A Likeli(hood) Story: Precise Modeling of Transcription Factor-DNA Interaction from High-Throughput Binding Assays

Curtis G. Callan, Jr.
Physics Department
Princeton University

Reporting on work with Justin Kinney and Gasper Tkacik

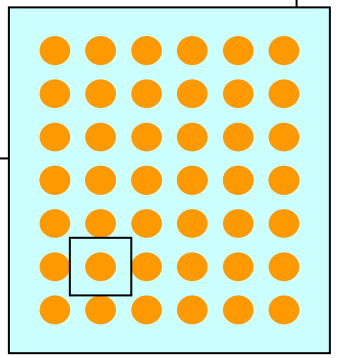
TF-DNA Energy Models from Binding Assays

- Transcription factors (TFs) are DNA-binding proteins which regulate gene transcription: key regulatory mechanism in *all* organisms.
- *A quantitative* understanding of gene regulation and its evolution, requires a *quantitative* understanding of TF-DNA interaction, i.e. sequence-dependent binding energy (SDBE).
- High-throughput experiments can give massive amounts of (rather noisy) information on TF-DNA binding. Popular examples are
  - PBM: protein binding microarrays (*in vitro)*
  - ChIP-chip: chromatin immuno-precipitation microarrays (*in vivo)*
- Usual goal: use the data to identify the TF binding sites (yes-no answer)
- Our goal: infer quantitative SDBE models from this noisy data. We will take the statistical inference approach used in physics to deal with WMAP/HEP data: we seek a probability distribution on model space.

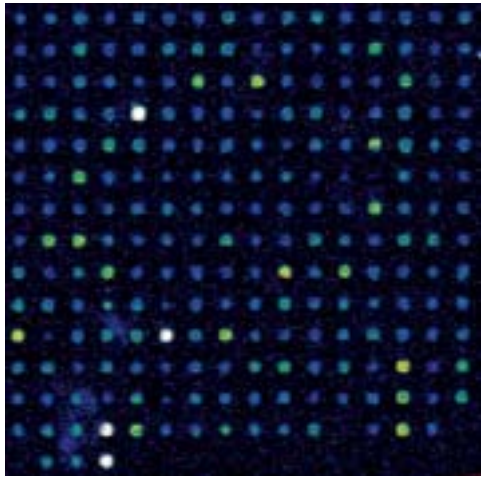
Some Philosophy: Lexical vs. Energetic Approach

- The binding specificity problem has two classic formulations:
  - Lexical: Is there a statistical sequence pattern (motif or pwm) that distinguishes true TF binding sites from “random” genomic background?
  - Energetic: Can we construct an accurate representation of the binding energy of the TF to general site sequences (an SDBE function)?
  - NB: Biological function is determined by energy, not p-value!
- But energy is hard to measure, while sequence is “easy”. Hence, more effort has gone into “motif-finding” than into energetics.
- B+vH algorithm for turning binding site sequences (of one TF) into an energy function (E-matrix) merges the two approaches. But …
  - Assumes that the sites evolve out of random background genome under the same selection pressure (a kind of ergodic hypothesis).
  - The conditions for lexical/energetic equivalence can easily fail (as far as random background goes, just think of Plasmodium!).
- Since binding assay experiments probe energy, it makes sense to try to model energy directly … sequence comes along for the ride.

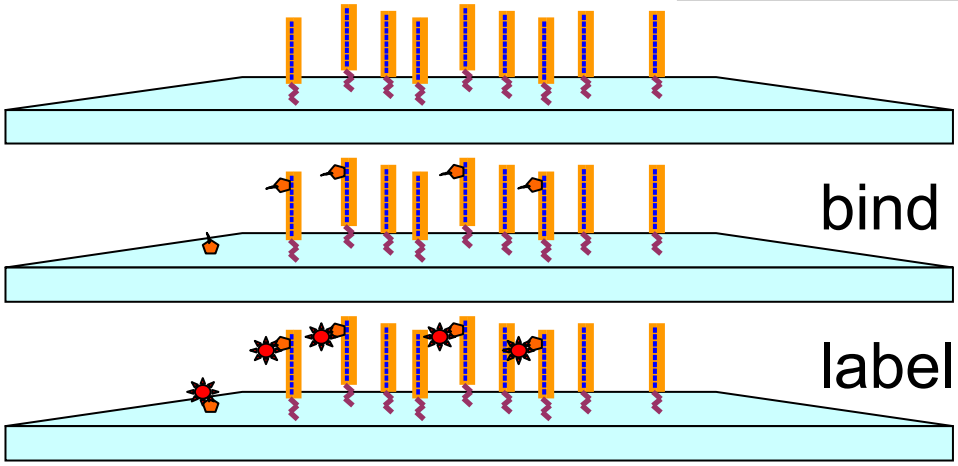
PBM Assay Overview (Mukherjee et al)



- Uses dsDNA microarrays to simultaneously assess TF binding to all intergenic regions of *S. cerevisiae.*



scan



- Fluorescence log-intensity ratios (LIRs) are filtered, averaged over replicates and normalized to taste. Each sequence S_i is assigned some best measured value Z_i (for $i = 1,2, \ldots, N$ intergenic regions).
- Connection between these measured values and whether a TF is bound to the region (or not) is very noisy due to the complicated and loosely-controlled chemistry. How to interpret the data?
- ChIP-chip assay (in vivo) produces similar-looking data.

Simple Binding Energy (and Binding) Model

- Bases within a site (length L) contribute *additively* to the binding energy. Model is a 4xL "energy matrix" *M*.
- A stretch of DNA is "bound" if it contains a site with E<μ (else "unbound"). Step function model of site occupancy.
- A model (*M*, μ) predicts whether any given DNA sequence *s* is bound (x=1) or unbound (x=0).
- How does this compare with what is seen in the experiment? Does any choice of model (*M*, μ) explain the data?

L = 6

Site = TGTGAC

Energy = 0.7 < 1

A	1.2	1.8	5.6	2.5	0.0	6.0
C	3.7	0.0	0.5	1.2	5.2	0.0
G	2.9	0.1	0.8	0.6	1.3	1.4
T	0.0	3.0	0.0	0.0	3.1	3.2

C A T G T G A C C T

Region is "bound"

$$\text{Model } M: \; s_i \mapsto x_i, \quad i = 1, 2, \ldots, N$$

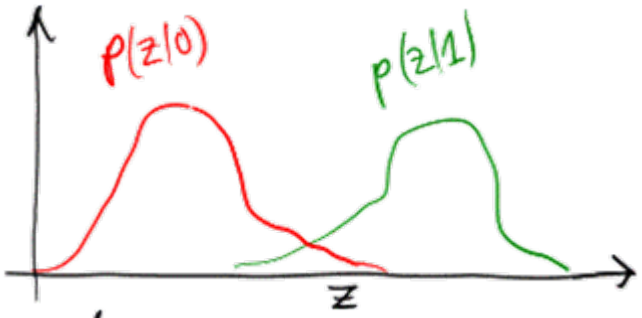
(binary *x*)

(continuous *z*)

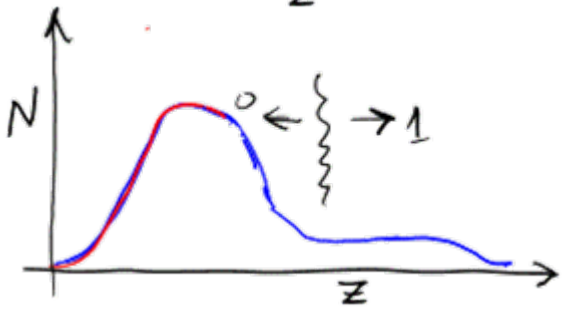
$$\text{Experiment}: \; s_i \mapsto z_i, \quad i = 1, 2, \ldots, N$$

Connecting "Theory" and Experiment:

- Fluorescence *z* of a bound (unbound) region is probabilistic (due to chemistry, etc.). Leads to a ``error models'' for the two states:



- Experiment sees only the histogram of net fluorescence $N(z)=N_0 p(z|0)+N_1 p(z|1)$ due to N_0 "unbound" + N_1 "bound" genes. Usually try to discriminate the two states by a "cut" on *z*.



- How good is model *M* ? If it predicts $\{x_i\}$, likelihood of actual data $\{z_i\}$ is:

$$p(\{z_i\}|M) = \prod_i p(z_i|x_i)$$ product over all regions

- Bayes' Rule then gives the likelihood of the model, given the data:

$$p(M|\{z_i\}) \propto p(M)p(\{z_i\}|M)$$ with model prior p(M)

- This is a prob dist'n on model space and a basis for statistical inference. Good! But … the actual error model is usually totally unknown!

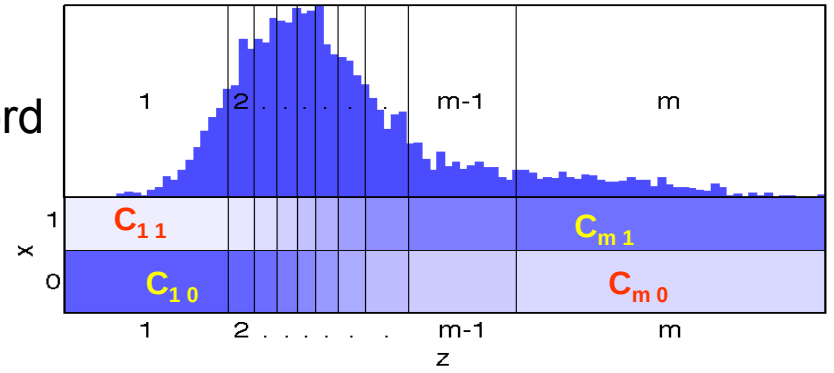
Options for More Sophisticated Modeling

- The energy matrix is just the most simple parametric model. We can allow for correlations if needed. Number of parameters grows …
- Bottlenecking through a binary model datum (bound vs not bound) is a first stab at the problem. We could do more:
  - Use Hwa's thermodynamic model of binding occupancy (K_d , [TF])
  - Parse predicted occupancy into multiple levels x_i (i=1,…N).
  - Analyze in terms of refined error model *p({z}|{x}) …*
- When does the number of parameters to be determined exceed the information content of the data? We don't know, but
- Will show that data (experimental numbers plus the genomic information) on wide-acting yeast TFs fixes a large number of parameters: we have not exhausted its information content.

Quenching the Error Model: EMA Likelihood

- In ignorance of the true error model, we will *average* data likelihood over *all* error models to get an error-model-averaged (EMA) likelihood.
- To actually *do* this average, we need to discretize the continuous data

- Bin each region s_i according to fluorescence (discretize N(z))
- Find predictions $\{x_i\}$ of model M, record counts c_{zx} per bin (divide bins into separate binding populations)
- EMA likelihood is a functional integra



m bins with equal #s of regions ..

each bin splits into two states

$$p(\{z_i\}|M) = \int [\mathcal{D}p(z|x)] \prod_i p(z_i|x_i) = e^{N[I(z;x) - H(z) - \Delta]}$$

Mutual information appears!

- Our binned data yield a simple formula:

$$p(\{z_i\}|\theta) \propto \frac{\prod_{zx} c_{zx}!}{\prod_x (m - 1 + \sum_z c_{zx})!}$$

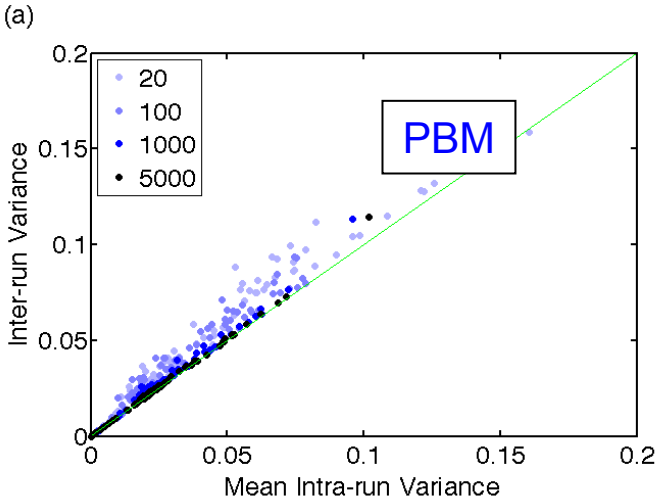
Practical algorithm for evaluating $p(M|\{z_i\})$ (up to normalization!)

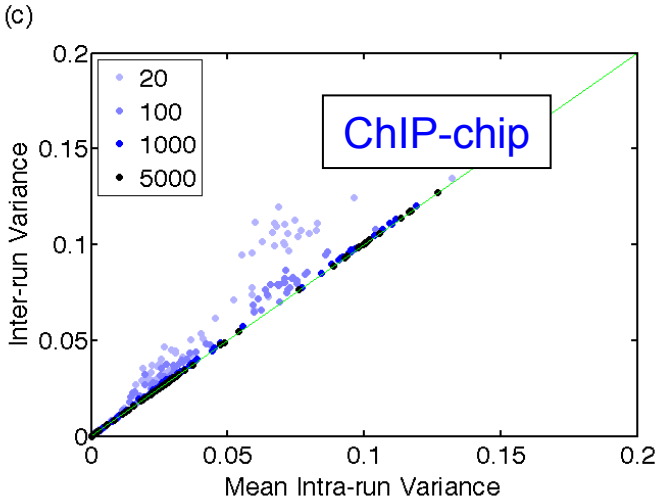
Markov Chain Monte Carlo Evaluation of $p(M|\{z_i\})$

- Let energy matrix elements live in $0 < M_{ib} \leq 1$ and let the cutoff take values in $0 \prec \mu \prec \mu_{\max}$.
- Choose a convenient starting point for the matrix, corresponding to a known motif if possible (to save time only).
- Go through a schedule of trying out small, normally-distributed increments to all matrix elements and the cutoff. Do Metropolis:
  - If increment improves $p(M|\{z_i\})$ (burdensome to compute), accept
  - If increment worsens $p(M|\{z_i\})$, accept with probability p_{old}/p_{new}
  - If increment takes you outside the box, reject and try again
- In long run, get an ensemble $\{M,\mu\}$ distributed according to $p(M|\{z_i\})$
  - Not normalized, but perfect for computing ensemble averages of ….
- At the end, shift and rescale so that lowest matrix element in each column is 0, cutoff μ=1 (leaving model predictions $\{x_i\}$ unchanged).
  - PBM/ChIP-chip data leave the absolute scale of energy undetermined!

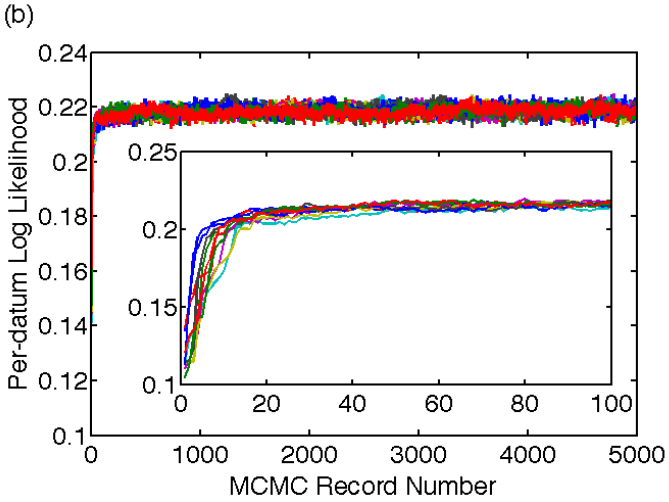
MCMC Estimation Converges *Fast* (for ABF1)

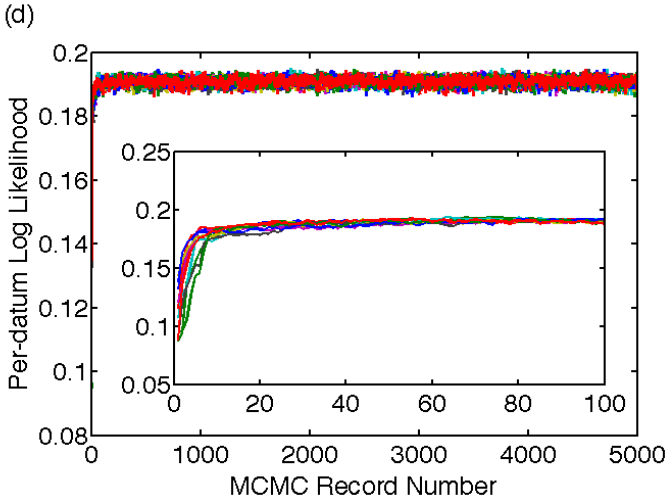
Burn-in test: do 10 runs, plot inter- and intra-run variance for each matrix element for larger and larger samples representing longer run times. Unit slope straight line is convergence signal.





Per-datum log-likelihood rises with MCMC time. Convergence to stable distribution is agreeably rapid.

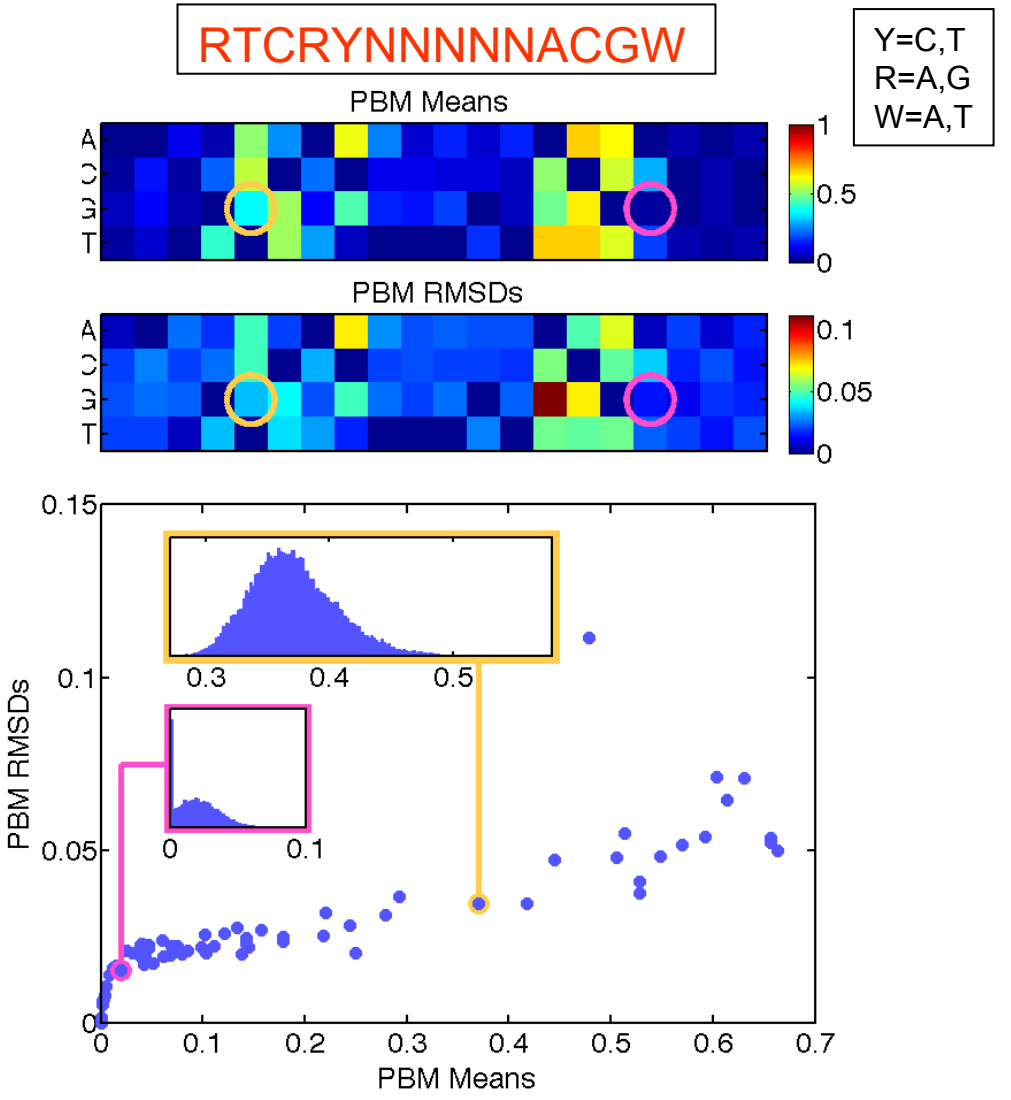




Key result: $p(M|\{z_i\})$ has a single smooth peak, easily found by MCMC!

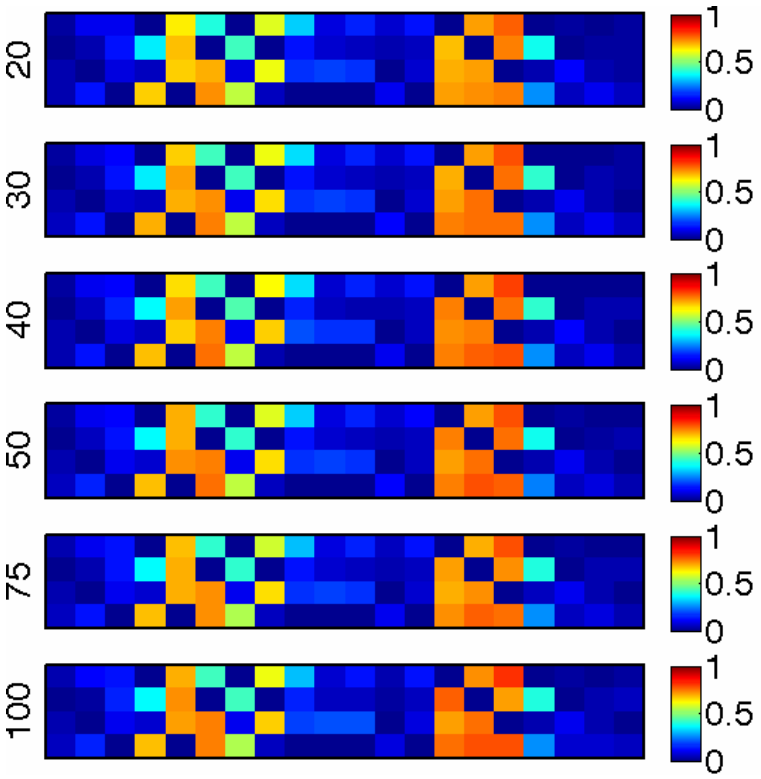
MCMC Results for TF ABF1p (Yeast)

- MCMC generates 40,000 matrices M sampled from $p(M|\{z_i\})$ using EMA likelihood.
- All 80 matrix elements have well-sampled distributions (see insets).
- Mean matrix makes perfect sense in terms of the known motif (more later)
- Distributions are amazingly tight: most RMSDs ≤ 5% of functional range.
- Meaningful structure, even in the middle of the binding site, where there is little specificity.
- That the data imply a smooth probability landscape in the high-dimensional model space is a surprise.
- No one model is the "best" model. We can now treat model predictions as clean probabilistic statements.

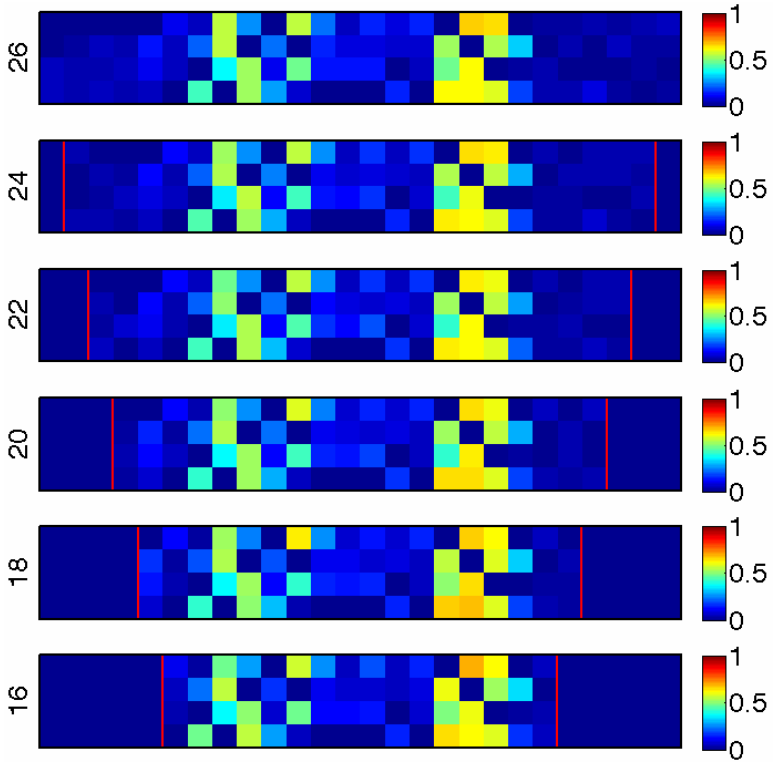


Results invariant to changed analysis parameters

Error model discretisation bin size (20-100 per bin)



Width of energy matrix (in bp)



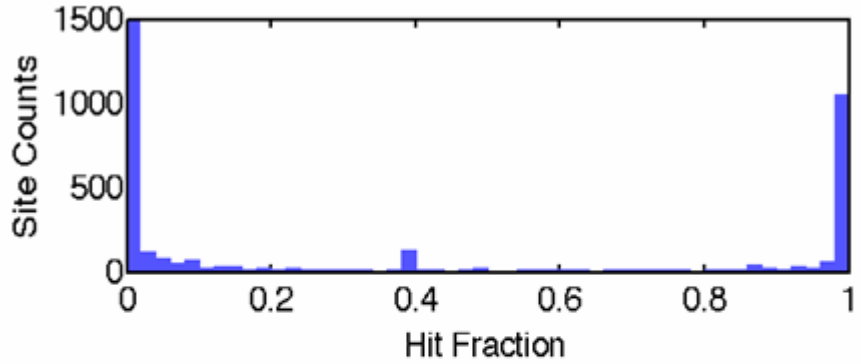
You can even divide the data (intergenic region LIRs) into randomly-chosen halves and compare the two mean energy matrices (overfitting test). They agree very well.

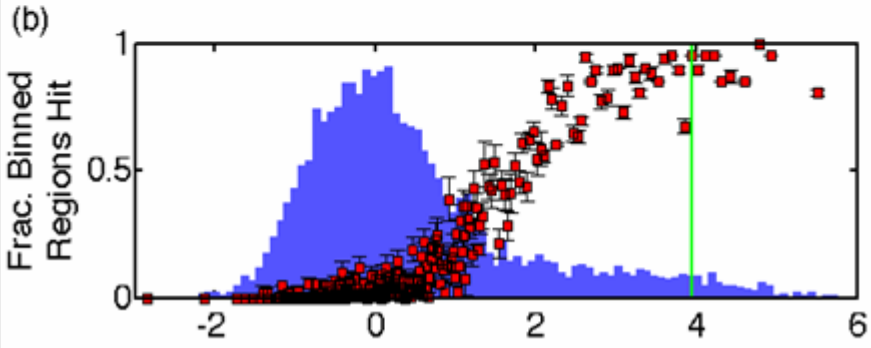
ABF1p MCMC Model Ensemble Predictions

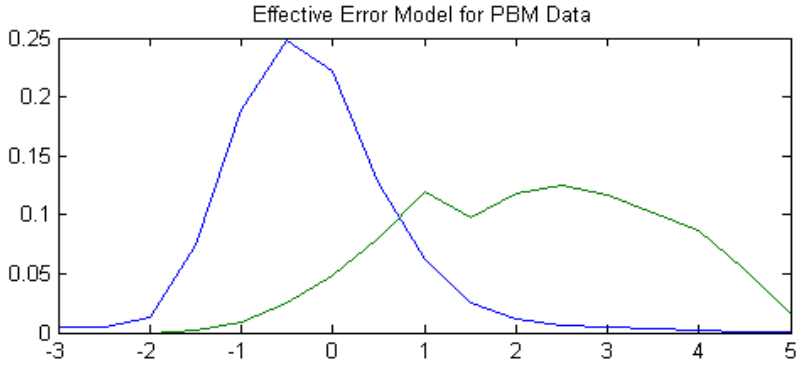
- Ensemble of ABF1p models lets us classify sites by *hit fraction*
- Strongly bimodal hit fraction dist'n cleanly discriminates bound sites
- We find > 1000 sites with h.f. >.5 (and result depends only weakly on cutoff)

- Compare expt'l LIR dist'n with h.f. of binned regions: consistent with credible error model
- Conservative Mukherjee et al LIR cutoff (green line) rejects many regions clearly bound by our criterion.

- Model predictions can be recast as an effective error model: green curve is p(z|1), blue curve is p(z|0) from mean energy model on the data.
- EMA method successfully determines an amazing number of parameters

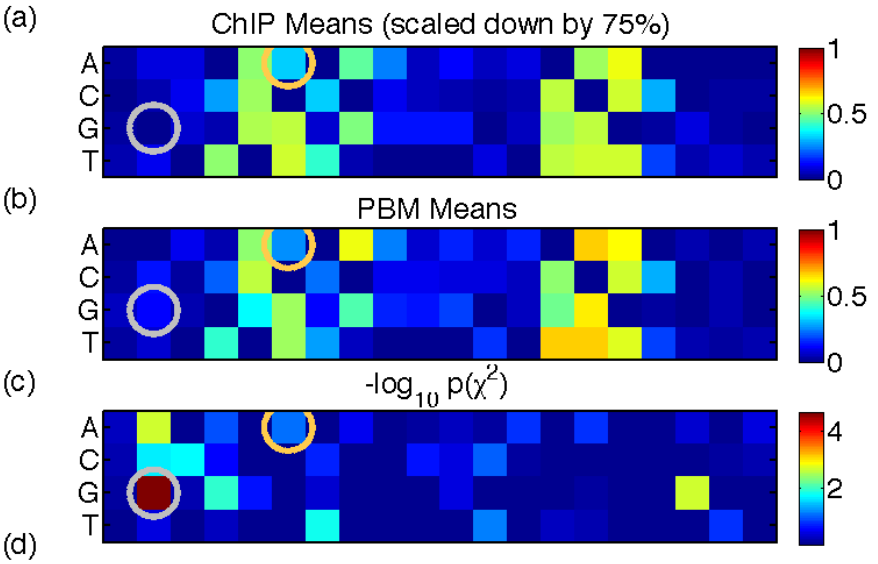


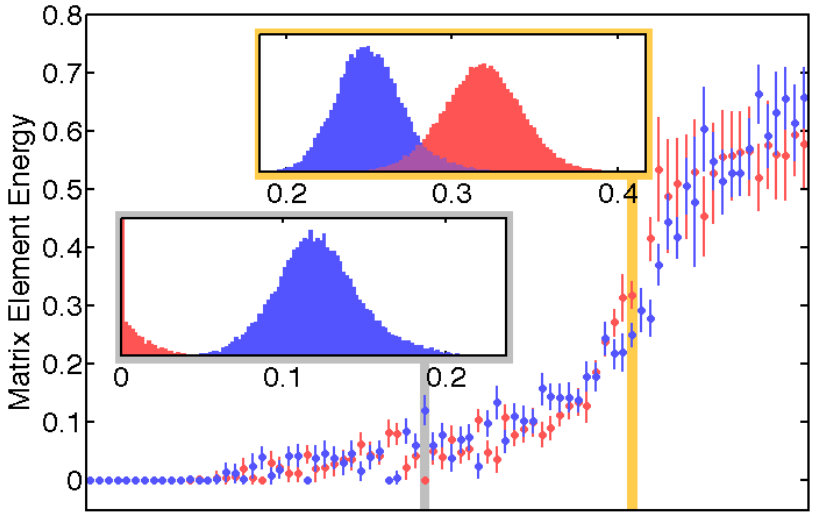




PBM vs. ChIP-chip Data Analysis

- PBM and ChIP-chip data give very similar matrices (but with the ChIP-chip cutoff set to .75 instead of 1).
- Cutoff stands in for the chemical potential of the TF: can vary between experiments (but the energy matrix should not!)
- Simple χ^2 test used to assess the overlap between the PBM and ChIP-chip distributions for each matrix element, i.e. test for consistency.
- Most elements have overlapping distributions. Only 3 don't, and those are outside the main site.
- Element by element match of mean and variance between the two analyses is impressive: No Free Parameters!
- The error models of the two exp'ts (as inferred from the data are very different); that the same energy matrix is inferred in both cases is a strong consistency check.

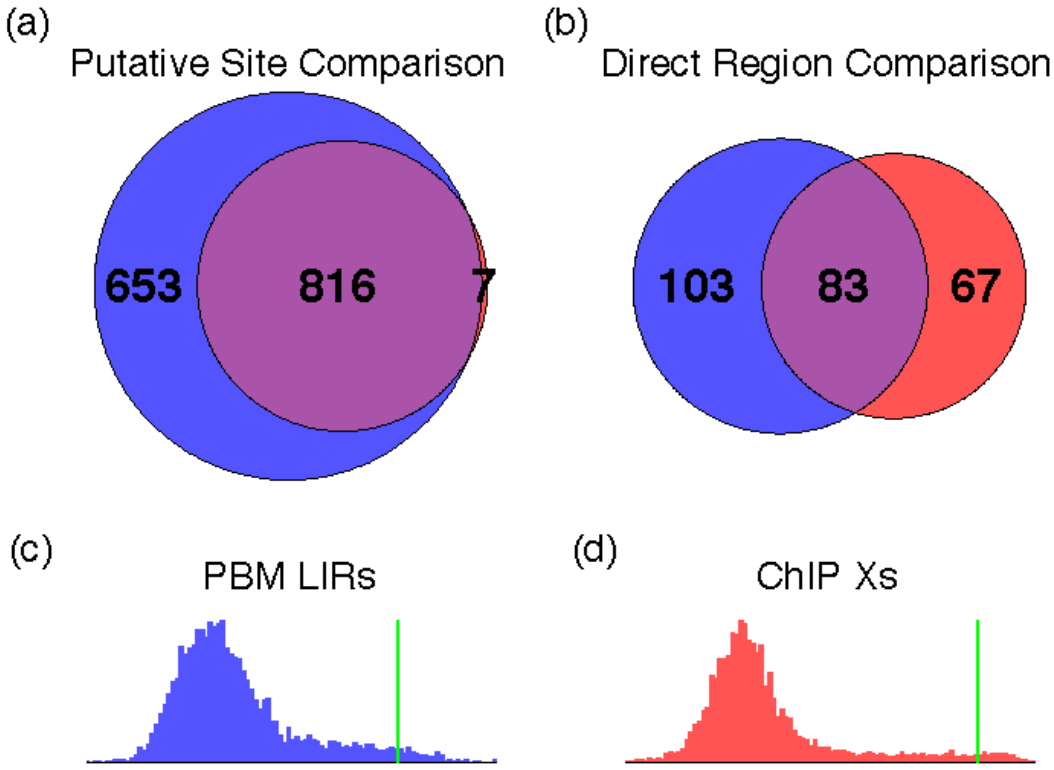




PBM vs. ChIP-chip Binding Predictions

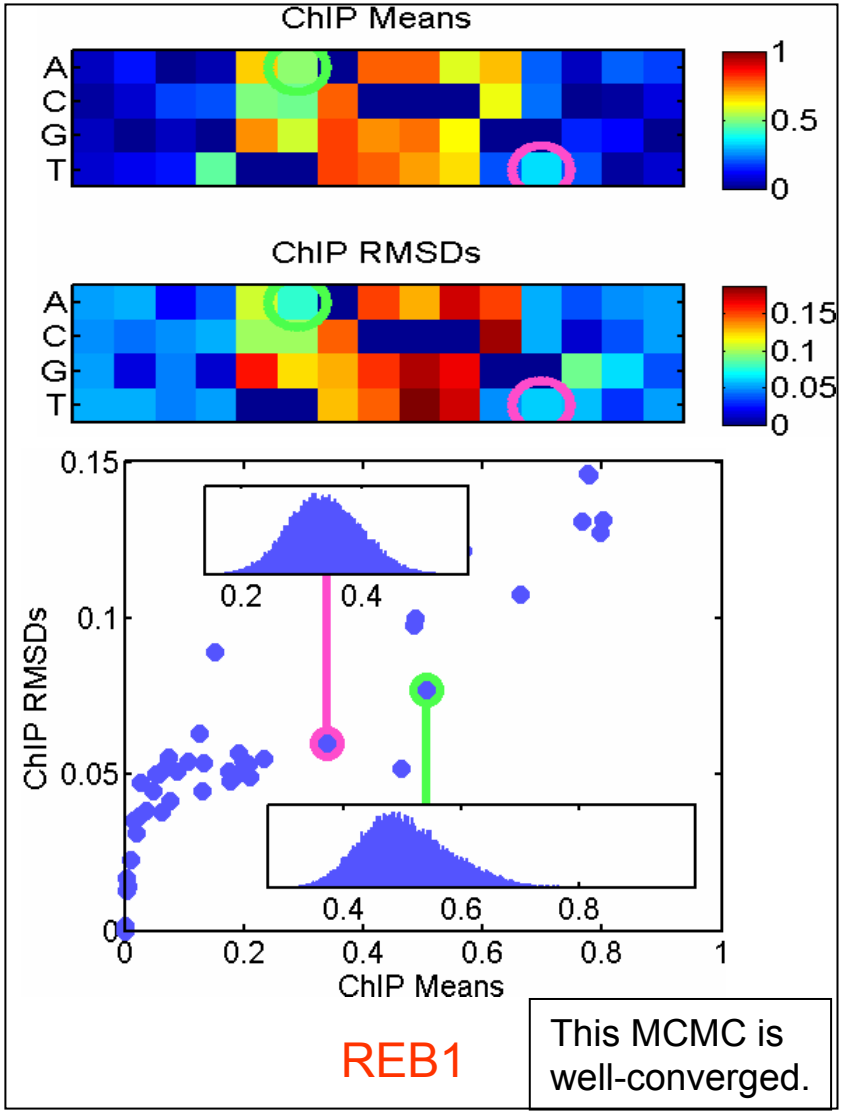
- We declare sites flagged by > 50% of energy matrices to be putative binding sites. Can make it tighter.
- Putative ChIP-chip sites are almost all predicted by the PBM matrices. As they should be!

- Experimenters identified putative sites by cutoff and found poor ChIP-chip/PBM prediction overlap.
- Changing the cutoff just admits more false positives: to do better, must decrease the expt'l noise.
- Our method for "understanding" the noise lets us flag more sites with little false positive penalty.



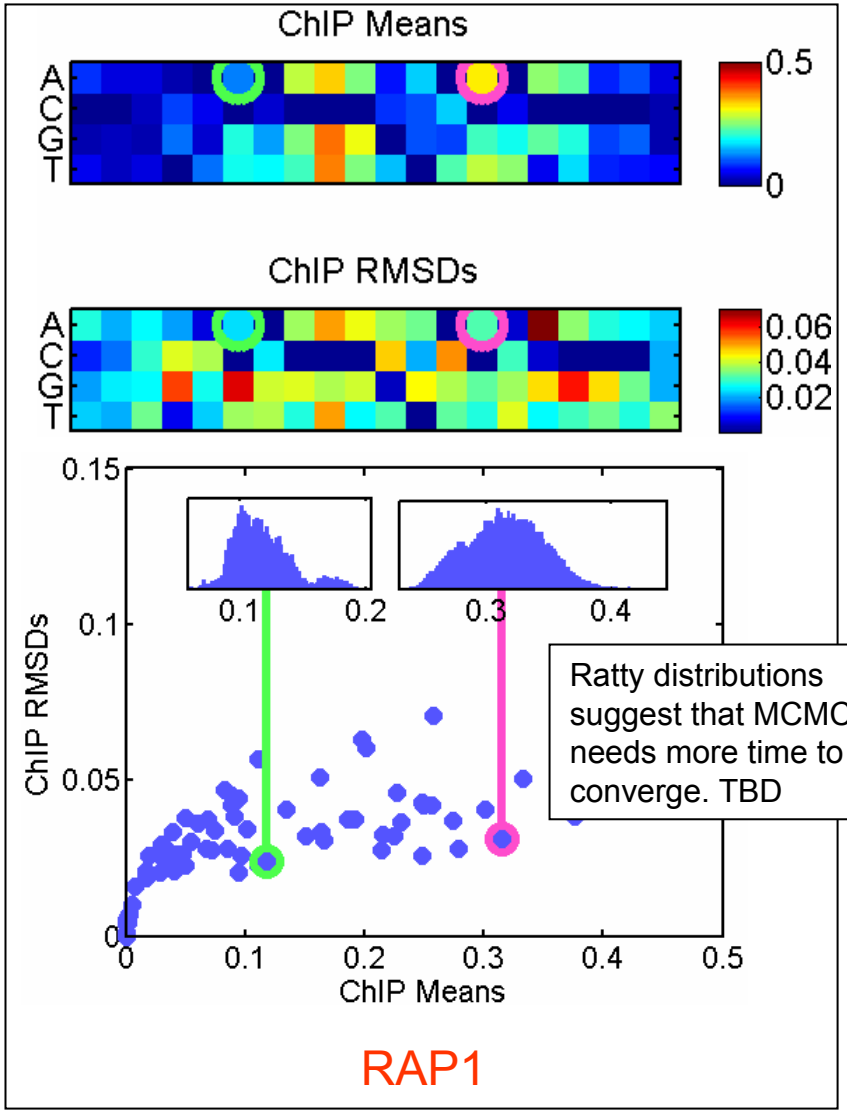
General lesson is that noise can be "understood" if the data is "bottle-necked" through a "good" parametric model. That the difference between *in vivo* versus *in vitro* experiment is captured this way is a nice surprise.

Some Results for Other Transcription Factors



REB1

This MCMC is well-converged.



RAP1

Ratty distributions suggest that MCMC needs more time to converge. TBD

How Well Do We Describe Binding Energies?

- Direct experimental evidence about TF-DNA binding energy is limited in scope. We really need hi-throughput direct energy measurements for a convincing test of our predictions (see Maerkl and Quake).
- We can predict scale-free energy differences between binding sites (explain). Knowing true K_d's for lots of sites would be informative.
- Functional properties of binding sites are governed in large part by energy: thus energy, not sequence, should be conserved in evolution if function is to be maintained.
- This suggests that comparing the predicted energies of orthologous binding sites would be a good indirect way of assessing whether our energy model is doing the right thing.
- If the model passes this test, our hundreds of binding sites would also provide the raw material for a quantitative study of evolutionary dynamics (genotype = site sequence; phenotype = site energy).

Orthology and Alignment of Genomes + Sites

Example intergenic region with predicted ecoli binding site for Crp:

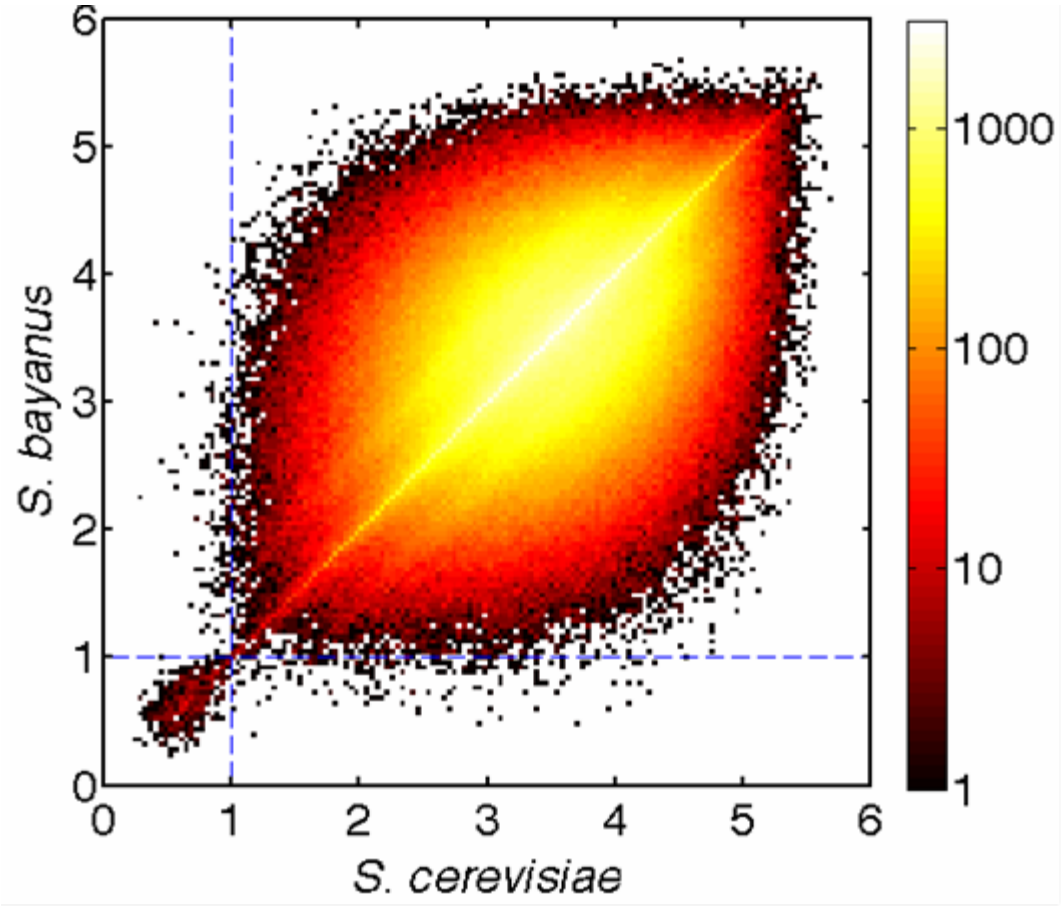
```
Sequence 1: ecoli          191 bp
Sequence 2: salm           198 bp

kefC folA   E=4.69->5.73, 8 mutations in the site   xxxxxxxxxxx marks the spot

          xxxxxxxxxxxxxxxxxxxxxxx
----TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTA
AGTAAAAAATGTGATGTTCTGCAAACTTTACTGCTAATTGGCTGTTTTTGAACTACTGTA
     ***  **** **    ** ******** ***** **  ********    * ****

ATGCGGCGAGTCCAGGGAGAGAGCGTGGACTCGCCAGCAGAATATAAAATTTTCCTCAAC
ATGCTGGCGCTCCACATCAAATGAGTGGCGTCGCCAGCAGAACGAAAAATTTTCGTGCTC
**** *    ****     *  * ****  ************   ********* *   *

ATCATCCTCGCACCAGTCGACGACGGTTTACGCTTTACGTATAGTGGCGACAATTTTTTT
ATCCTCTTTTCGTCAGTCGACGAAAGATTGCGCTTTACGTATAGTGGCGGCAATTTTTTT
*** ** *  *  **********  * ** ******************** **********
```

Alignment of related sequences amounts to finding the most parsimonious way of mutating one into the other (including possibility of creating gaps). Powerful software readily available. The red box identifies a binding site in Ecoli and shows the sequence of the orthologous site in Salmonella, Note that *sequence* conservation by itself doesn't do well at picking out likely TF binding sites.

Binding Energy is Conserved (Yeast ABF1)

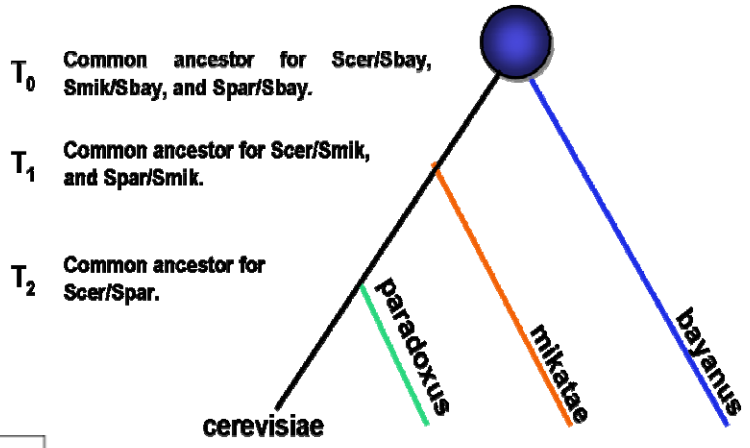
Use the energy model derived by likelihood analysis from *S.cerevisiae* assays to assign energies to sites (an *a priori* genotype-phenotype map):

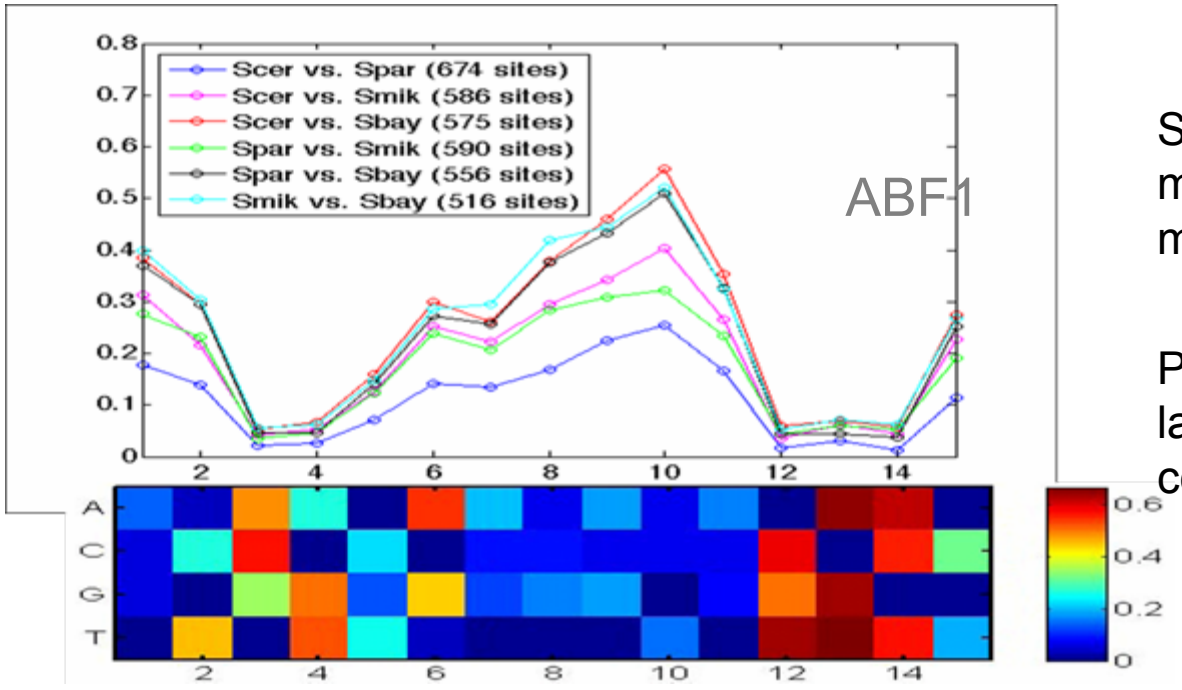
- 676 (!) intergenic sites in *S. cerevisiae* with Abf1p binding energy < 1 have clear orthologs in *S. bayanus* (the two proteins have high degree of homology)
- ~75% of these orthologs also have energy < 1. Conservation this strong is not an accident! No Free Parameters!
- Sites with energy > 1 have little or no correlation between energies in the two genomes: mutation randomizes energy.
- Clear evidence of selection pressure on binding site *energy;* selection pressure on *sequence* is indirect.
- Precise genotype-phenotype map is good starting point for a quantitative understanding of how TF binding sites and regulatory networks evolved.
- This will be pursued in the next talk.



Energy Clearly Imprinted on Sequence Evolution

We have a rich yeast phylogenetic tree and many orthologous site pairs. Can ask pop'n average questions about the likelihood of base changes at diff't locations in thte ABF1 site. Selection on the basis of energy ….not site-by-site, but on average in population.





Substitution probability pattern matches structure of energy matrix (bigger ΔE's disfavored).

Pattern evolves with time (from last common ancestor) as if under control of common `Hamiltonian'.

Comments and Conclusions

- The opportunities for quantitative study of evolution with a large population of binding sites and a clean quantitative phenotype are pretty exciting.
- That a minimalist energy model so accurately describes complex DNA binding data is a big surprise. Arbitrary sets of 100s of genes *cannot* be so regulated; how rare are large sets which *could* be?
- Different binding sites have different affinities (for good reasons?). We make specific predictions about how affinities are ordered. How does this concord with biochemical reality?
- The output of this effort is meant to be input to an effort to construct proper stat mech models of interesting networks.
- The “paradigm” that data determines a probability distribution on a space of parameterized models is much more general than this particular implementation. Our work is a “proof of principle”.

Acknowledgements

- Collaborators (responsible for most of the good ideas):

Justin Kinney
(PU Grad Student)

Gasper Tkacik
(PU Grad Student)

- Support:

  Burroughs-Wellcome Program in Quantitative Biology
  NIH Center for Systems Biology at Princeton University

- Reference: J Kinney, G Tkacik & C Callan PNAS104:501(2007)