What physics-inspired bio-informatics can tell us about transcription factor binding.

Curtis G. Callan, Jr.
Physics Department
Princeton University

Reporting on work with Justin Kinney and Gasper Tkacik

Acknowledgements

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

Justin Kinney
(PU Grad Student)

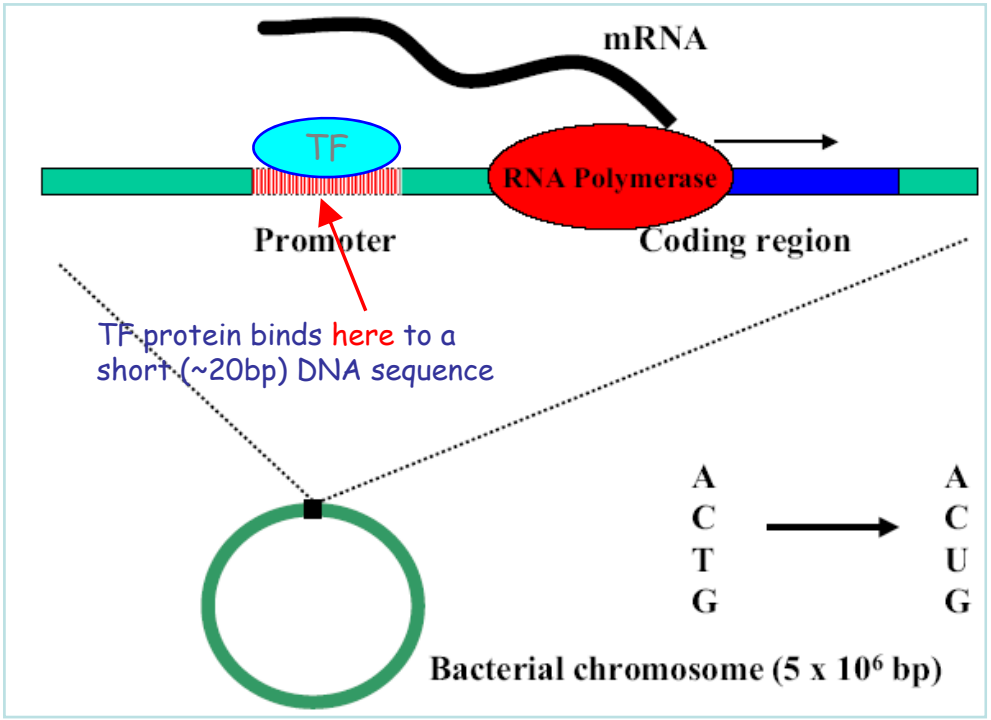
Gasper Tkacik
(PU Grad Student)

- 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)

Cartoon Overview of Gene Expression



Transcription factor proteins (TFs) bind to promoter to help (hinder) RNAP copy gene to mRNA.

Sequence-dependent TF-DNA binding thermodynamics controls which TF binds to which gene. Specificity is governed by energy.

RNAP protein complex makes an mRNA copy of the gene. Ribosome translates triplets of bases into amino acids via the "genetic code".

Coding Problem: Same TF binds to many different sequences. No analog of 3bp codons. Sites are statistically defined at best (next slide)

Binding Energy: Generic non-specific binding is weak; sequence-dependent binding provides a random energy landscape; strongest binding sites are probably the biologically significant ones.

TF Binding Sites: Statistical At Best

Sequences of some of the 48 Crp sites (19bp)			
Location	'Energy'	Sequence	Flanking Genes
70158	6.187863	AAGTGTGACGCCGTGCAAATAA	araB araC
431345	6.356798	AACTGTGAAACGAAACATATTT	tsx yajI
431384	9.872654	GTGTGTAAACGTGAACGCAATC	tsx yajI
702991	6.714032	TTTTGTGAGTTTTGTCACCAAA	nagB nagE
791335	6.900346	AAGTGTGACATGCAATAAATTA	galE modF
1019443	7.764454	ATGCCTGACGGAGTTCACACTT	ompA sulA
1236678	5.007025	AGATGTGAGCCAGCTCACCATA	ycgB dadA
2229736	6.836420	ATTTGCGATGCGTCGCGCATTT	yohK cdd
2229786	4.217979	TAATGAGATTCAGATCACATAT	yohK cdd
2350502	4.463704	ATGTGTGCGGCAATTCACATTT	glpT glpA
2350552	11.720174	AAACGTGATTTCATGCGTCATT	glpT glpA

Sequences of the three LacI sites (21bp)		
Location	'Energy'	Sequence
365546	0.809	AATTGTGAGCGGATAACAATT
365546	0.799	AATTGTTATCCGCTCACAATT
365145	4.068	AAATGTGAGCGAGTAACAACC
365145	4.058	GGTTGTTACTCGCTCACATTT
365638	6.449	GGCAGTGAGCGCAACGCAATT
365638	6.439	AATTGCGTTGCGCTCACTGCC

Left and Right Operator Sites in ϕ_λ				
Name	Location	ρ_{cI}	ρ_{cro}	Sequence(s)
OL1	35589	0.4810	0.0210	GTATCACCGCCAGTGGTAT
				ATACCACTGGCGTCGATAC
OL2	35613	0.0910	0.0470	TCAACACCGCCAGAGATAA
				TTATCTCTGGCGGTGTTGA
OL3	35633	0.0670	0.1160	TTATCACCGCAGATGGTTA
				TAACCATCTGCGGTGATAA
OR3	37949	0.0025	0.6850	CTATCACCGCAAGGGATAA
				TTATCCCTTGCGGTGATAG
OR2	37972	0.0125	0.0150	CTAACACCGTGCGTGTTGA
				TCAACACGCACGGTGTTAG
OR1	37996	0.3460	0.1160	TTACCTCTGGCGGTGATAA
				TTATCACCGCCAGAGGTAA

Binding Energy from Site Statistics (BvH)

Sequences of some of the 48 Crp sites (19bp)			
Location	'Energy'	Sequence	Flanking Genes
70158	6.187863	AAGTGTGACGCCGTGCAAATAA	araB araC
431345	6.356798	AACTGTGAAACGAAACATATTT	tsx yajI
431384	9.872654	GTGTGTAAACGTGAACGCAATC	tsx yajI
702991	6.714032	TTTTGTGAGTTTTGTCACCAAA	nagB nagE
791335	6.900346	AAGTGTGACATGGAATAAATTA	galE modF
1019443	7.764454	ATGCCTGACGGAGTTCACACTT	ompA sulA
1236678	5.007025	AGATGTGAGCCAGCTCACCATA	ycgB dadA
2229736	6.836420	ATTTGCGATGCGTCGCGCATTT	yohK cdd
2229786	4.217979	TAATGAGATTCAGATCACATAT	yohK cdd
2350502	4.463704	ATGTGTGCGGCAATTCACATTT	glpT glpA
2350552	11.720174	AAACGTGATTTCATGCGTCATT	glpT glpA

TF contacts an L-base-pair DNA string.
Uncorrelated additive model for affinity:
$E(b_1 b_2 \ldots b_L) = e_1(b_1) + e_2(b_2) + \ldots + e_L(b_L)$

4xL energy matrix $e_i(b_a)$ contains info about sequence specificity of binding. BvH estimates it from sequence data:

Pseudocounts for small statistics. Null model for bkgd stats.

$$e_i(b) = \log \frac{\max_a N_i(a) + 1}{N_i(b) + 1} > 0$$

Ex: position i=9:
$N_A = 2$, $N_C = 3$
$N_G = 3$, $N_T = 3$

Transcription factor binding statistics: LacI

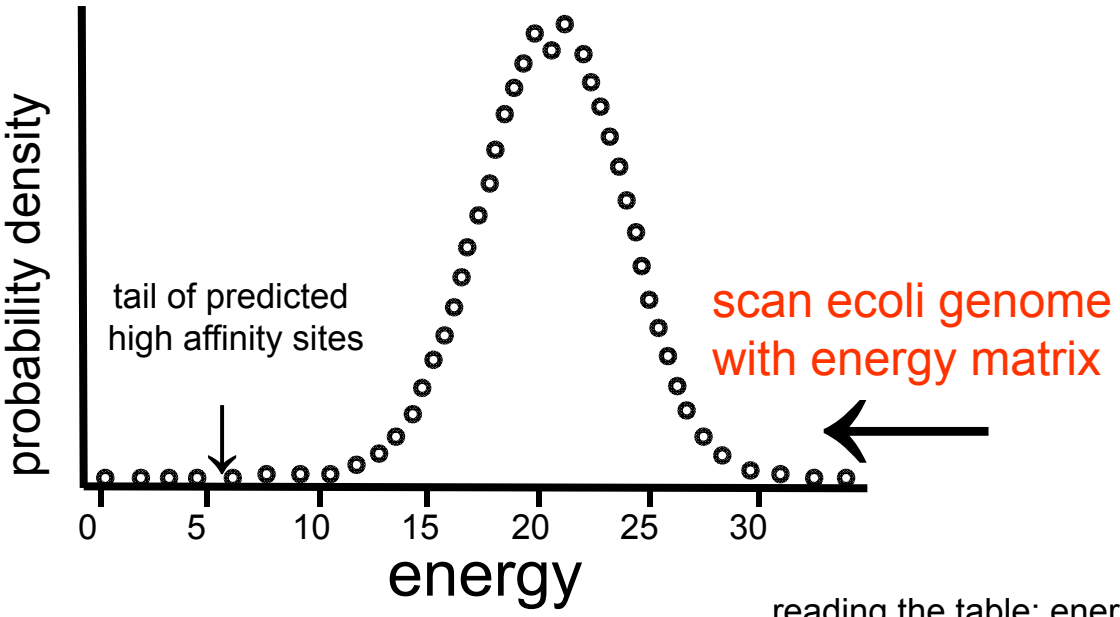
Given sequences for the strongest sites, construct the sequence-dependent energy function; run it over the genome to find site binding energy distribution; ….

sequences of the three LacI sites (21 bp)

location	"energy"	sequence
365546 365446	0.809 0.799	AATTGTGAGCGGATAACAATT AATTGTTATCCGCTCACAATT
365145 365145	4.068 4.058	AAATGTGAGCGAGTAACAACC GGTTGTTACTCGCTCACATTT
365638 365638	6.449 6.439	GGCAGTGAGCGCAACGCAATT AATTGCGTTGCGCTCACTGCC

BvH rule →

A	C	G	T
0.000	1.609	0.511	1.609
0.000	1.609	0.511	1.609
0.916	0.916	1.609	0.000
1.099	1.792	1.792	0.000
1.946	1.946	0.000	1.946
1.792	1.099	1.792	0.000
1.609	1.609	0.000	0.511
0.000	1.792	1.792	1.099
1.386	0.693	0.000	0.288
1.609	0.000	0.916	0.916
1.386	0.000	0.000	1.386
0.916	0.916	0.000	1.609
0.288	0.000	0.693	1.386
1.099	1.792	1.792	0.000
0.511	0.000	1.609	1.609
0.000	1.792	1.099	1.792
1.946	0.000	1.946	1.946
0.000	1.792	1.792	1.099
0.000	1.609	0.916	0.916
1.609	0.511	1.609	0.000
1.609	0.511	1.609	0.000



scan ecoli genome with energy matrix ←

reading the table: energy(ACCG ...) = 0.000 + 1.609 + 0.916 + 1.792 ...

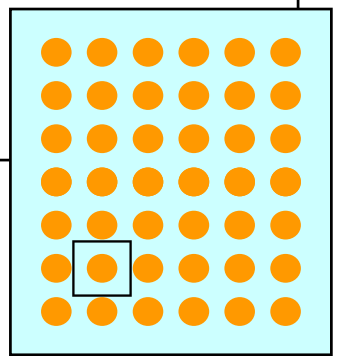
TF-DNA Energy Models from Binding Assays

- 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.

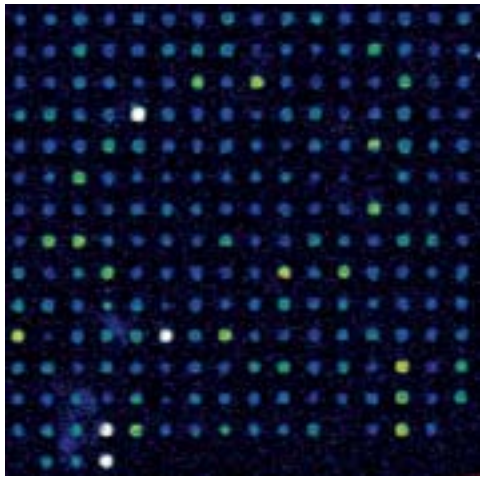
Some Philosophy: Lexical vs. Energetic Approach

- 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.

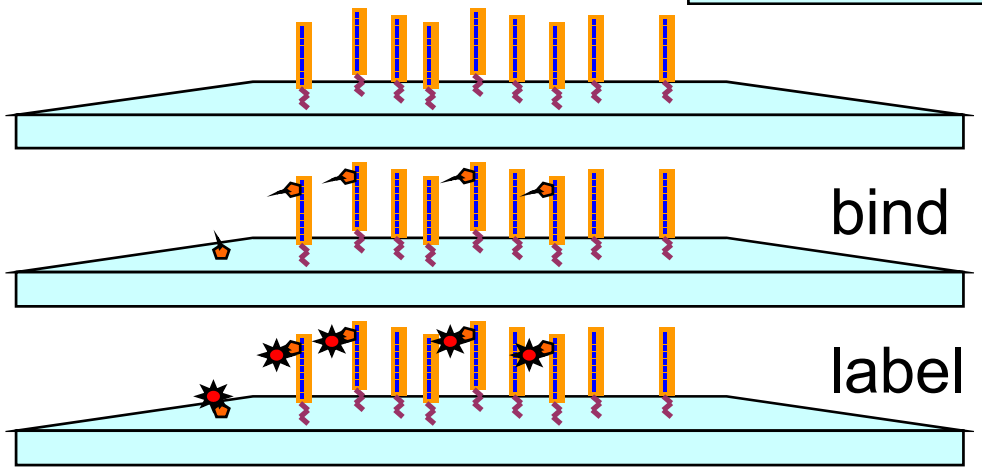
PBM Assay Overview (Mukherjee et al)



- 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 Seq-Specific Binding Energy 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"

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

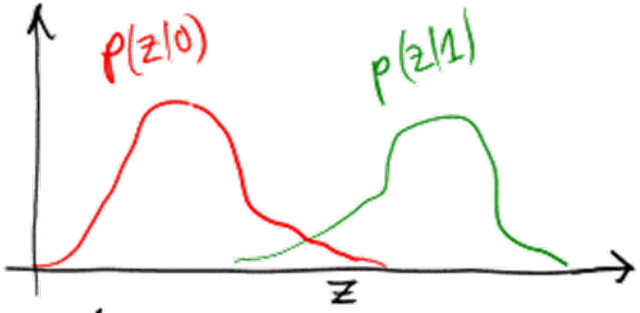
(binary *x*)

(continuous *z*)

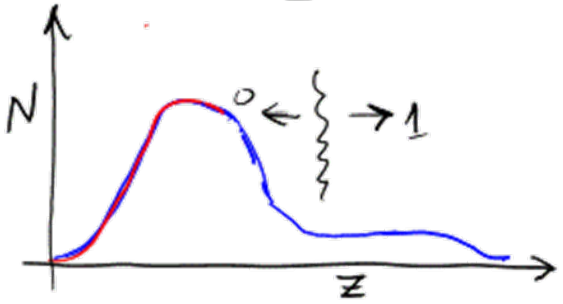
$$\text{Experiment}: \; s_i \mapsto z_i, \quad i = 1, 2, \ldots, N$$

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!

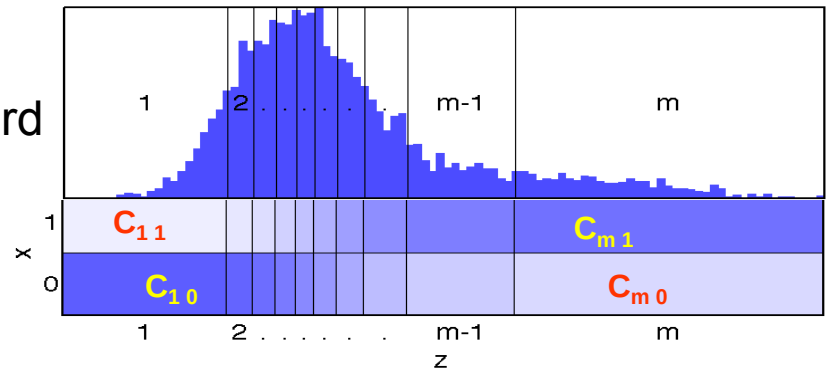
Options for More Sophisticated Modeling

- 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 or parametrized error model *p({z}|{x}) …*
- When does the number of parameters to be determined exceed the information content of the data? Hard to give a priori answer …
- Will show that the data (experimental numbers plus the genomic information) on wide-acting yeast TFs succeeds in fixing *many* parameters. Information content of the data *not* exhausted.

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 $\{x_j\}$ 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!)

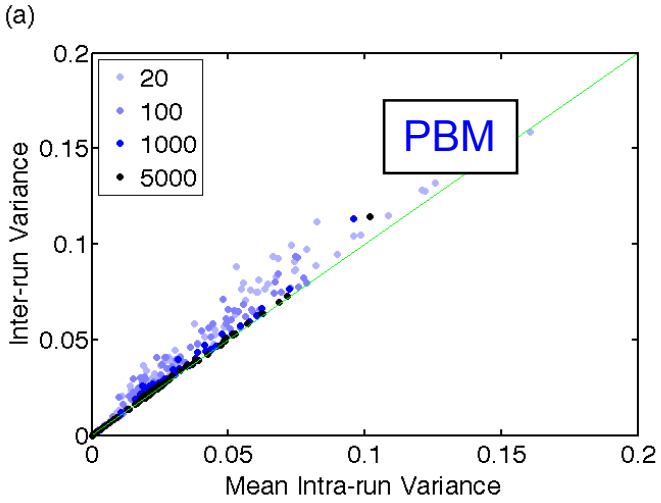
Digression on Mutual Information

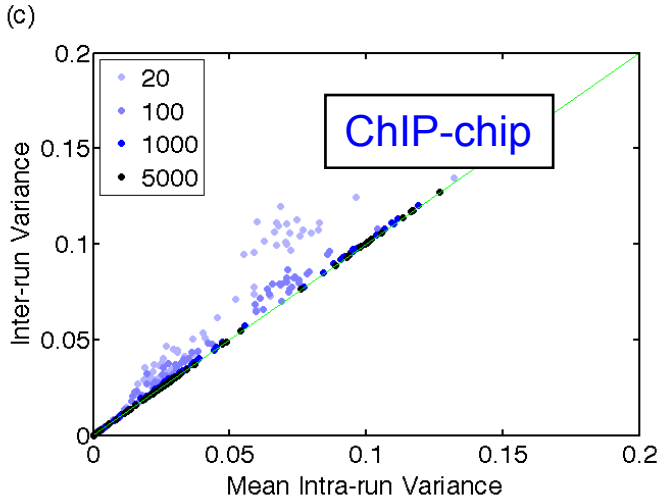
Markov Chain Monte Carlo Evaluation of $p(M|\{z_i\})$

- 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!

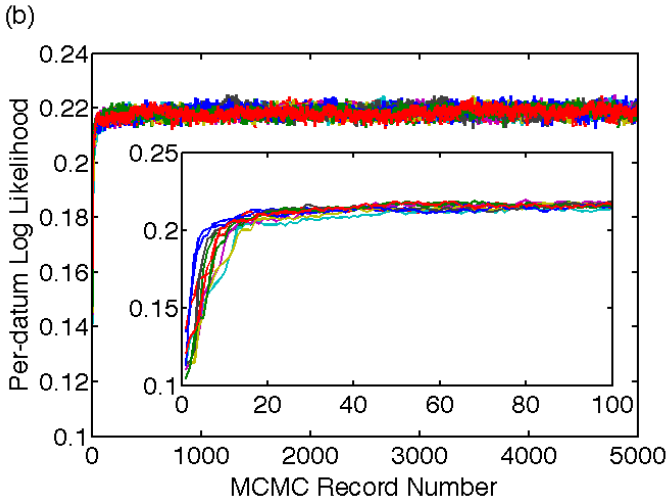
MCMC Estimation Converges *Fast* (for ABF1)

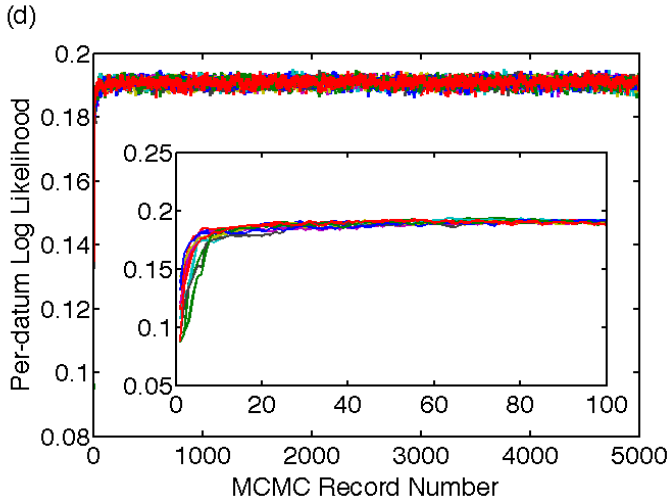
Burn-in test: do 10 runs, plot inter- and intra-run variance for each matrix element for larger and larger samples representing longer run times. Unit slope straight line is convergence signal.





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

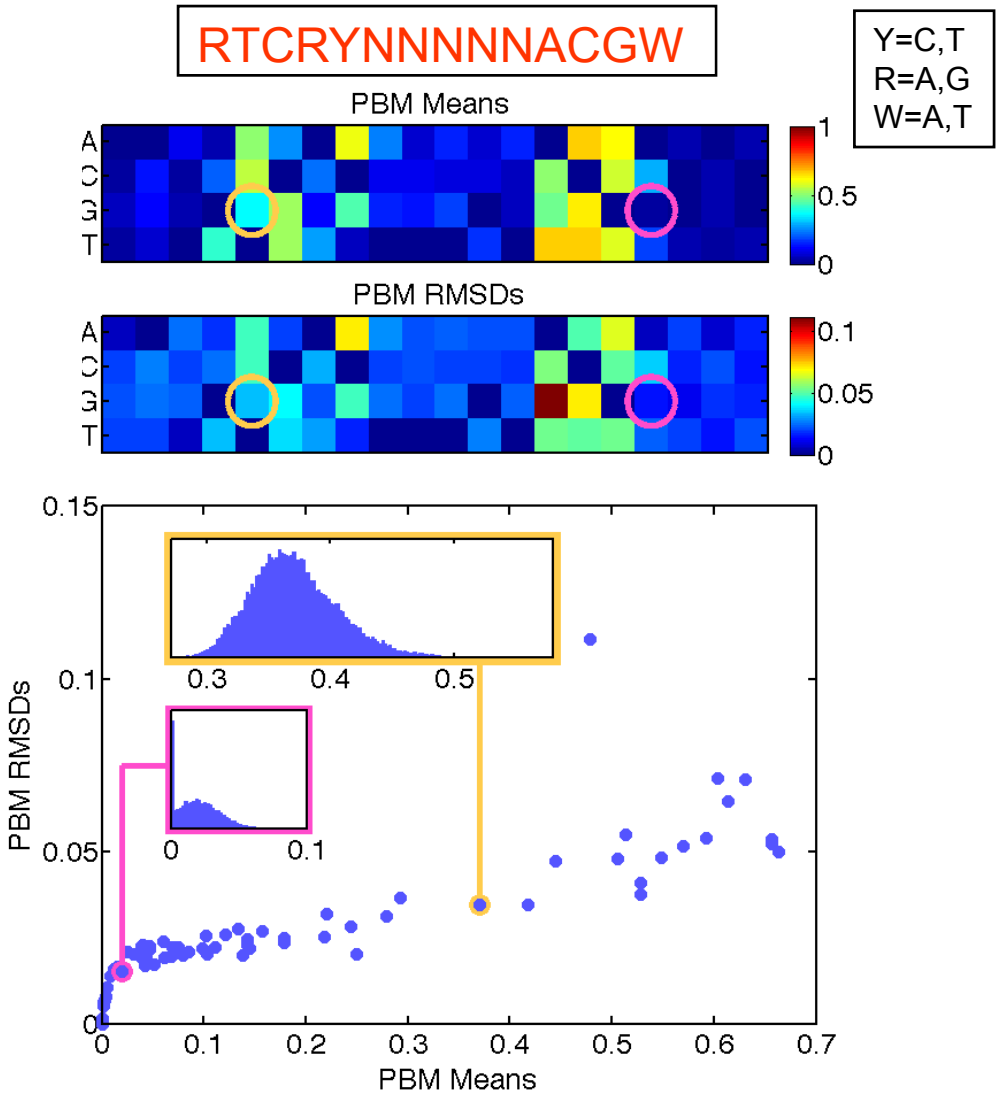




Key result: $p(M|\{z_i\})$ has a single smooth peak, easily found by MCMC!

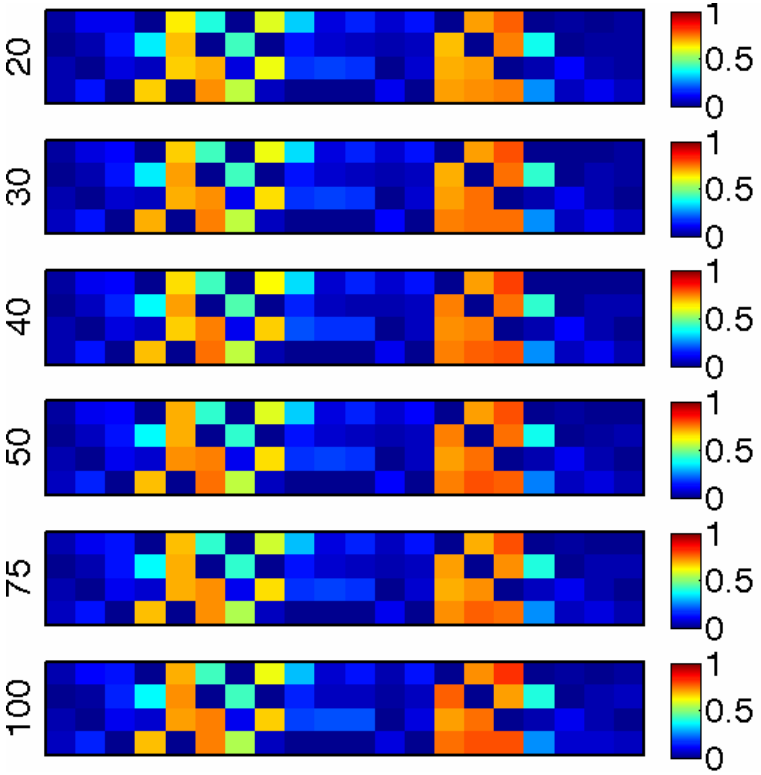
MCMC Results for TF ABF1p (Yeast)

- 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.

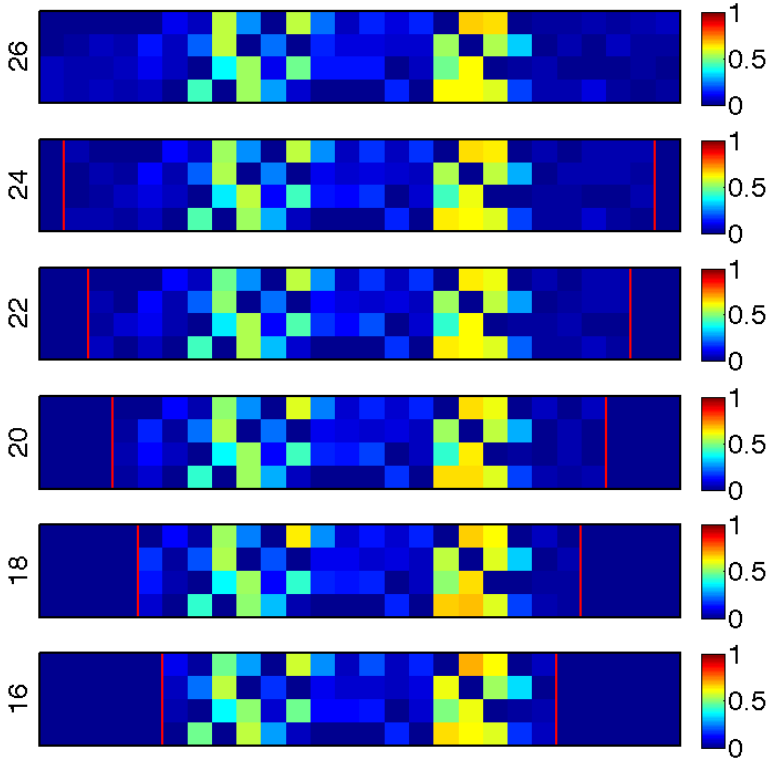


Results invariant to changed analysis parameters

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



Width of energy matrix (in bp)



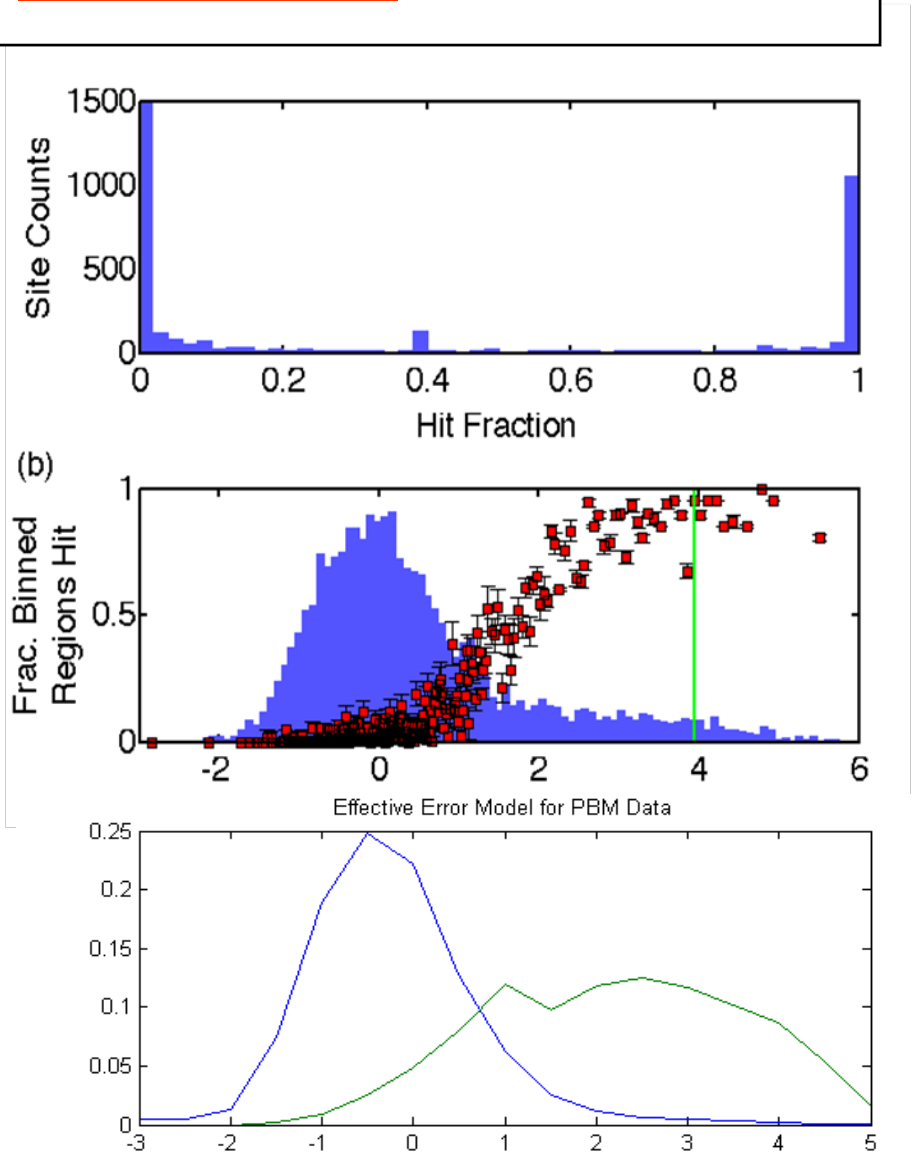
You can even divide the data (intergenic region LIRs) into randomly-chosen halves and compare the two mean energy matrices (overfitting test). They agree very well.

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)

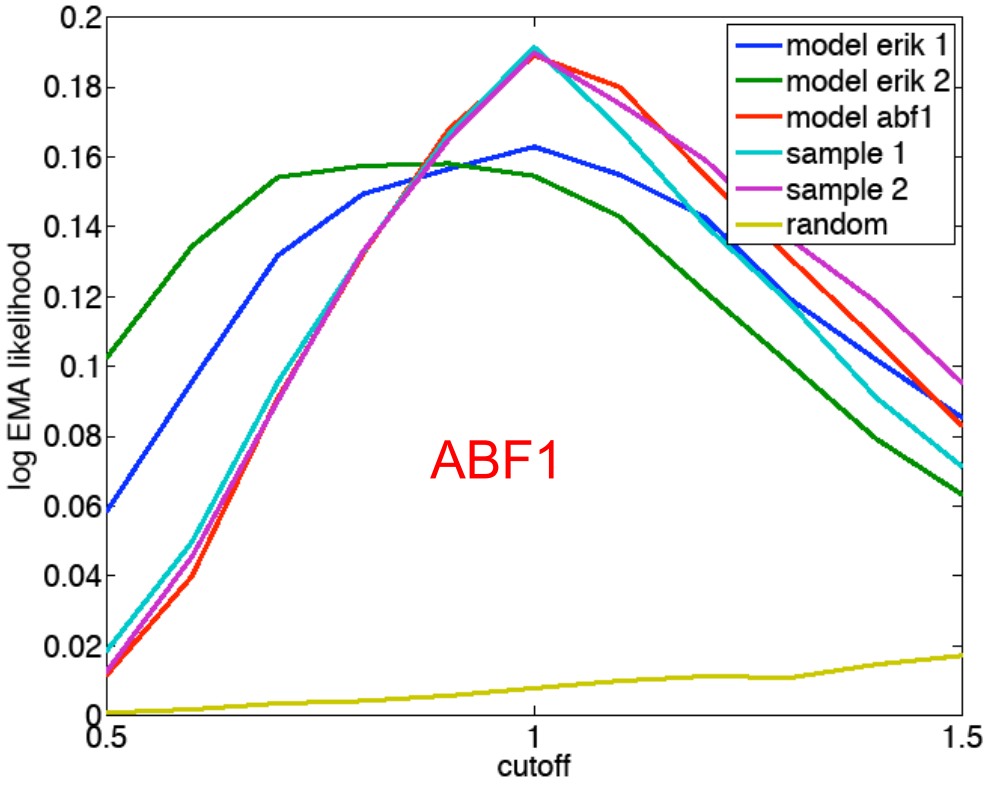
- Compare expt'l LIR dist'n with h.f. of binned regions: consistent with credible error model
- Conservative Mukherjee et al LIR cutoff (green line) rejects many regions clearly bound by our criterion.

- 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



MCMC Ensemble vs Std Bioinformatics

- Bioinformatic lists of ABF1 binding sites lead to Berg von Hippel emats.
- Test how well they describe the PBM data by computing log EMA likelihood versus choice of cutoff
- NB: BvH emat does not come with a cutoff. You have to choose one.
- For comparison, do the same for our ensemble average energy emat
- Not bad: max for BvH based on the larger set of sites is same as MCMC!

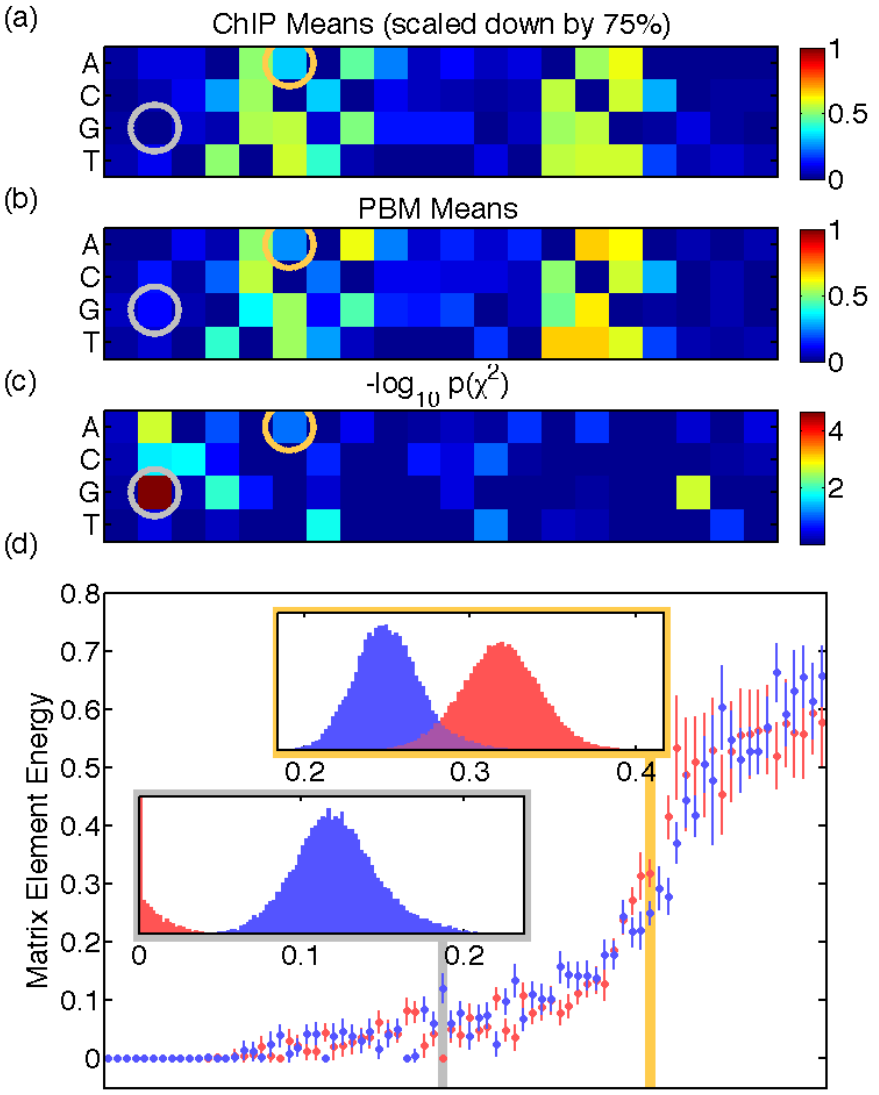


erik1 (erik2): list of 140 (670) abf1 sites (bioinformatics with different cutoffs)

But the comparison of relative likelihood of the two solutions is very bad. We plot per datum log likelihood and there are 6000 data points: 6000x(.19-.16)=180! Global fit of bioinformatic energy model is very poor in comparison.

PBM vs. ChIP-chip Data Analysis

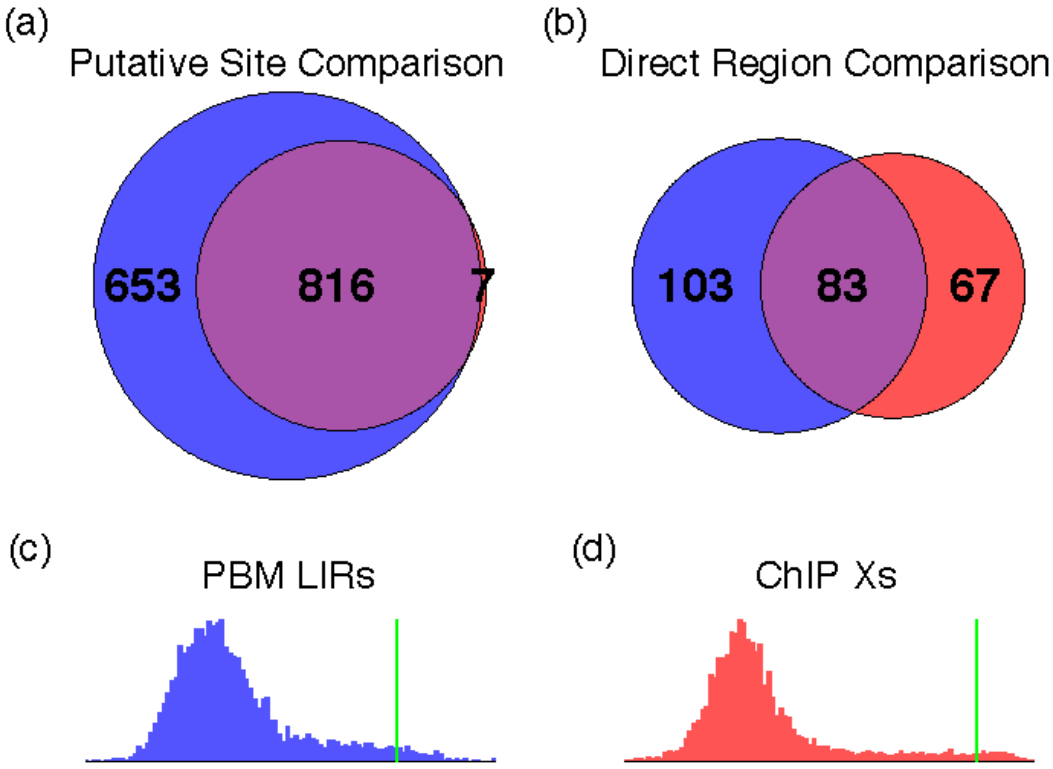
- PBM and ChIP-chip data give very similar matrices (but with the ChIP-chip cutoff set to .75 instead of 1).
- Cutoff stands in for the chemical potential of the TF: can vary between experiments (but the energy matrix should not!)
- Simple χ^2 test used to assess the overlap between the PBM and ChIP-chip 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.

Beyond the Energy Matrix Model?

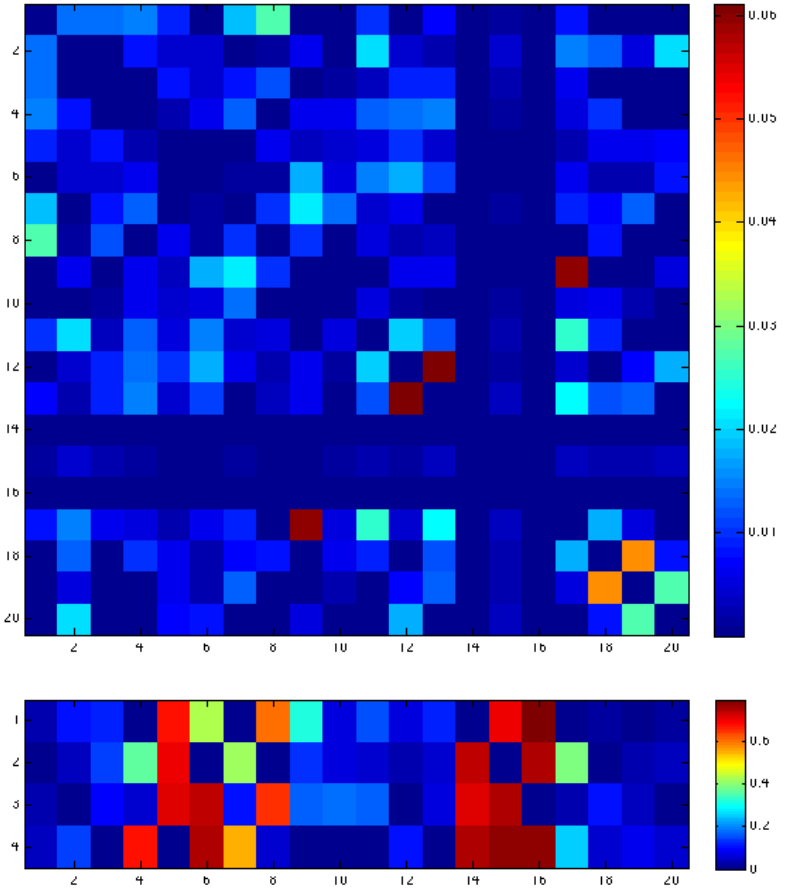
- Is the matrix model enough? Do we need inter-base correlations in the SDBE?
- Indirect evidence comes from the MI of the joint distribution of bases at sites *i* , *j* in the ensemble of putative ABF1 sites:

> 600 entries!

```
AAGTGTGACGCCGTGCAAATAA
AACTGTGAAACGAAACATATTT
GTGTGTAAACGTGAACGCAATC
TTTTGTGAGTTTTGTCACCAAA
AAGTGTGACATGGAATAAATTA
ATGCCTGACGGAGTTCACACTT
AGATGTGAGCCAGCTCACCATA
ATTTGCGATGCGTCGCGCATTT
```

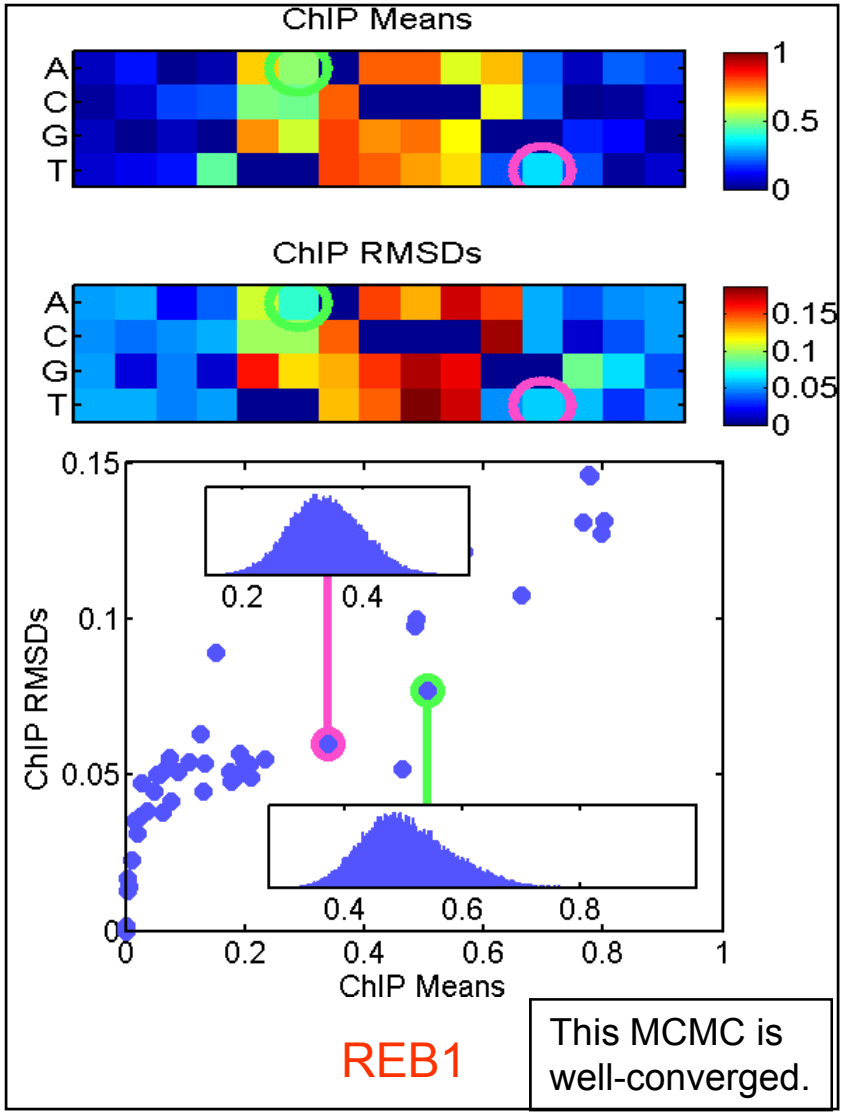
- Need pro-grade method of estimating MI due to small count effects, etc.
- MI is very nearly zero: max value on the color bar is .06 bits! Most entries < .01 bit

Sequence MI between position pairs



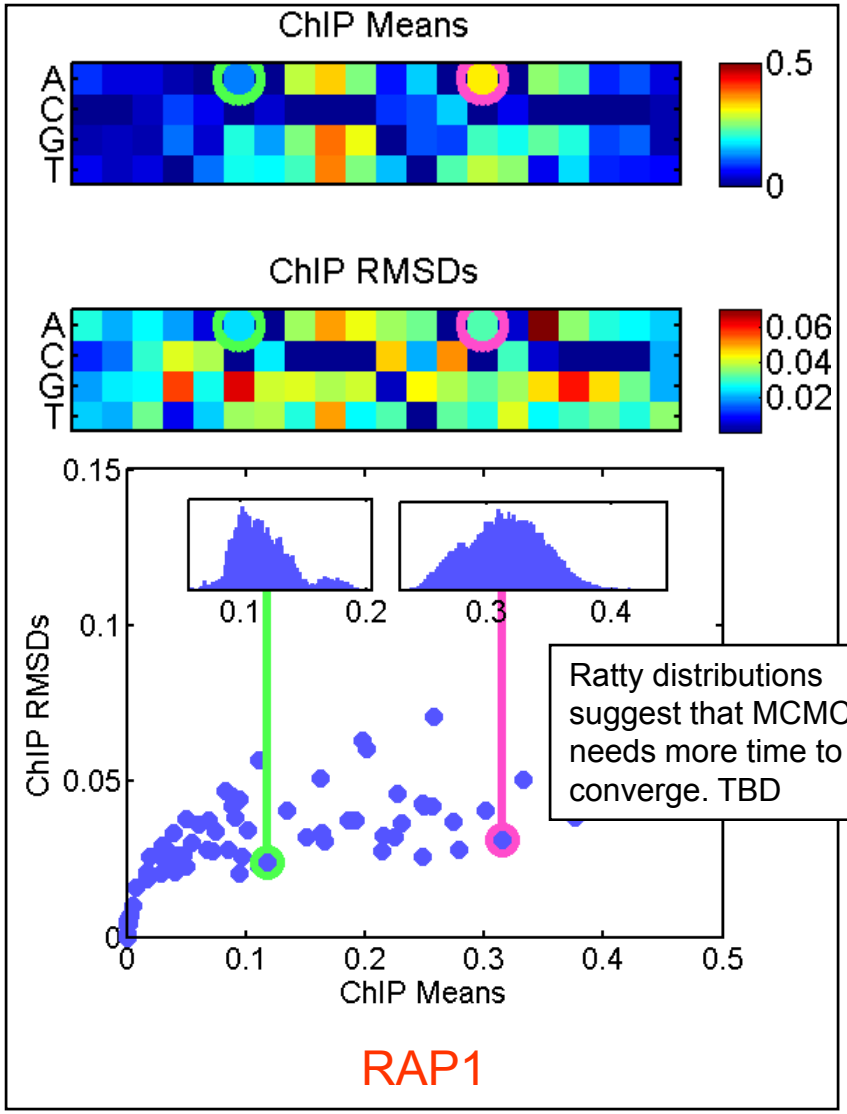
ABF1 Energy Matrix (for reference)

Some Results for Other Transcription Factors



REB1

This MCMC is well-converged.



RAP1

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

Comments and Conclusions

- 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?
- 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.
- 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 "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".

What physics-inspired bio-informatics can teach us about evolution.

Curtis G. Callan, Jr.
Physics Department
Princeton University

Reporting on work with Justin Kinney, Michael Lassig & Ville Mustonen

Some interesting connections with evolution?

- 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 many binding sites would provide the raw material for a quantitative study of evolutionary dynamics
- We bring something new to the party: accurate genotype-phenotype map (genotype = site sequence; phenotype = site energy).
- The opportunities for quantitative study of evolution with a large population of binding sites and a clean quantitative phenotype are very promising.

Orthology and Alignment of Genomes & Sites

Example ecoli intergenic region with predicted binding site for Crp TF:

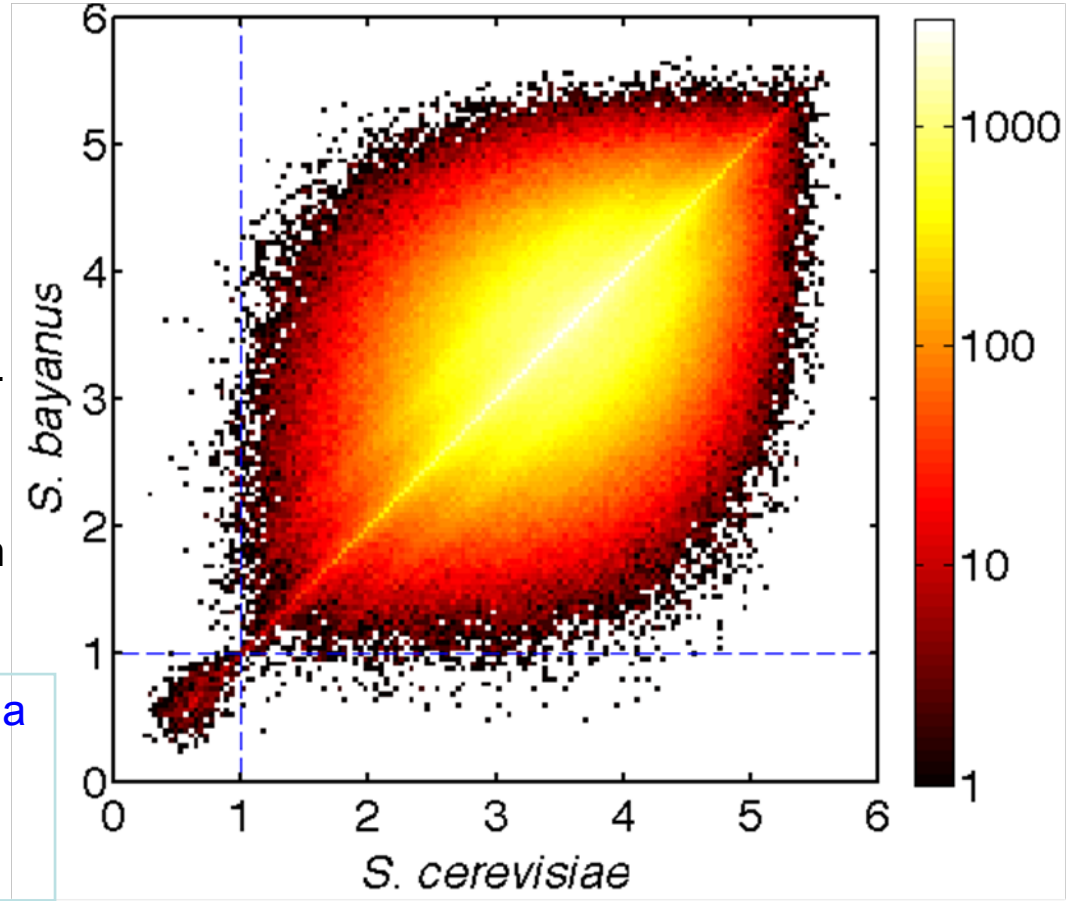
```
Sequence 1: ecoli          191 bp
Sequence 2: salm           198 bp

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

          xxxxxxxxxxxxxxxxxxxxxx
----TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTA
AGTAAAAAATGTGATGTTCTGCAAACTTTACTGCTAATTGGCTGTTTTTGAACTACTGTA
     ***  **** **    ** ******** ***** **  ********    * ****

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

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

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

Binding Energy is Conserved (Yeast ABF1)

Energy model derived from *S.cerevisiae* ABF1 assays assigns energies to site pairs in related yeasts *S. cerevisiae* and *S. bayanus.* The two are quite diverged (~40% intergenic region substitution rate) but the ABF1 proteins are very close.

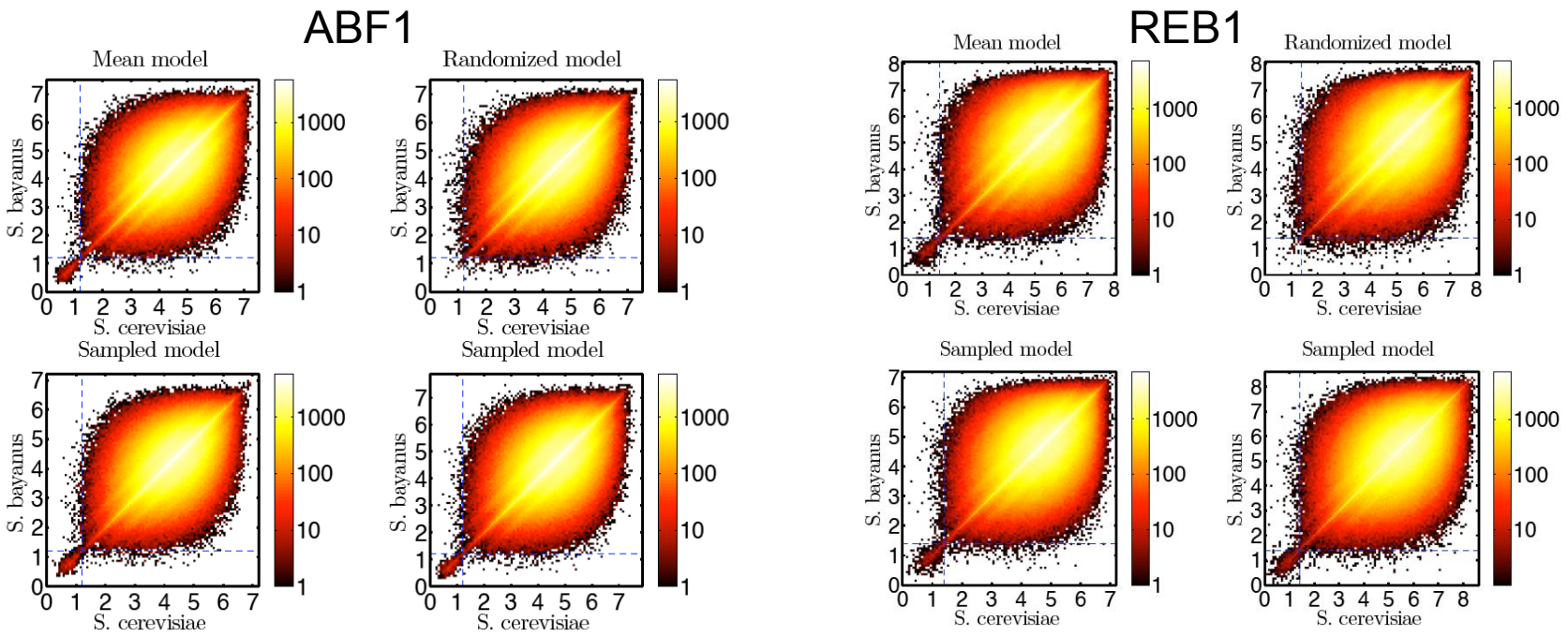
- 676 intergenic sites in *S. cerevisiae* with Abf1p binding energy E< 1 have ungapped orthologs in *S. bayanus.*
- Most of these orthologous sites (more than 75%) also have energy < 1: below E_{cut} mutation conserves energy.
- Sites with energy > 1 have little or no correlation between energies in the two genomes: mutation randomizes energy.
- Conservation this strong is no accident. Note that E_{cut} = 1 was derived from ChIP-chip data, not adjusted.
- Clear evidence of selection pressure on binding site *energy;* Selection pressure on *sequence* is indirect.



A precise genotype-phenotype map is a good starting point for a quantitative understanding of how TF binding sites and regulatory networks evolved.

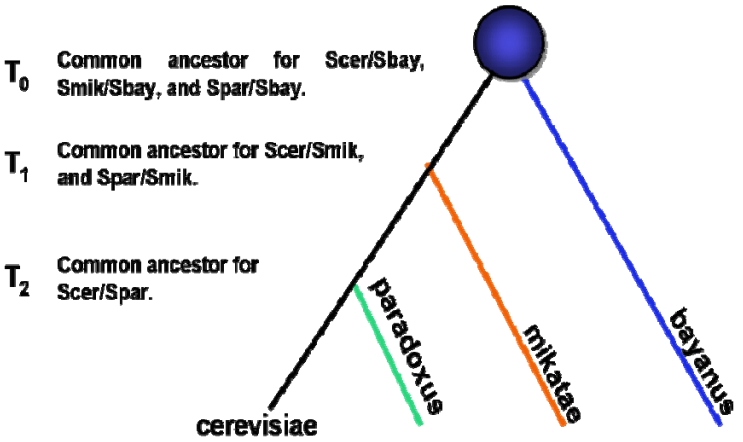
Conservation is an ensemble property

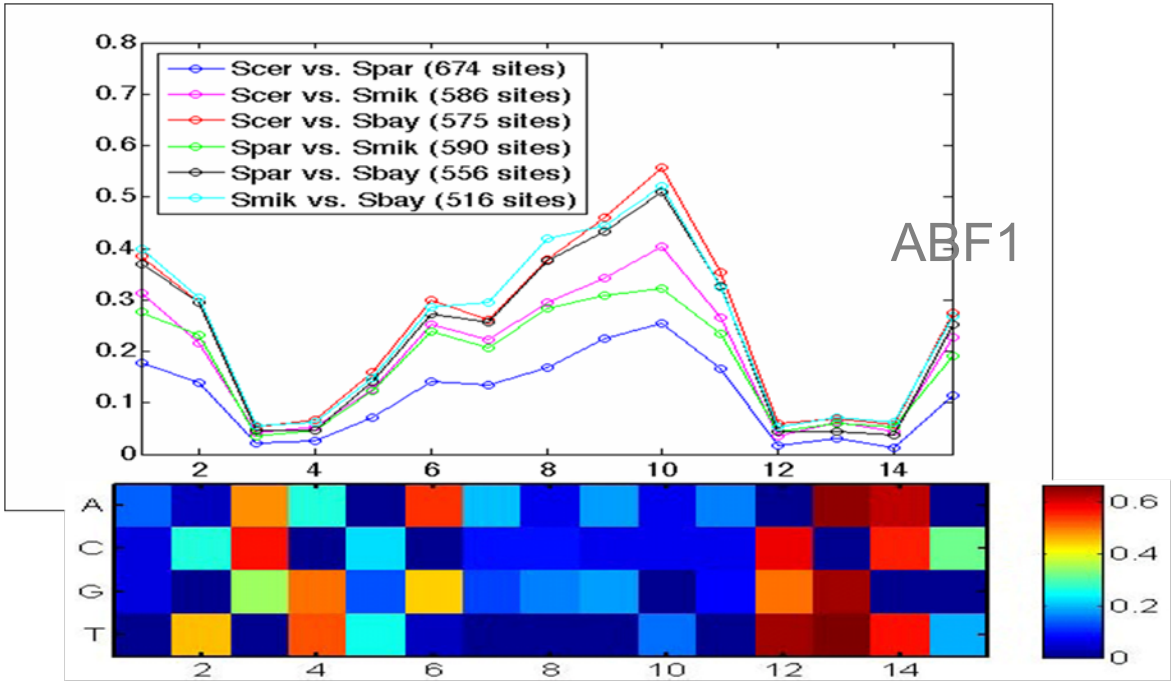
Conservation characterizes an energy matrix that accurately represents a real TF. Matrices in our MCMC ensemble should all be “good” models and its instructive to compare conservation plots of randomly chosen sample models. For comparison we show the mean MCMC model and a sampled model with randomized columns: it has the same base pair usage patterns, but positional correlations are lost.



Energy Imprints Itself on Sequence Evolution

We have a rich yeast phylogenetic tree and hundreds of below-threshold orthologous site pairs. Look at population average of likelihood of base *changes* between such pairs at different locations in the ABF1 site.





- Selection effects on sequence hard to see site-by-site, but visible in population average.
- 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'.
- Can we convert this metaphor into a precise analysis?

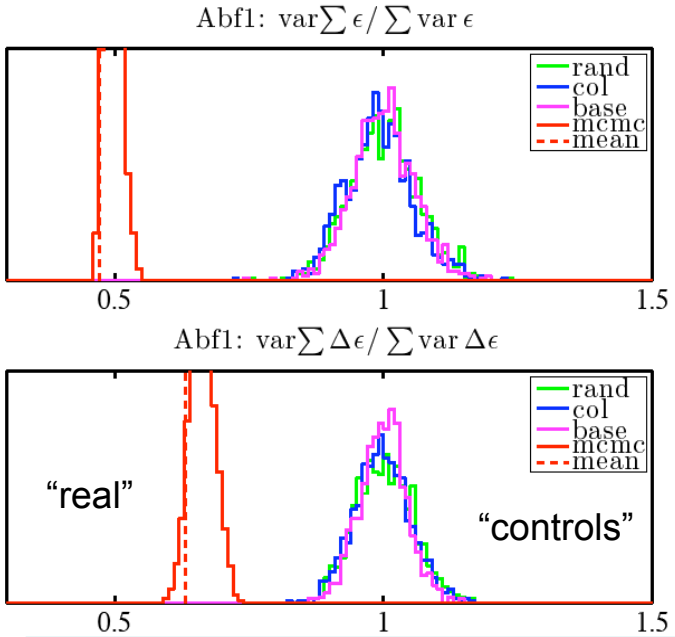
Another view of site energy conservation

Site positional energy contributions ε_i (i=1-20) & their changes $\Delta\varepsilon_i$ between orthologous sites usually assumed to vary independently in any site population (BvH). Simple diagnostic test:

$$\frac{\mathrm{var}\left(\sum_i \epsilon_i\right)}{\sum_i \mathrm{var}\left(\epsilon_i\right)} = 1, \qquad \frac{\mathrm{var}\left(\sum_i \Delta\epsilon_i\right)}{\sum_i \mathrm{var}\left(\Delta\epsilon_i\right)} = 1$$

Evaluating these quantities on the ensemble of putative Abf1 sites in *S. cerevisiae* (and their orthologs in *S. bayanus*) gives:

$$\frac{\mathrm{var}\left(\sum_i \epsilon_i\right)}{\sum_i \mathrm{var}\left(\epsilon_i\right)} = .473 \qquad \frac{\mathrm{var}\left(\sum_i \Delta\epsilon_i\right)}{\sum_i \mathrm{var}\left(\Delta\epsilon_i\right)} = .632$$



Variances for the Abf1 *Scer* sites computed with many different emats

In fact, variance of total energy is much less than sum of variances of positional energies. True for single species and for changes between species. Sites in one species are long-time outcome of mutation process connecting species: simple interpretation is that substitutions occur under selective pressure on full site energy

Population genetics of binding site evolution

Assume: a) particular TF locus σ_1 evolves over time into a succession of sites $\{\sigma_1\}$ selected according to a fitness function F(E(σ)), and b) multiple sites [σ_i] of same TF in one species are equivalent to the ensemble of one locus over time. Implicit assumption that all loci of our TF are subject to same fitness F(E) (TBD).

Kimura-Ohta population genetics: finite population, normally dominated by one allele (specific site seq'ce); random mutations fix with probability set by fitness relative to current allele

Basic K-O formula: neutral substitution rate μ is modified by fitness change $\Delta F_i = F(\sigma_{i+1})-F(\sigma_i)$ to

$$r_i \equiv r(\sigma_i \rightarrow \sigma_{i+1}) = \mu_{ab}\frac{\Delta F_i}{1-e^{-\Delta F_i}}$$

Let $P_0(\sigma)$ be the site sequence distribution satisfying detailed balance under the bkgd mutation rate μ (call it the null dist'n):

Functional dist'n Q(σ) satisfying detailed balance under KO fitness-modified rate is:

$$Q(\sigma) = \exp(F(\sigma))P_0(\sigma)$$

Fitness depends on sequence via energy E(σ); so project distributions on E:

$$Q(E) = \exp(F(E))P_0(E)$$

Read off fitness function from energy distribution of TF binding sites! Neat idea of Mustonen & Lassig (2005)

Fitness functions derived from real genomes

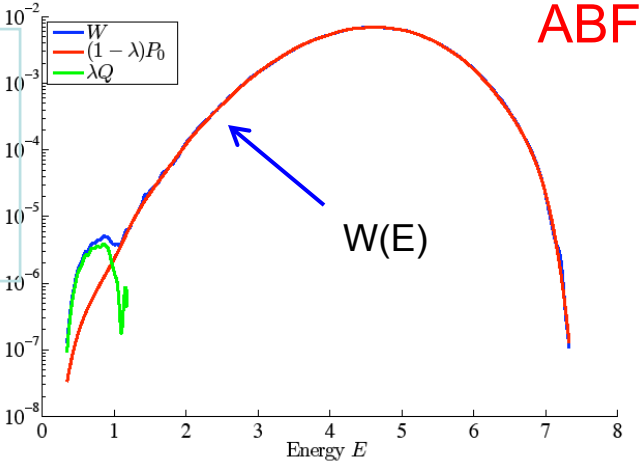
Basic idea: wrt a given TF, sites belong to either to P_0 (bkgd) or Q (func'l) distributions. Mixture model (HMM) for full site dist'n:

$$W(E) = \lambda Q(E) + (1-\lambda)P_0(E)$$
$$= \lambda P_0(E) e^{F(E)} + (1-\lambda)P_0(E)$$

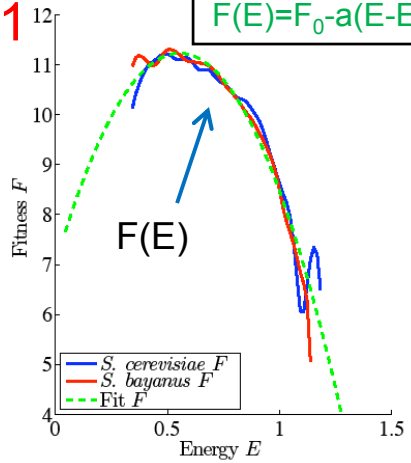
Construct $P_0(\sigma)$ to match tri-base statistics of intergenic regions; read off λ from large E behavior of W; then read off Q from low E; finally use ML formula to get F.

Amazingly clean fit for yeast TFs; diff't genomes give same fitness for same TF; simple polynomial fit to F(E) seems adequate (more anon!)

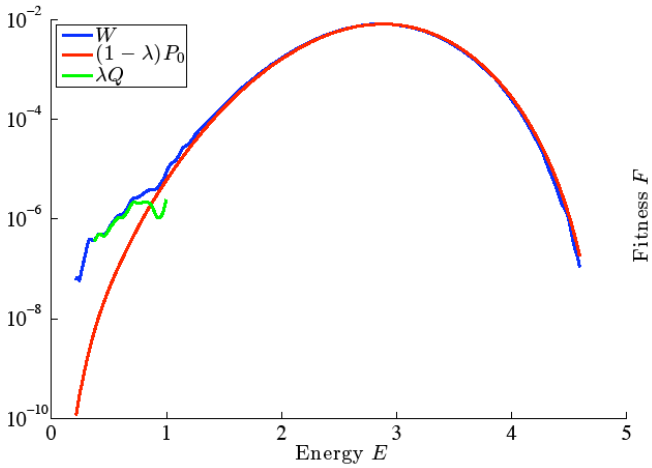
ABF1



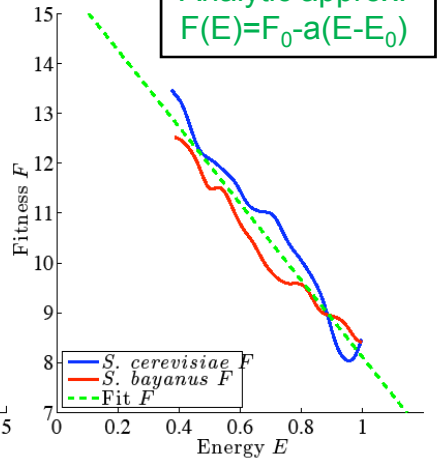
Analytic approx:
$F(E)=F_0-a(E-E_0)^2$



RAP1



Analytic approx:
$F(E)=F_0-a(E-E_0)$



Linear fitness, BvH energy & bioinformatics

- Fitness F(E) linear in E has special properties when the energy E is a sum of positional contributions!
  - The "functional" distribution Q(σ) on sequences factorizes into a product of independent base distributions for each position
  - In single base mutation events, the change in fitness ΔF depends only on the mutating base: no dependence on context
  - The K-O fixation probability of a single-base mutation in a site is independent of the identity of all other bases in the site
- These properties of site ensembles are the basis of most of what is done in bioinformatics (esp. Berg-von Hippel)
  - Our analysis shows that linear-in-energy fitness is by no means a universally valid assumption!
  - Non-linear fitness is easily incorporated in simulations of mutation dynamics and its effects are easy to diagnose.

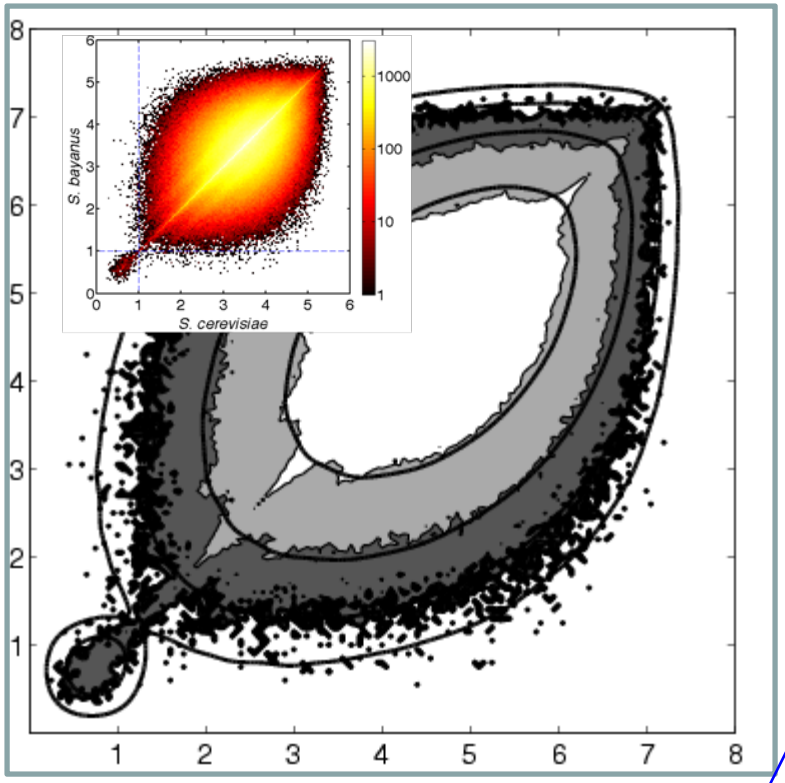
Simulating Abf1 binding site evolution

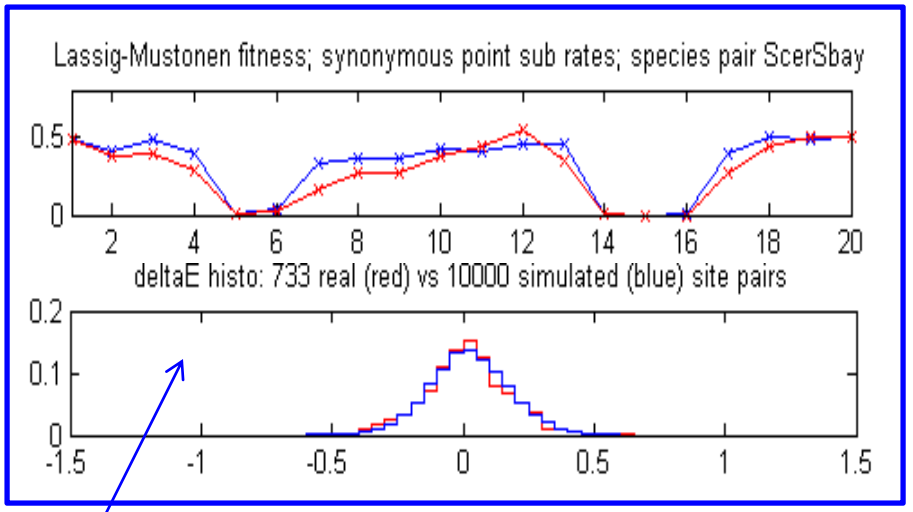
- Use $P_0(\sigma)$ and the fitness function $F(E(\sigma))$ in an MCMC to generate sample $\{\sigma\}_{Q1}$ of sites from functional distribution $Q(\sigma) \sim e^{F(E)}P_0(\sigma)$.
  - This simulates the distribution of Abf1 sites in the initial genome.
- Single point mutation rate matrix **r** is derived from intergenic region average *Scer/Sbay* substitution rate (or synonymous codon usage).
  - Invert observed sub rate $p(b|a) = \exp(t\mathbf{r})|_{ba}$ over *Scer/Sbay* divergence time t.
- Use Gillespie algorithm to evolve sites in $\{\sigma\}_{Q1}$ over time t using K-O rates based on $\mathbf{r}_{ba}$ and F(E) for the various single base substitutions.
  - This generates a simulated sample $\{\sigma\}_{Q2}$ of Abf1 sites in the evolved genome.
- Individual simulated site pairs carry no useful information, but statistics of site ensembles can be usefully compared with data
- For energy variances, we find remarkable agreement (esp. for interspecies comp'n)

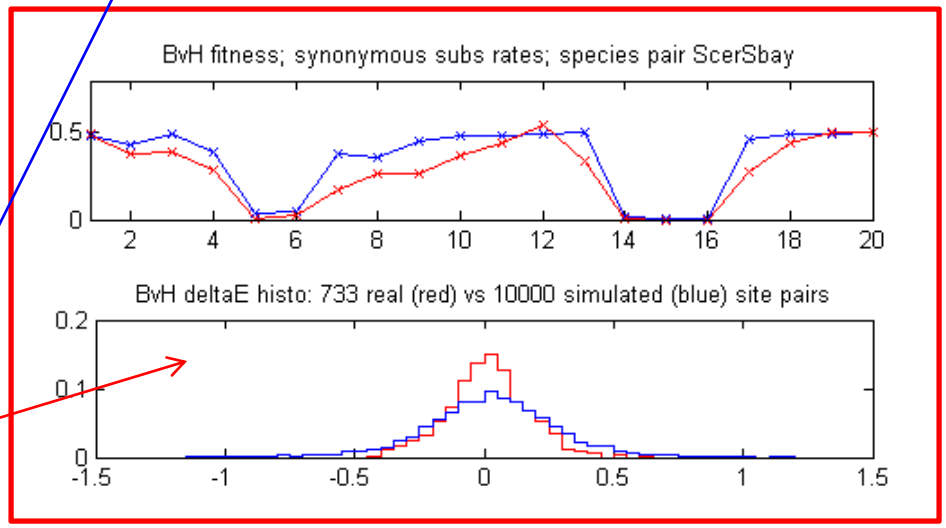
$$\frac{\mathrm{var}\left(\sum_i \epsilon_i\right)}{\sum_i \mathrm{var}\left(\epsilon_i\right)} = .36 \qquad \frac{\mathrm{var}\left(\sum_i \Delta\epsilon_i\right)}{\sum_i \mathrm{var}\left(\Delta\epsilon_i\right)} = .60$$

Evolution Simulation Results

Scer /Sbay simulated energy propagator is remarkably like the real thing:

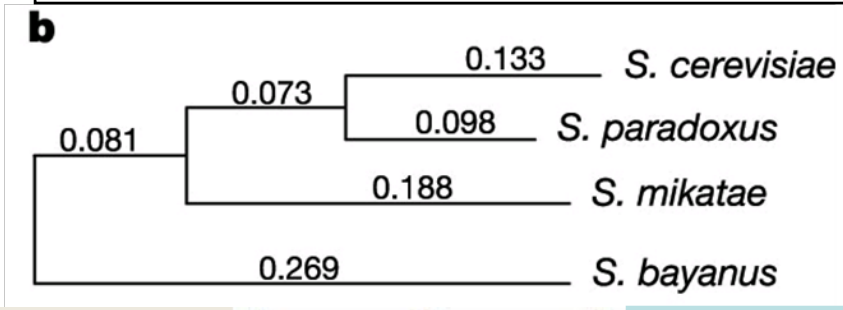






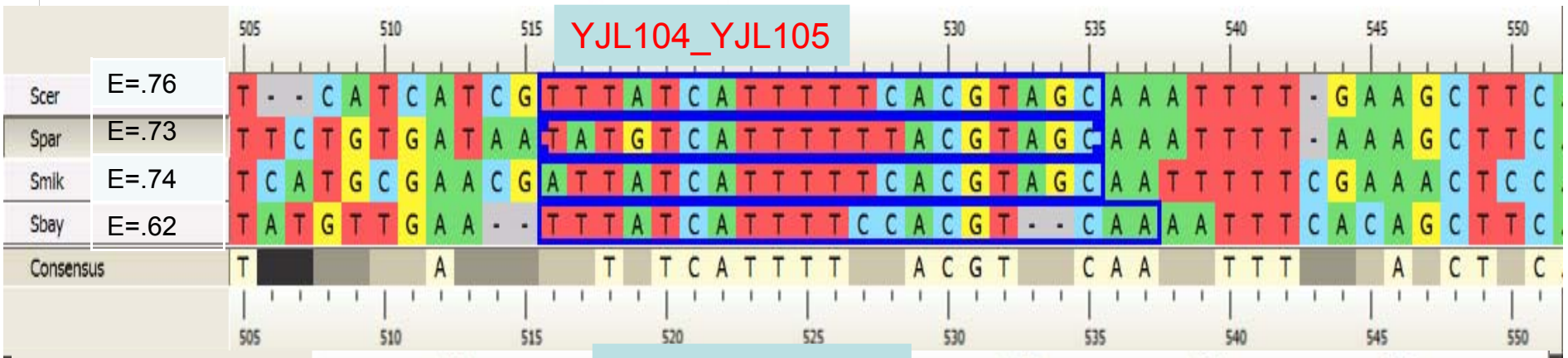
Simulated statistics on ortholoIgous site pairs for Abf1 (true vs BvH fitness):

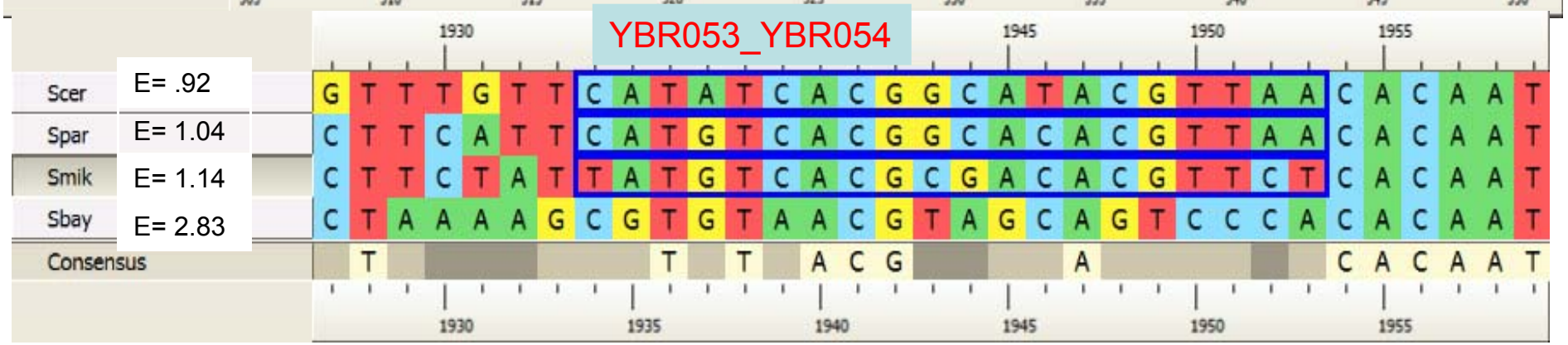
Four-way alignment & energy conservation



Define site clusters by alignment (5bp wobble) and relaxed energy cutoff (E<1.2)

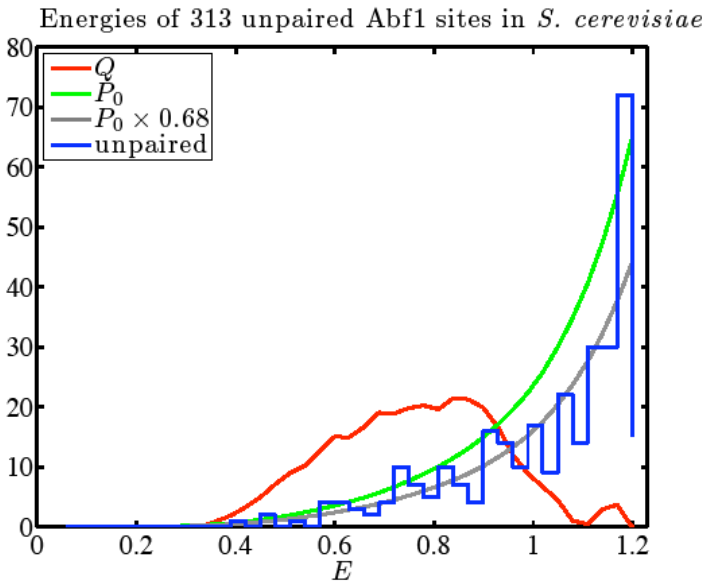
Q4: 660 instances in 606 intergenic regions
Q3: 17/12/66/52 cases w Scer/Spar/Smik/Sbay missing

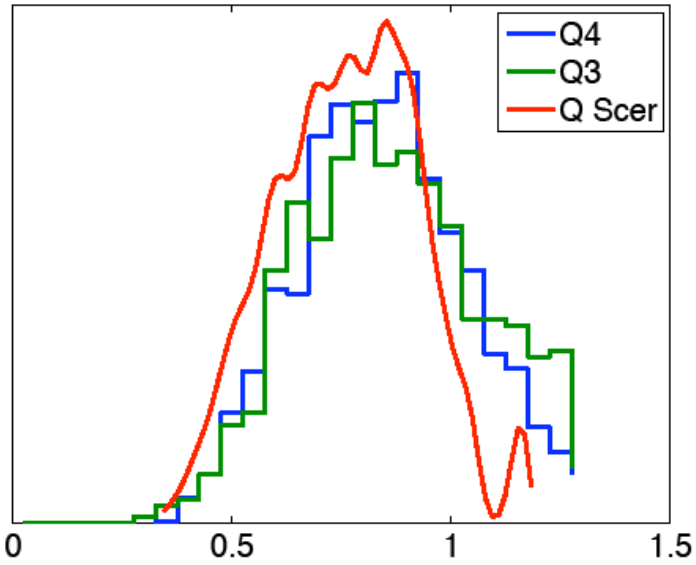




Multi-species analysis weeds out null sites

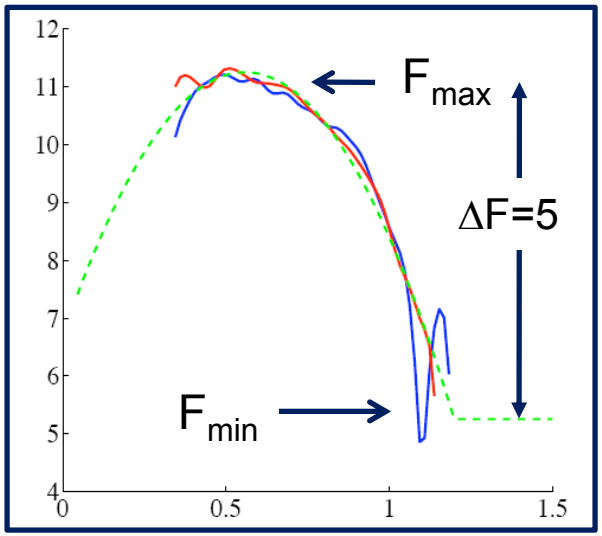
Focus on Scer/Sbay site pairs where one site has "ionized" ($E>E_{cut}$). Are they loss of function events? Does E dist'n of the $E< E_{cut}$ sites match the functional dist'n Q or the null dist'n P_0? It looks as if these sites are low-E tail of null dist.





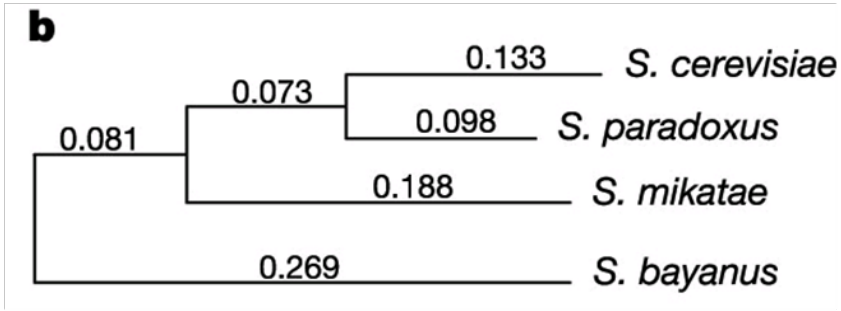
This analysis gives the opposite result when applied to orthologous triples or quads of $E< E_{cut}$ sites in the *Saccharomyces* tree. The E histogram of either site ensemble clearly matches the functional energy distribution Q (derived from the *Scer* genome). No free parameters!

Phylogenetic Tree Simulation Strategy



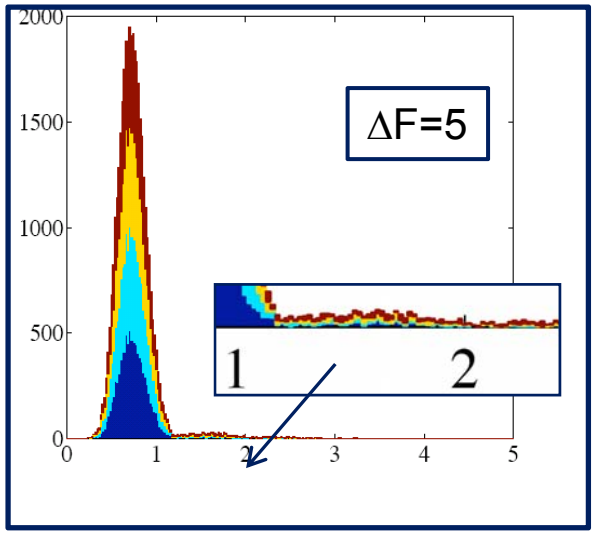
Fitness F(E) can't become arbitrarily negative! At some F_{min}, the site must lose function and we will set F= F_{min} for all higher energies. Evolution is thermal motion (kT=1) in potential U=-F (don't forget that density of states rises rapidly with E). States initially "in the well" have a Kramers-Wannier escape rate (loss of function eventw!) depending sensitively on the choice of ΔF. How to determine?

Attempt to estimate ΔF by simulating multi species evolution via KO: choose ancestor sequence from Q(σ), evolve two copies independently along first two branches; repeat duplication and evolution at later branches. Obtain four progeny sites (and energies). Repeat to get an ensemble of locus histories. Adjust ΔF to match observed loss rate?



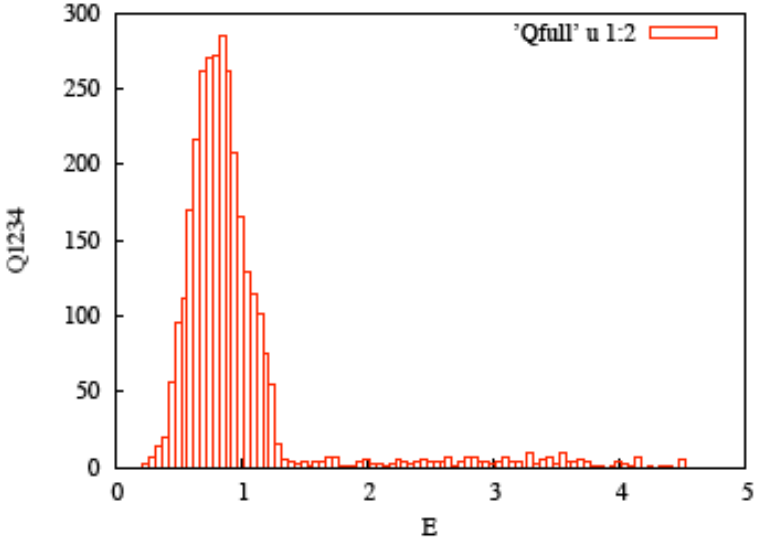
Kellis et al phylogenetic tree with intergenic-based branch times

Quad Evolution Simulation Results



"Well-motivated" choice of ΔF allows a small fraction of sites to "escape" (one species out of the four losing the locus dominates). We can dial that fraction up or down by small changes in ΔF.

Real histogram of quad site energies when at most one site has "escaped". Lots more to say about evidence for loss of function vs gain of function and point mutation vs more nasty processes.



Conclusions

- That's up to you!