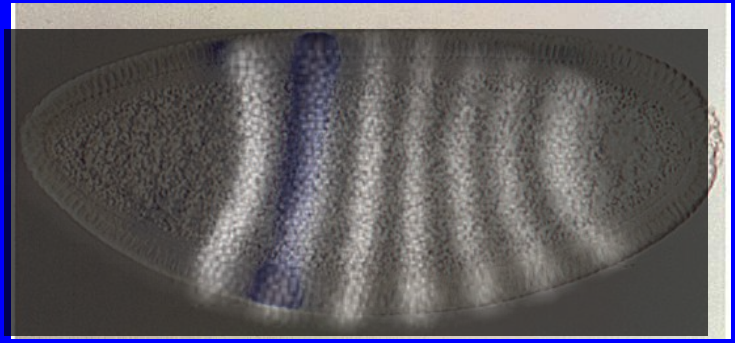
COMPUTATIONAL FRAMEWORKS FOR UNDERSTANDING THE FUNCTION AND EVOLUTION OF DEVELOPMENTAL ENHANCERS IN DROSOPHILA

Saurabh Sinha,

Dept of Computer Science, University of Illinois

Cis-regulatory modules (enhancers)

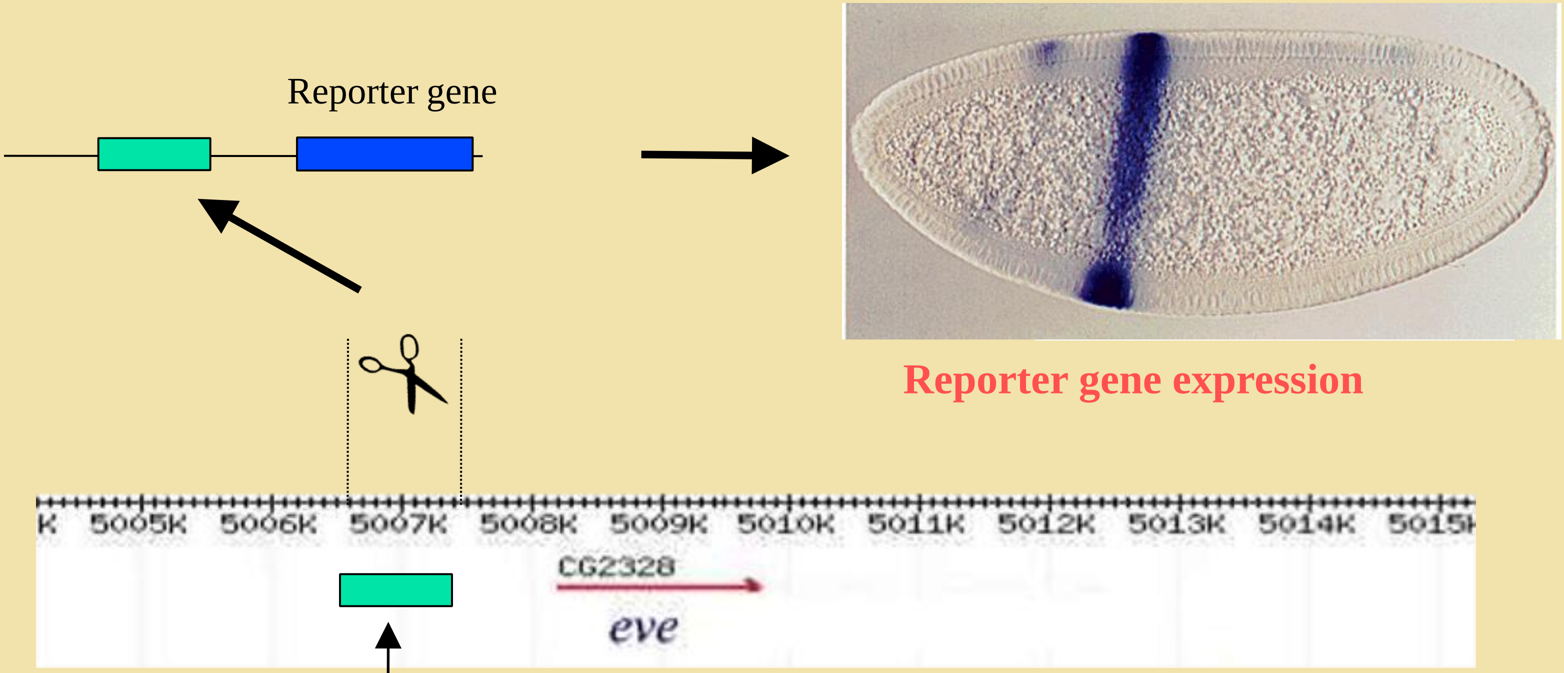
Even-skipped ("eve") gene expressed in seven stripes in the trunk region



Different stripes driven by different cis-regulatory sequences

"Eve stripe 2"

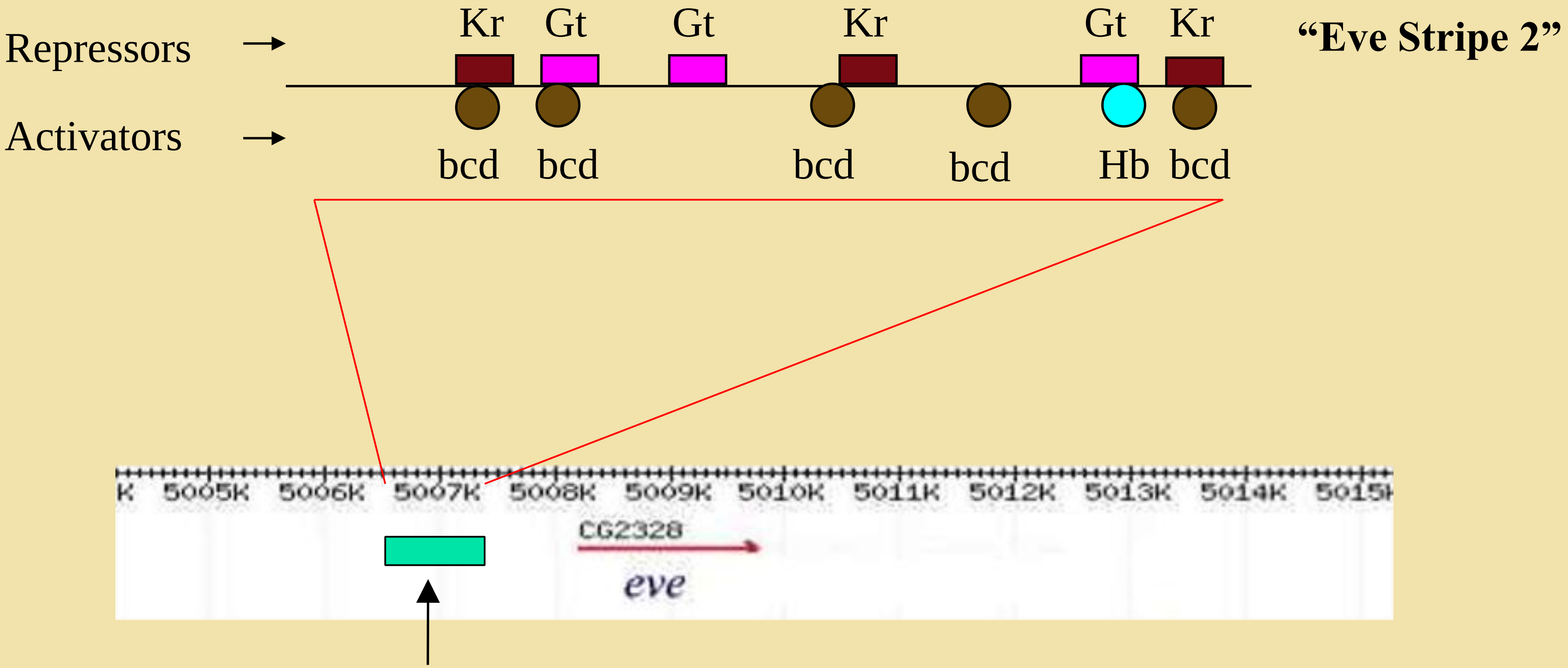
Cis-regulatory modules



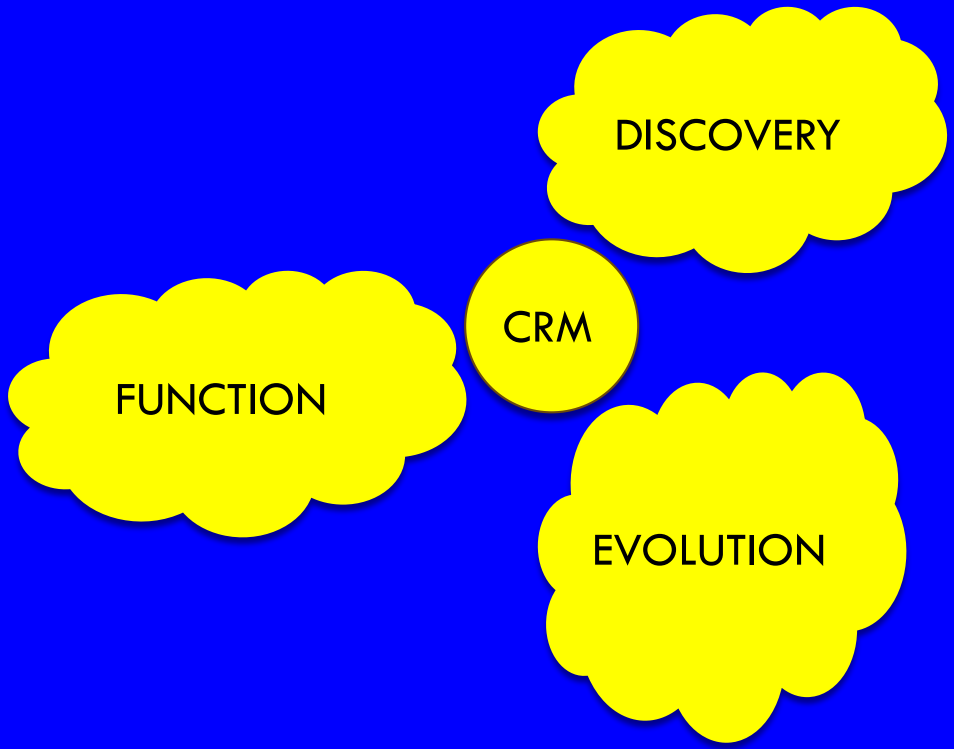
Reporter gene expression

Regulatory sequence associated with eve Stripe 2

Cis-regulatory modules



Regulatory sequence associated with eve Stripe 2



DISCOVERY

CRM

FUNCTION

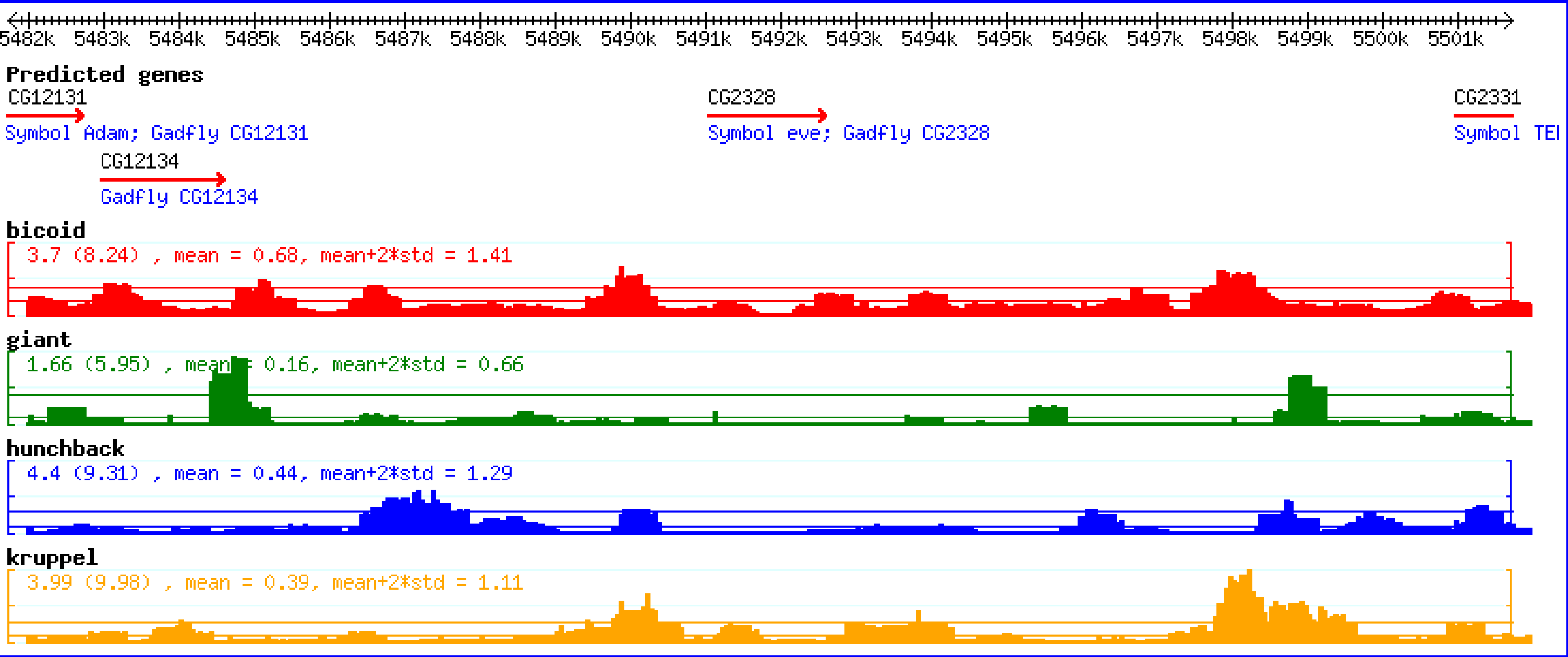
EVOLUTION

Drosophila Genome Surveyor

Genome Browser tracks for motifs for ~ 300 TFs. (HMM-based.)

In each of 12 Drosophila genomes, as well as multi-species averages

Can combine tracks for any subset of motifs (for CRM discovery)



Predicted genes

CG12131
Symbol Adam; Gadfly CG12131

CG12134
Gadfly CG12134

CG2328
Symbol eve; Gadfly CG2328

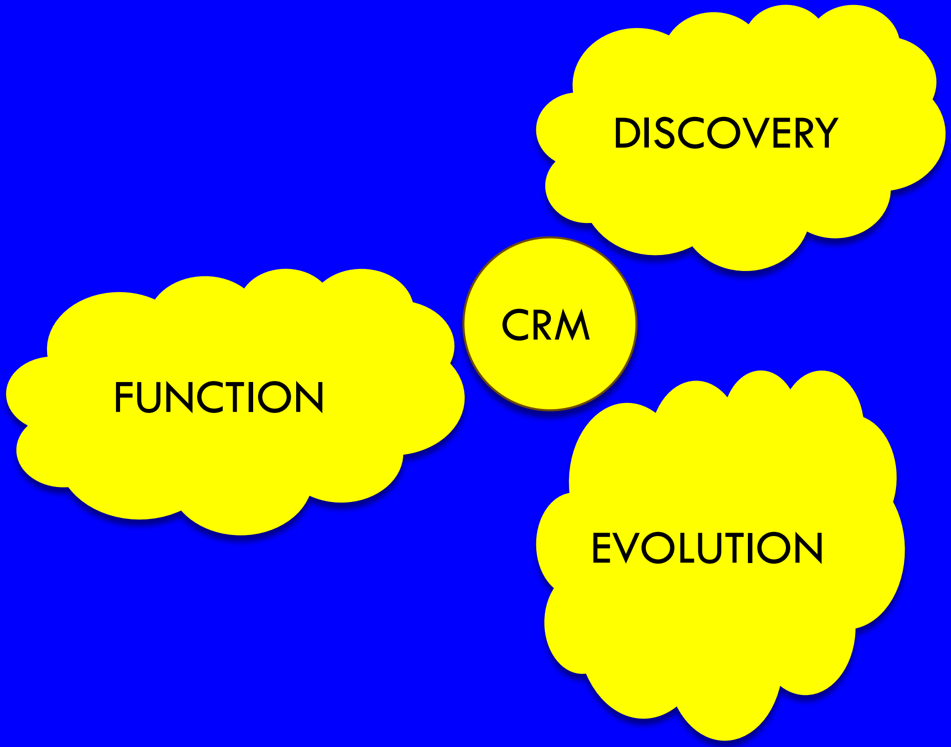
CG2331
Symbol TEI

bicoid
3.7 (8.24) , mean = 0.68, mean+2*std = 1.41

giant
1.66 (5.95) , mean = 0.16, mean+2*std = 0.66

hunchback
4.4 (9.31) , mean = 0.44, mean+2*std = 1.29

kruppel
3.99 (9.98) , mean = 0.39, mean+2*std = 1.11



DISCOVERY

CRM

FUNCTION

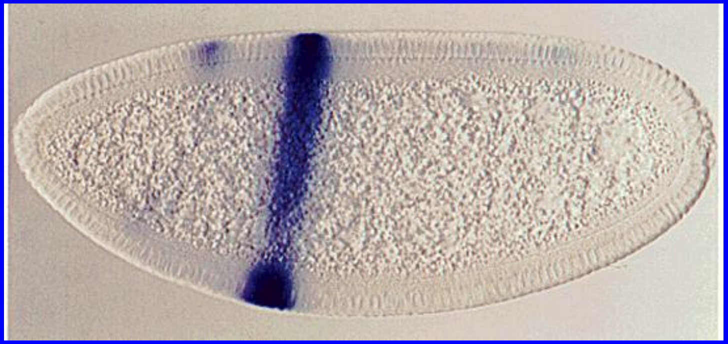
EVOLUTION

PLoS Biology 2010
PLoS Comp Bio 2010

CRM Function

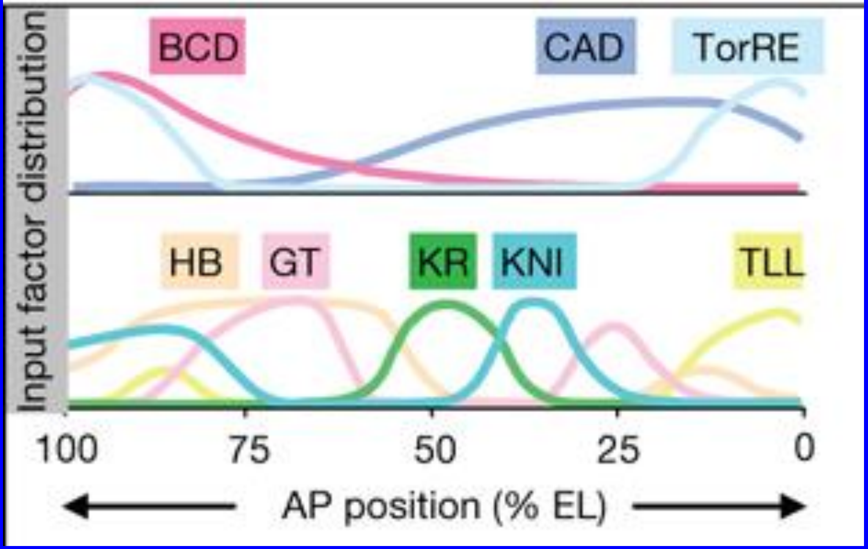
- □ How do we go from sequence to expression?

ACGGATCGACA….CGACGACGATCG



PLoS Biology 2010
PLoS Comp Bio 2010

Well, we'll use more than just the sequence

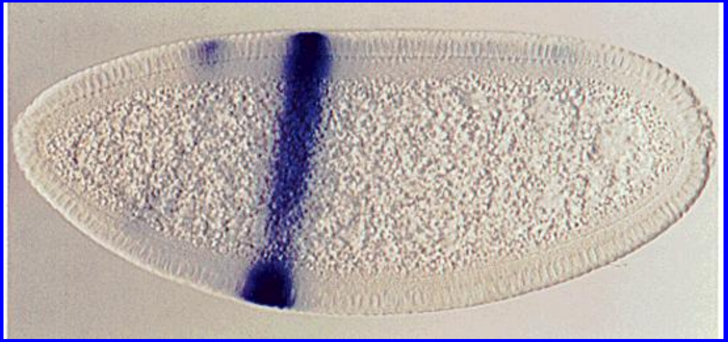


Assume that TF concentration profile known

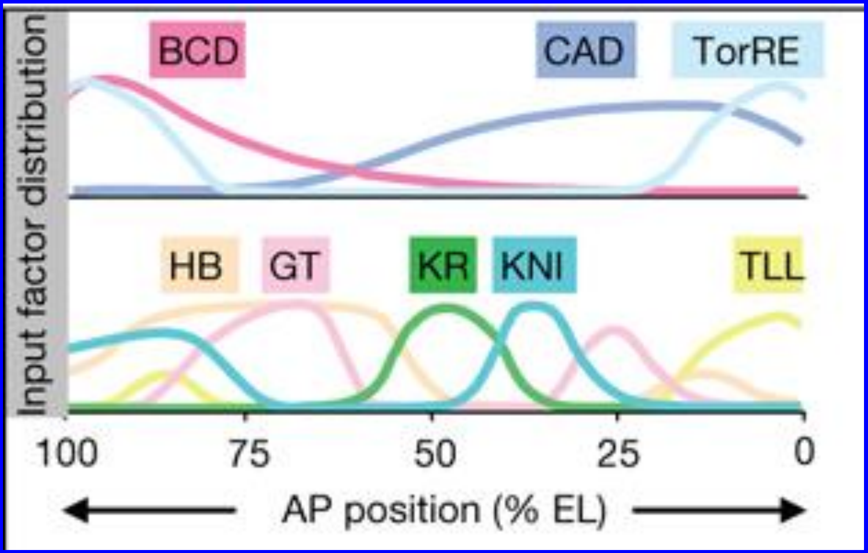
Segal et al. Nature 451, 535-540(31 January 2008)doi:10.1038/nature06496

+

ACGGATCGACA….CGACGACGATCG
(eve stripe 2 CRM)



The A/P patterning regulatory network

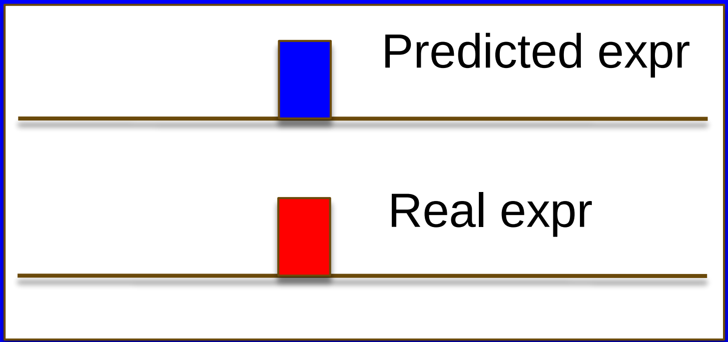


Assume that TF concentration profile known

Segal et al. Nature 451, 535-540(31 January 2008)doi:10.1038/nature06496

+

ACGGATCGACA….CGACGACGATCG
(eve stripe 2 CRM)



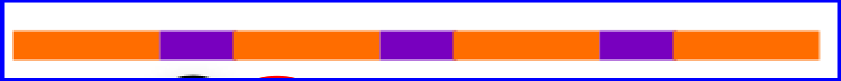
Statistical Thermodynamics-based models

- Shea & Ackers (1985). "*The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation.*" J Mol Biol 181: 211–230.
- Buchler NE, Gerland U, Hwa T (2003). "*On schemes of combinatorial transcription logic*". Proc Natl Acad Sci U S A 100: 5136–5141.
- Gertz J, Siggia ED, Cohen BA (2009). "*Analysis of combinatorial cis-regulation in synthetic and genomic promoters*". Nature 457: 215–218.
- This is what our framework will be based on.

Statistical Thermodynamics-based models

- Other quantitative models:
- Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, et al. (2006). Nat Genet 38: 1159–1165.
- Zinzen RP, Papatsenko D (2007). PLoS Comput Biol 3: e84.
- A general purpose software implementation missing.

Model based on equilibrium thermodynamics



CRM with 3 binding sites. Two activator sites and one repressor site.

$2^3 = 8$ possible configurations of bound/unbound factors.

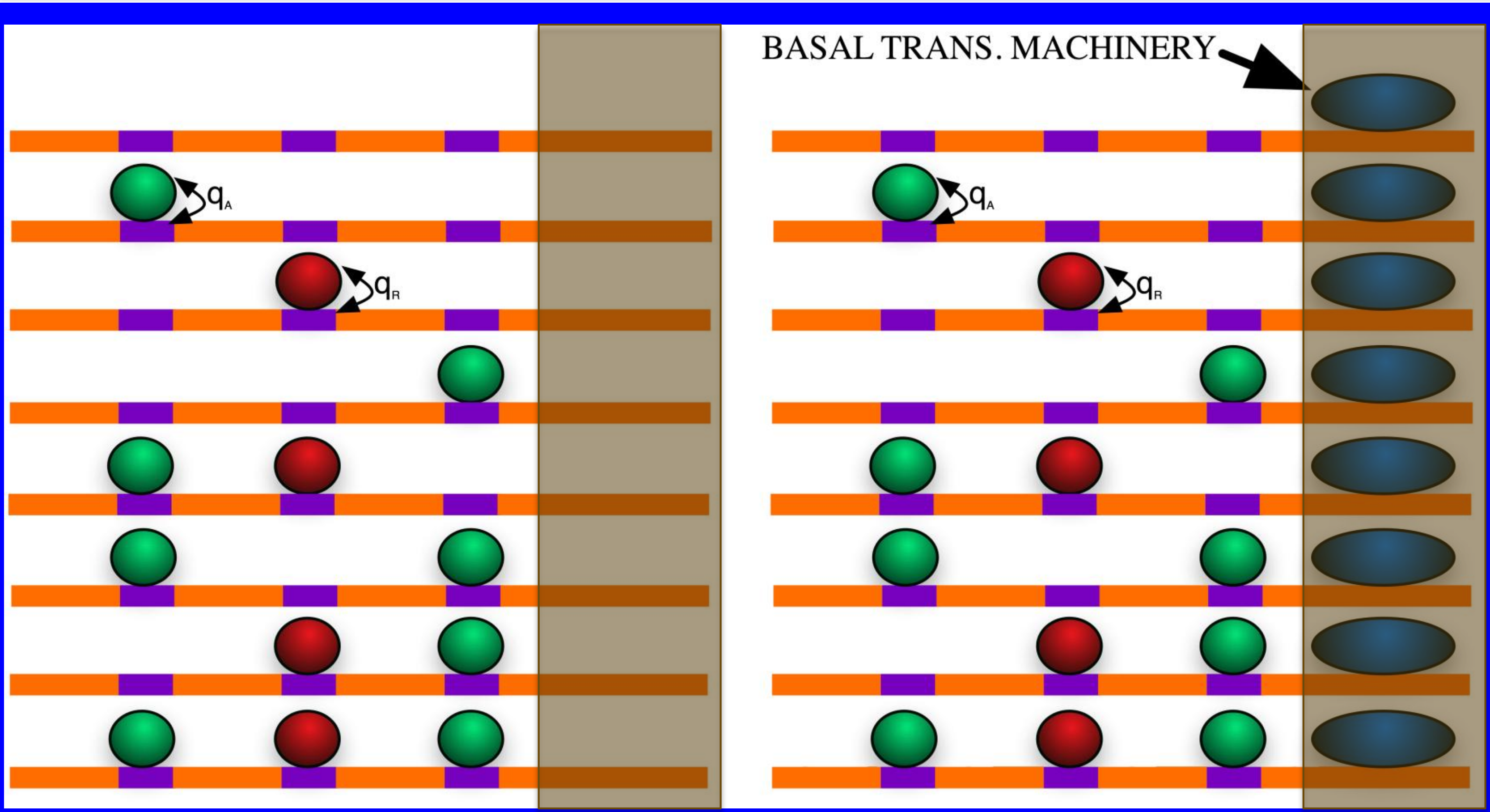
Statistical weight of a bound site (q) given by sequence and TF concentration

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

Statistical weight of a configuration comes from product over bound sites.

Statistical weight = relative probability of a configuration

Modeling Gene Expression

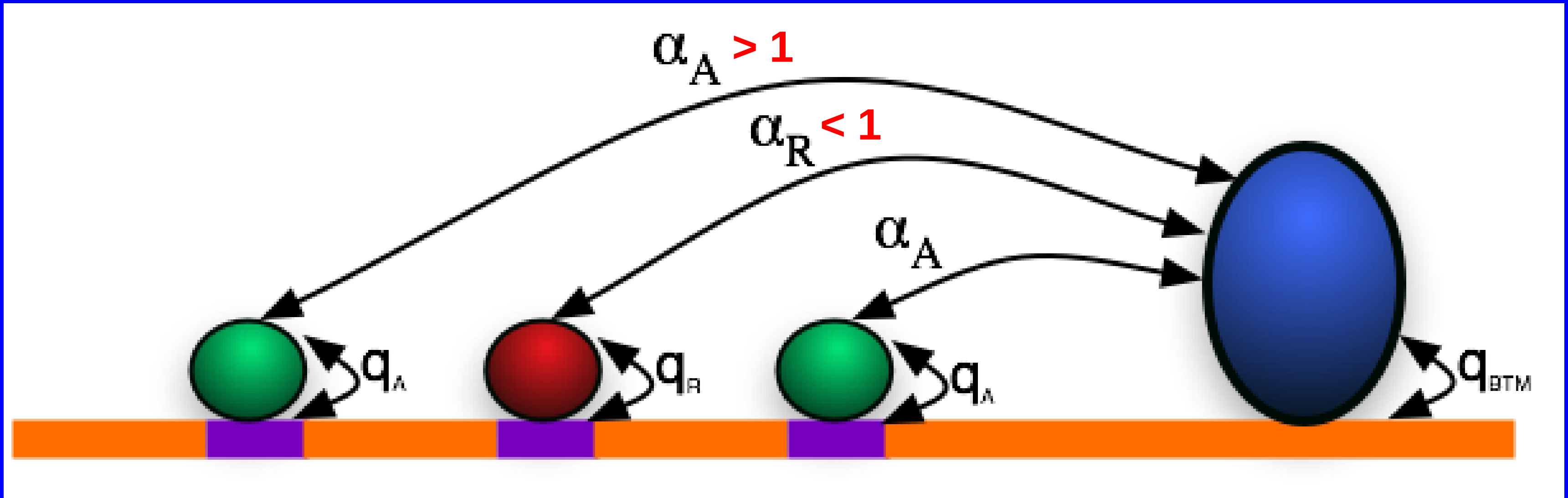


BTM may be bound (at promoter) or not

Gene expression ∝ probability of bound BTM.

Shea &Ackers, 1985
Gertz et al, 2009.

TF effect on gene expression



Each bound trans. factor interacts independently with BTM.

Activators have stabilizing effect. Repressors destabilize.

Implementation

- Two free parameters per TF:
  - one for TF-DNA interaction
  - one for TF's activation or repression strength
- Given these parameters, *relative* TF concentrations and any sequence, compute the predicted expression level (fractional occupancy of BTM) in time proportional to length of sequence. Dynamic programming.
- Note that predicted expression levels are *relative*.

Implementation

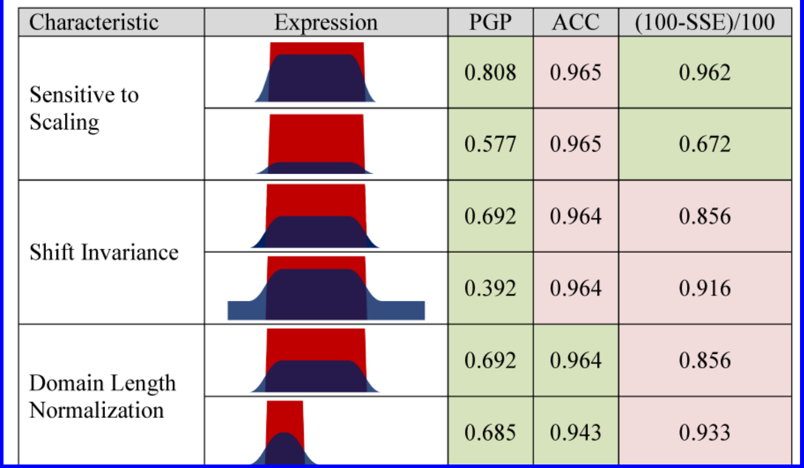
- Given any set of enhancer sequences and their output expression profiles, learn parameter values such that model output best fits data.

- ~40-50 CRMs that drive A/P pattern
- “Pattern” here is the expression in each of ~100 “bins” along the A/P axis
- ~6-10 TFs that are known to be “relevant”

Issues: Objective function

- What does it mean for model output to "fit" data? That is, what is the objective function?
- Sum of squared errors
- Average Correlation coefficient

- Each has problems

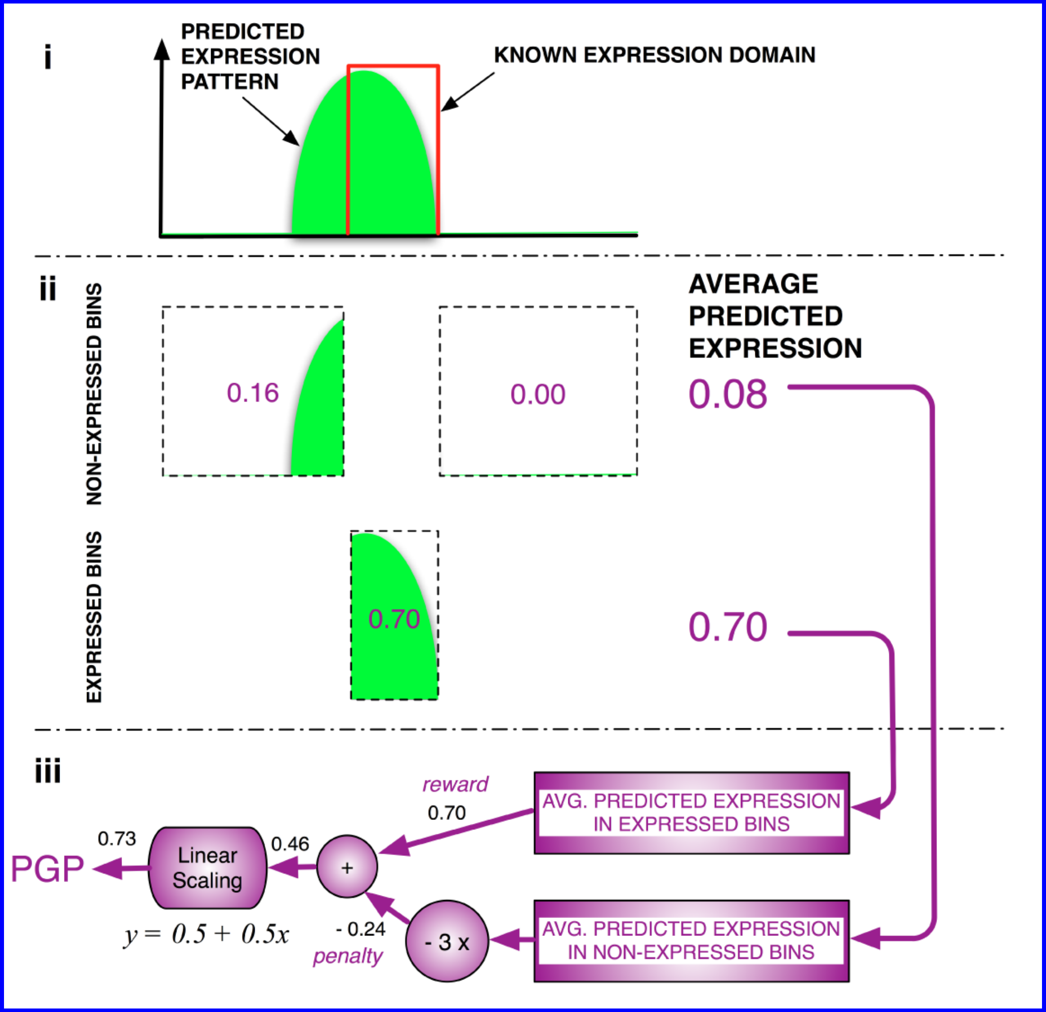


Characteristic	Expression	PGP	ACC	(100-SSE)/100
Sensitive to Scaling		0.808	0.965	0.962
		0.577	0.965	0.672
Shift Invariance		0.692	0.964	0.856
		0.392	0.964	0.916
Domain Length Normalization		0.692	0.964	0.856
		0.685	0.943	0.933

Issues: Objective function

- What does it mean for model output to “fit” data? That is, what is the objective function?
- Objective function is really a subjective choice.
- Public implementation *alternates* between Average CC and RMSE.
- New implementation uses “Pattern Generating Potential” or PGP (from Kazemian et al 2010). We “engineered” an objective function that we were least unhappy about.

Issues: Objective function



Issues: Simultaneous fit to all CRMs

- Important that we fit the parameters to many CRMs simultaneously.
- Generally easy to fit to a single CRM or a handful.
- Therefore need an objective function that can
  - not only tell when one prediction is a better fit than another prediction for the same CRM,
  - But also compare the fit on one CRM to the fit on another CRM.

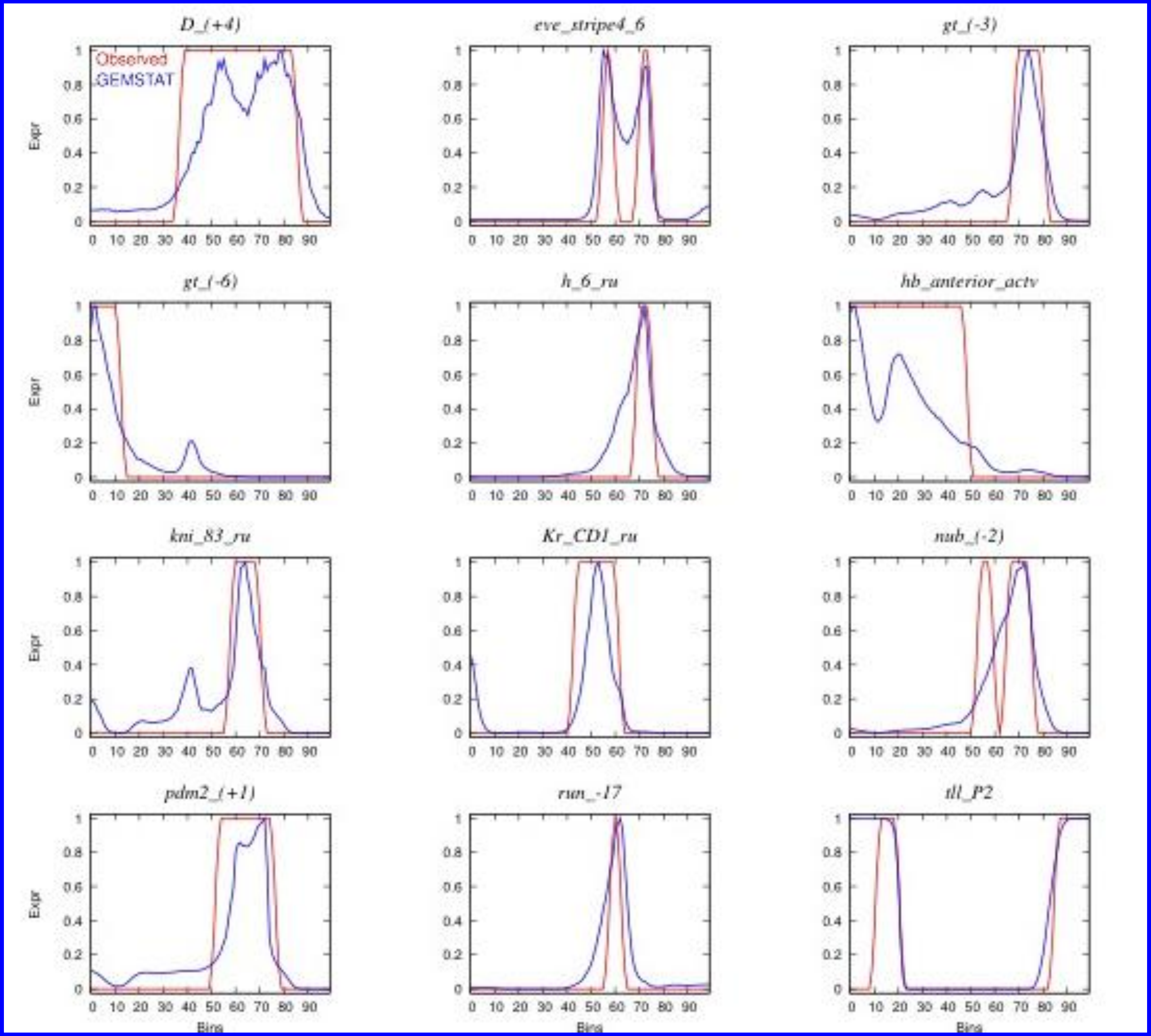
Issues: Simultaneous fit to all CRMs

- Important that we fit the parameters to many CRMs simultaneously.
- In practice, there will be some (or many) CRMs for which we are missing key TFs, or CRMs that are "weird".
- But a simultaneous fit will try to find parameters that produce best fits *overall*. Perhaps we'd like to allow the optimization to "pass" on some (or many) CRMs of its choice. We do that now.

Issues: Optimization algorithm

- Tried a few things; public implementation uses a combination of a gradient descent method and a simplex algorithm.
- Also tried an “evolutionary strategy”
- On real data, similar results; on realistic but simulated data, similar results.
- Also, a fairly exhaustive search of parameter space done before choosing 1000 best “starting points”

Visuals of model fits (some good ones)



Mechanistic inferences?

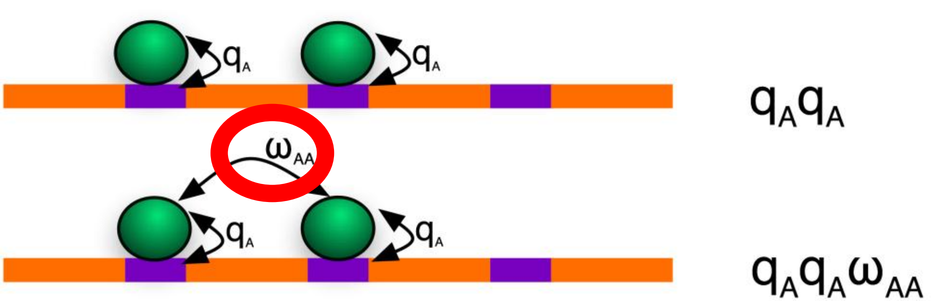
- Test if particular mechanistic aspects improve the fit of model to data. For example, the model I described vs model that includes short range repression.

Comparing model fits

- Compare the optimized objective function under each of the two models
- Sum of squared errors
- Average correlation coefficient
- PGP
- Same, but under cross validation, if models differ in complexity
- Statistical significance of the difference?

Testing mechanistic aspects

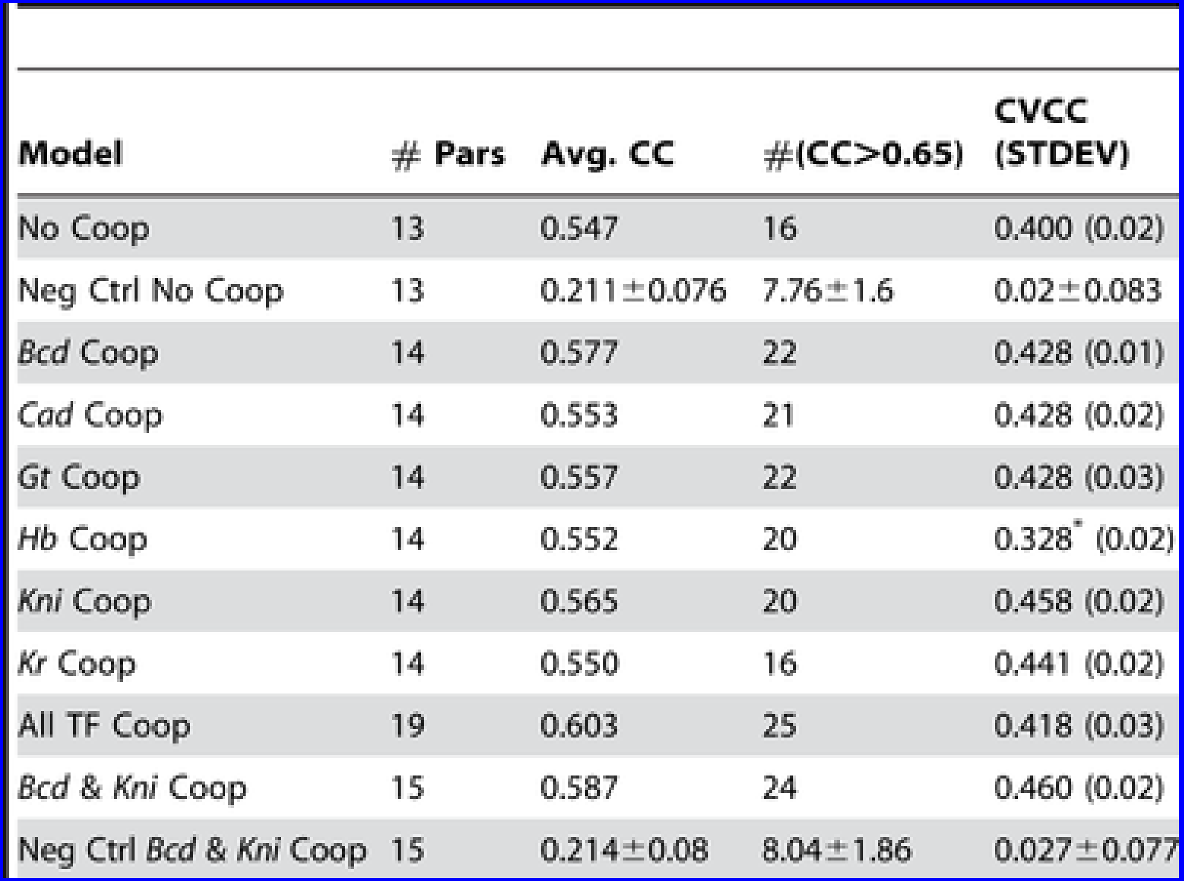
- Effect of cooperative DNA binding by pairs of TFs



Cooperativity in DNA-binding (between adjacent bound trans. factors) contributes a term ω to the weight of a configuration

Testing mechanistic aspects

- Effect of cooperative DNA binding by pairs of TFs

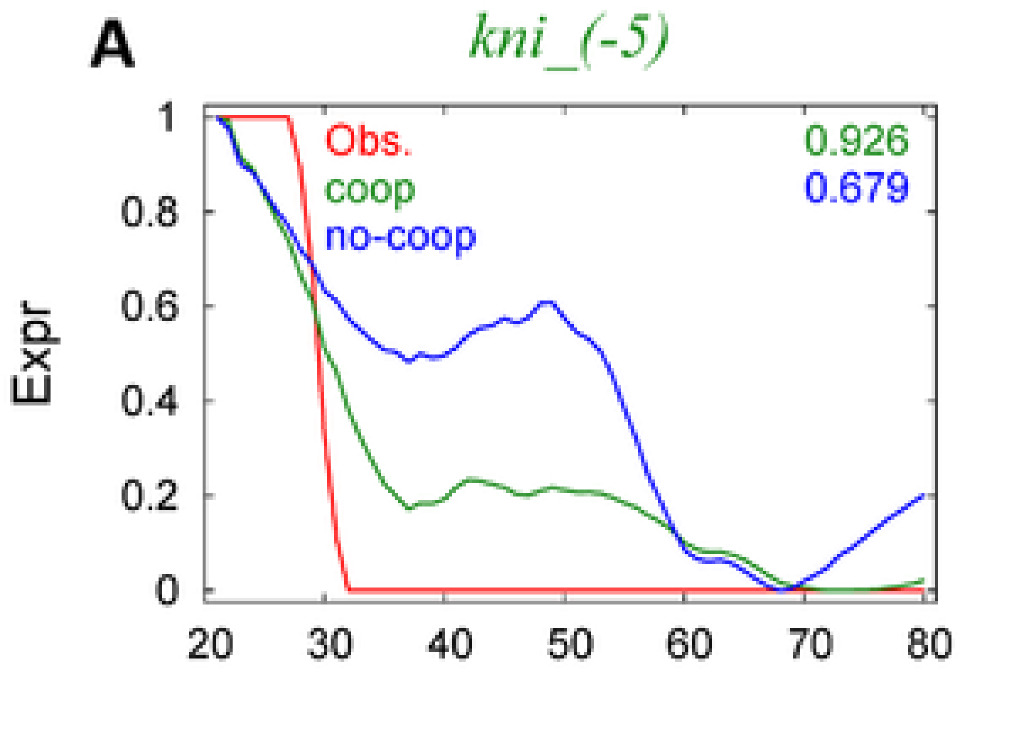


Model	# Pars	Avg. CC	#(CC>0.65)	CVCC (STDEV)
No Coop	13	0.547	16	0.400 (0.02)
Neg Ctrl No Coop	13	0.211±0.076	7.76±1.6	0.02±0.083
Bcd Coop	14	0.577	22	0.428 (0.01)
Cad Coop	14	0.553	21	0.428 (0.02)
Gt Coop	14	0.557	22	0.428 (0.03)
Hb Coop	14	0.552	20	0.328* (0.02)
Kni Coop	14	0.565	20	0.458 (0.02)
Kr Coop	14	0.550	16	0.441 (0.02)
All TF Coop	19	0.603	25	0.418 (0.03)
Bcd & *Kni* Coop	15	0.587	24	0.460 (0.02)
Neg Ctrl *Bcd* & *Kni* Coop	15	0.214±0.08	8.04±1.86	0.027±0.077

BCD and KNI self cooperativity helps

Testing mechanistic aspects

- Effect of cooperative DNA binding by pairs of TFs

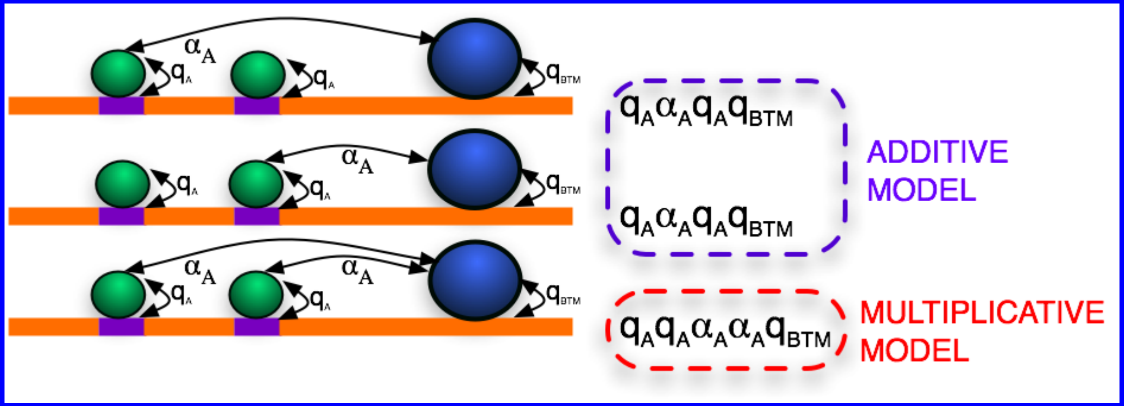


Back to model comparison

- Typically, both models do similarly on many CRMs, one does better on some, the other does better on some others. Comparing overall quality of fit often misses the mark.
- Compare fits on each CRM separately, quantify as a p-value, see if a significant number of CRMs have a significant improvement under one model vs another.

Testing mechanistic aspects

- Effect of synergistic activation
  - multiple bound activators simultaneously contacting the basal transcriptional machinery



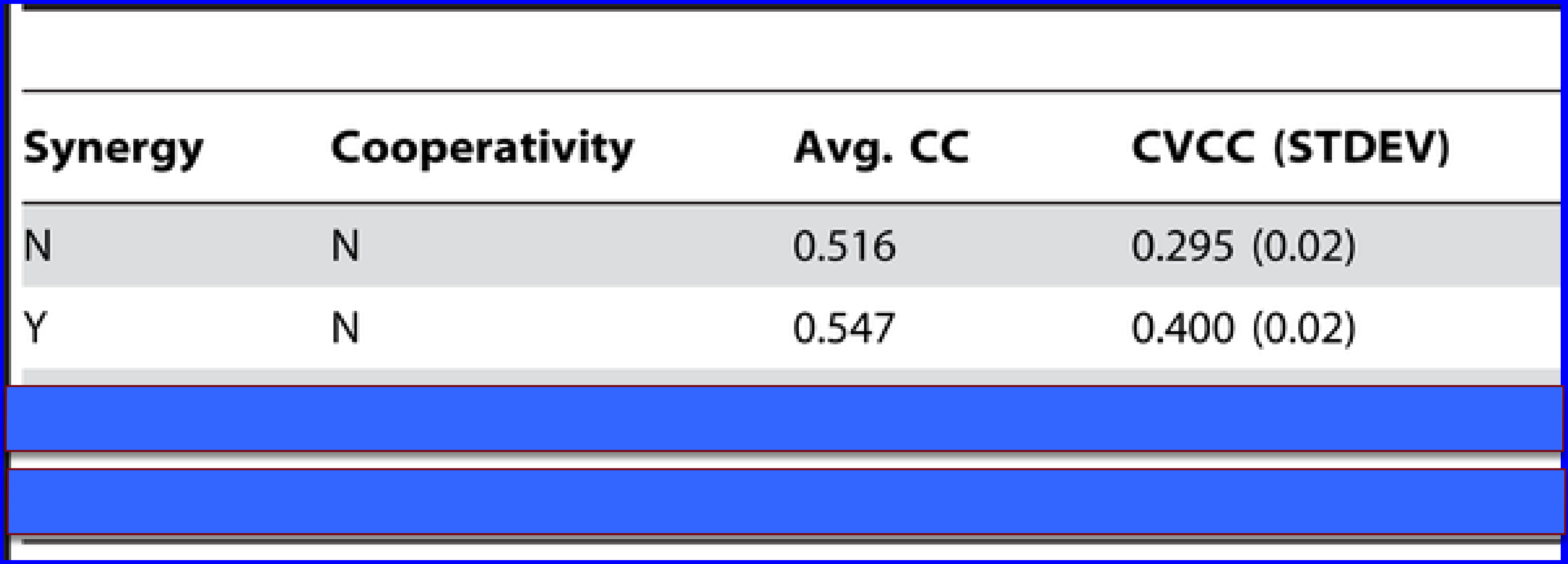
With two bound activators, there are two possibilities:

1) Both interact simultaneously with the BTM: leads to “synergistic” activation

2) Only one interacts with the BTM at a time: no synergy

Testing mechanistic aspects

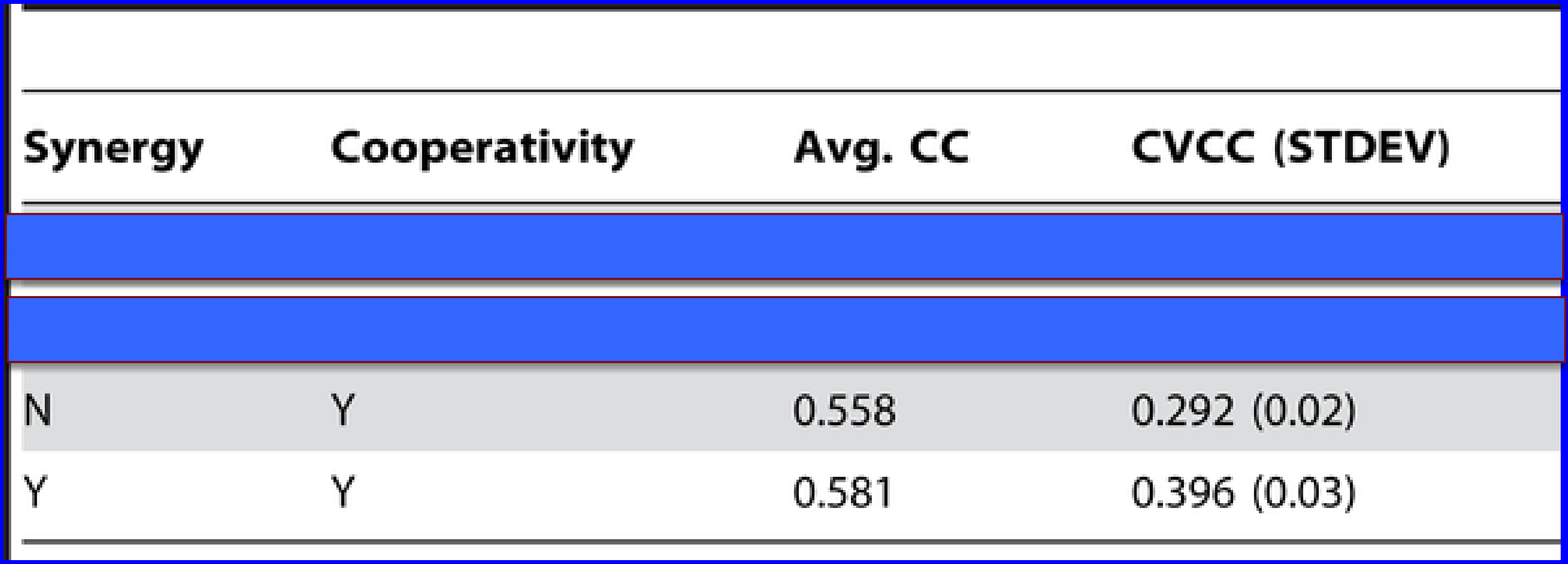
- Effect of synergistic activation
  - multiple bound activators simultaneously contacting the basal transcriptional machinery



Synergy	Cooperativity	Avg. CC	CVCC (STDEV)
N	N	0.516	0.295 (0.02)
Y	N	0.547	0.400 (0.02)
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

Testing mechanistic aspects

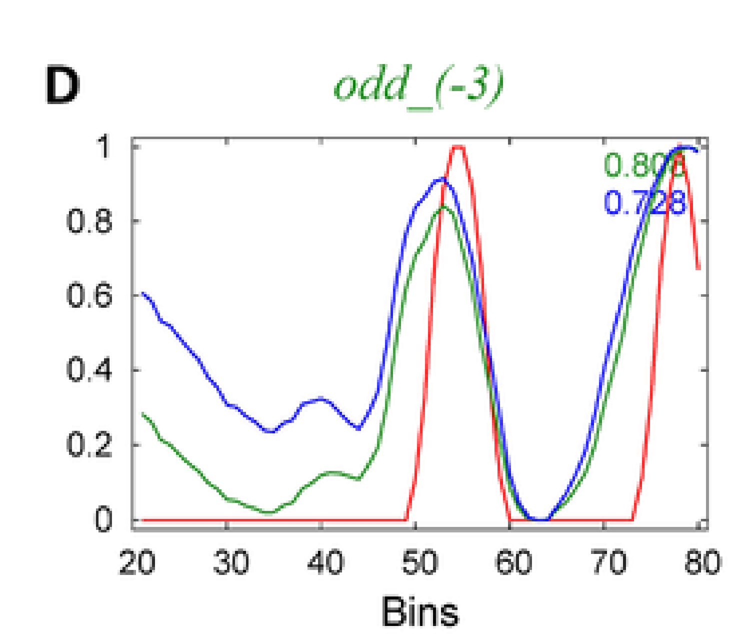
- Effect of synergistic activation
  - multiple bound activators simultaneously contacting the basal transcriptional machinery



Synergy	Cooperativity	Avg. CC	CVCC (STDEV)
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
N	Y	0.558	0.292 (0.02)
Y	Y	0.581	0.396 (0.03)

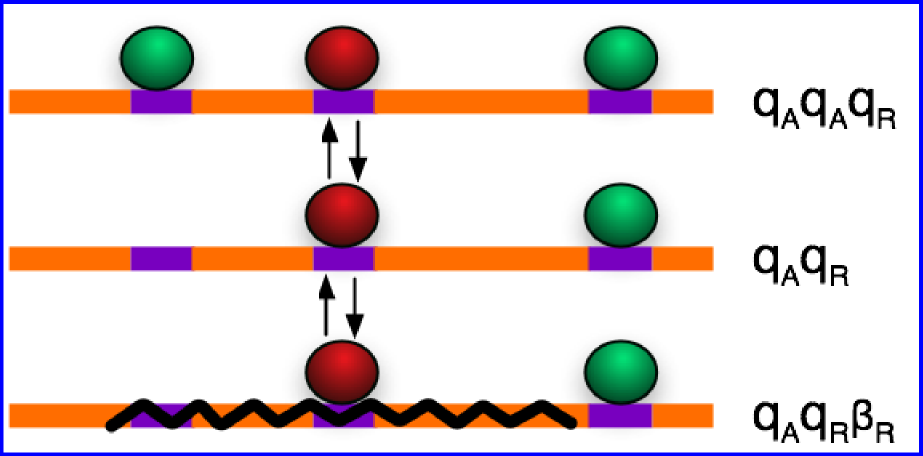
Testing mechanistic aspects

- Effect of synergistic activation
  - multiple bound activators simultaneously contacting the basal transcriptional machinery



Testing mechanistic aspects

- Effect of short range repression



Repressor will work without direct interaction with BTM

If bound, creates a new configuration where its locality is rendered "inaccessible" to other factors

- For KR, HB: short range repression model as effective as the baseline model

Issue: missing TFs

- An effect that is really due to a missing TF may be incorrectly assigned due to a mechanistic aspect in a model.
- So we need to be most diligent about including the relevant TFs.

<slides deleted>

Issue: Gene locus modeling

- Trained model can be used to predict expression pattern driven by any CRM sequence
- Ideally, would like to predict gene expression pattern from entire locus
- In this scenario, we don’t know the CRMs in the locus

Issue: Gene locus modeling

- Given the 16 Kbp eve locus, can we predict its pattern correctly ?
- Predicting on the entire 16 Kbp locus as one sequence will not work.

<slides deleted>

Acknowledgements

Collaborators

Drosophila

David Arnosti	**(MSU)**
Michael Brodsky	**(U.Mass. Med)**
Marc Halfon	**(SUNY Buffalo)**
Stas Shvartsman	**(Princeton)**
Scot Wolfe	**(U.Mass. Med)**

Statistical thermodynamics

Eric Siggia	**(Rockefeller)**

Students

Md. A.Hassan Samee
Xin He
Jaebum Kim
Thyago Duque
Majid Kazemian
Charles Blatti

Funding

NSF CAREER

NIH (NIGMS) R01

Dept. of Computer Science, UIUC.

Institute for Genomic Biology, UIUC