Title: Cellular biology for the theoretical physicist

Date: Dec 01, 2004 02:05 PM

URL: http://pirsa.org/04120000

Abstract:

Cellular Biology and Theoretical Physics

Do the tools, methods (and people) of theoretical physics have anything useful to contribute to modern cellular biology? Or did Delbruck and Szilard have all the fun?

Curtis Callan (Princeton)
with lots of help from W. Bialek

Abstract

Each cell in our body contains the same genetic information, coded in a single DNA molecule. Via gene regulation, each cell controls which proteins are made and what the cell actually `does'. The core mechanism of regulation is that the expression of genes is influenced by the binding of protein molecules (transcription factors) to particular short segments of DNA sequence lying near the sequences which code for protein. Despite nearly fifty years of rapid progress in unraveling this mechanism, deep physical questions, regarding specificity, kinetics and noise, remain imperfectly understood. Although these questions emerged from the study of a particular biological system, they apply broadly across biology.

From our point of view as physicists, these are questions about the way in which biological function is constrained by physical principles and are fair game for study by theoretical physics (and physicists). These questions have not been resolved by the relentless advance of molecular biology over the last fifty years. However, the recent explosion of quantitative data, due to genome sequencing, expression profiling, etc. have placed these questions in a new context--one which presents an opportunity to address central theoretical problems of biology from a physicist's point of view.

I will discuss some aspects of this vast and fascinating topic as seen from my own limited experience in dabbling in biology.

Some reasons why the question is timely

Biology as it is practiced today looks more and more like physics: quantitative experiments, large volumes of data, sophisticated data analysis, models, ...

Physics teaches us that models and data analysis must be guided by formal theory: Qualitatively striking phenomena demand new mathematical structures

Physics is not just a methodological model: Cells often operate in a regime where physical constraints are important - limits to specificity, precision, noise, ...

The genomic revolution (organism sequencing, expression profiling, ...) has brought these issues into sharper focus .... we need much more than "bio-informatics" to extract meaning from the mass of data being produced.

Theoretical physics provides a reservoir of people and ideas which are well-suited to take up the challenges of the new biology (that's our opinion anyway).

To make this more concrete, I will describe a few cases where sophisticated theoretical approaches are being used to address real problems in cellular biology.

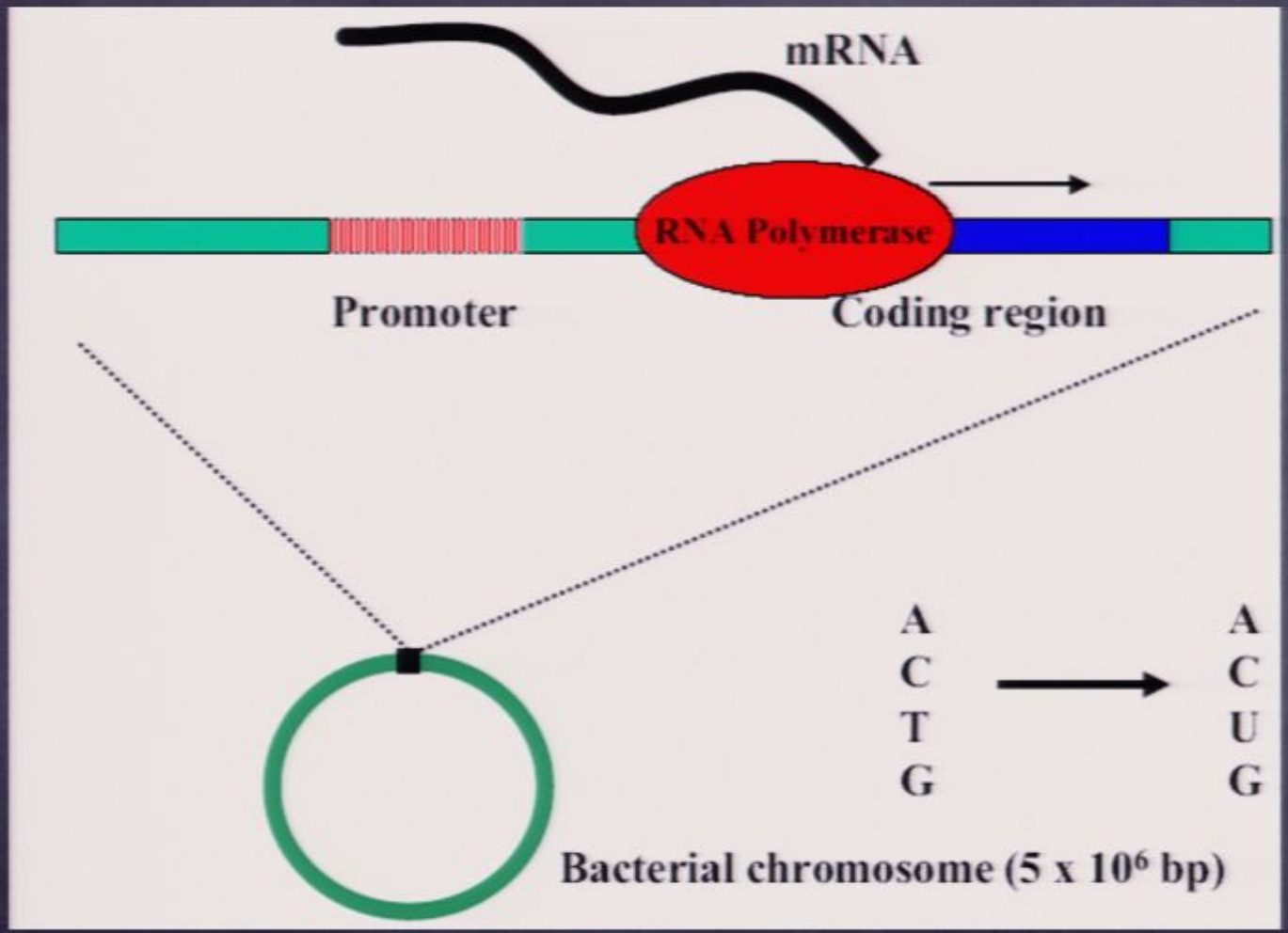
Neuroscience as "existence proof" that theoretical physics has something to say about biology:

theoretical physics ideas	new fields of experiment
coding of sensory signals in neural spike trains should be an efficient - perhaps optimal - code ... must be matched to the distribution of inputs (Bialek et al)	information in spike timing; neural code adapts to input statistics; info adaptation is as fast as possible (de Ruyter, Meister, Berry, et al)
electrical dynamics of neurons determined by ion channels, but overly sensitive to numbers of different channel types ... "self tuning" mechanisms needed for robustness (Abbott et al)	neurons do "remodel" their channel numbers; novel homeostatic regulation mechanisms observed (Marder, Turrigiano, et al)
computing with attractors ... network dynamics for memory, recall, optimization, ... (Hopfield, Seung, et al)	stabilization of eye position as the prototypical short-term memory (Tank et al)

Illustrative examples from current work on problems in cell biology:

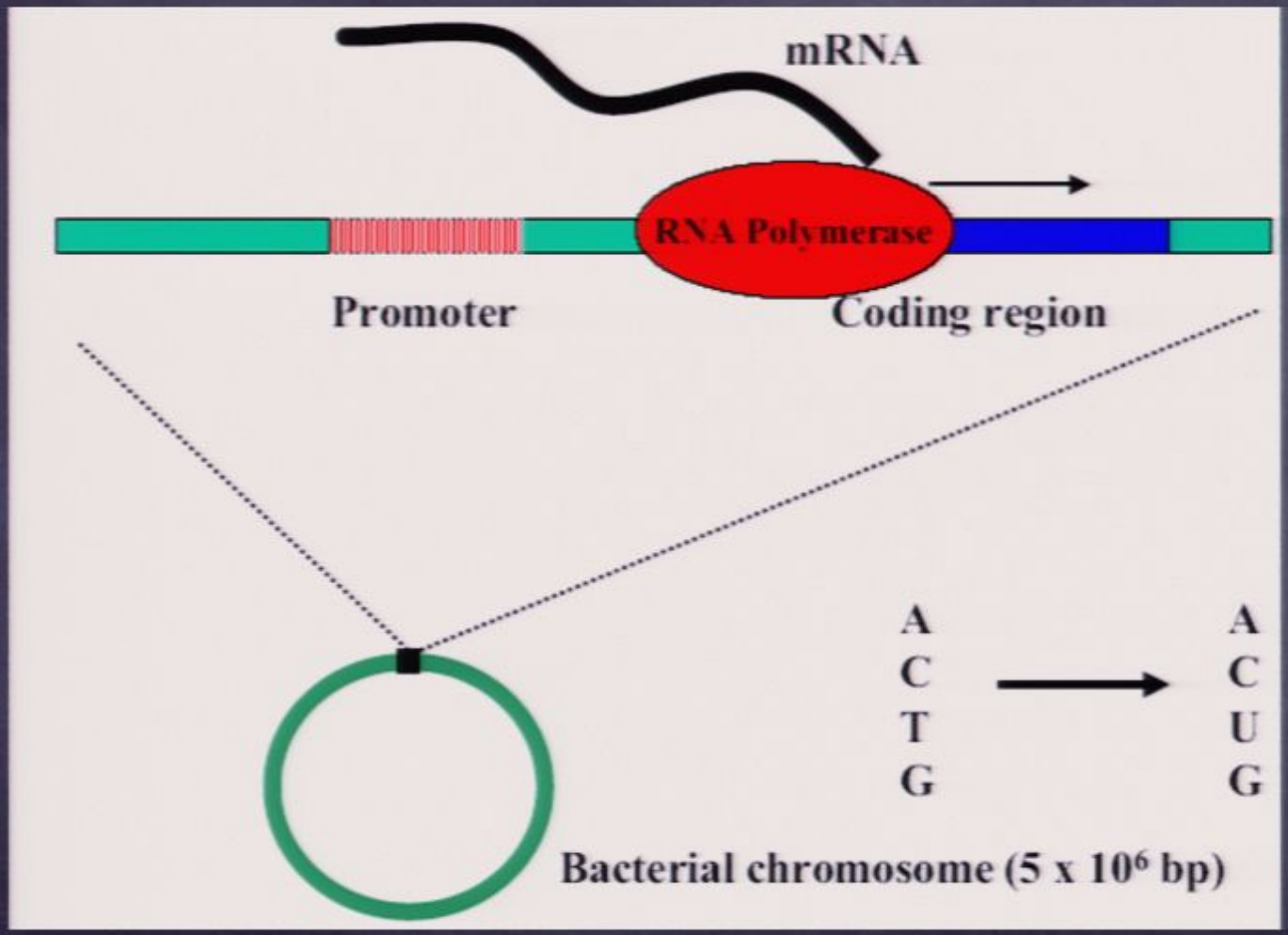
- Optimization principles for identifying transcription factor targets
- Evolutionary comparisons as a tool for learning about transcription factor binding
- Noise, small numbers and stochastic aspects of gene expression dynamics
- Coordinate-independent approaches to finding groups of co-regulated genes

Cartoon Overview of Gene Expression



Genes are transcribed by a protein complex (RNAP) and ultimately translated into protein by the ribosome (triplets of bases are read as one out of twenty amino acids via the "genetic code"). Special transcription factor proteins (TFs) control RNAP binding to promoters.

Cartoon Overview of Gene Expression



Genes are transcribed by a protein complex (RNAP) and ultimately translated into protein by the ribosome (triplets of bases are read as one out of twenty amino acids via the "genetic code"). Special transcription factor proteins (TFs) control RNAP binding to promoters.

Transcription Factors: Proteins that Regulate Gene Expression

Basic Mechanism: TFs bind to short (noncoding?) DNA sequences to modify expression level of nearby genes. Complex circuits are made.

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

PWM Method: One-point correlation model of site statistics/binding energy (Berg+vonHippel). Useful reduced-dimension approach.

Significance: To analyze GRNs at next level of complexity, need quantitative model for how TF finds its DNA. PWM is the only game in town!

Issues: Basic stat mech of TF binding; going from sequence to energy by optimization; problems/solutions; constraining parameters

Masses of Genomic Information are Available

256 336	++	thrL thrA
5021 5233	++	thrC b0005
5531 5682	+-	b0005 yaaA
6460 6528	--	yaaA yaaJ
7960 8237	-+	yaaJ talB
9192 9305	++	talB mog
9894 9927	+-	mog yaaH
10495 10642	--	yaaH b0011
11316 11381	+-	htgA yaaI
11787 12162	-+	yaaI dnaK
14080 14167	++	dnaK dnaJ
15299 15444	++	dnaJ yi81_1
16178 16750	--	yi82_1 gef
16961 17488	-+	gef nhaA
18656 18714	++	nhaA nhaR
19621 19810	+-	nhaR insB_1
20509 20814	--	insA_1 rpsT
21079 21180	-+	rpsT b0024
21400 21406	++	b0024 ribF
22349 22390	++	ribF ileS
25702 25825	++	lspA slpA

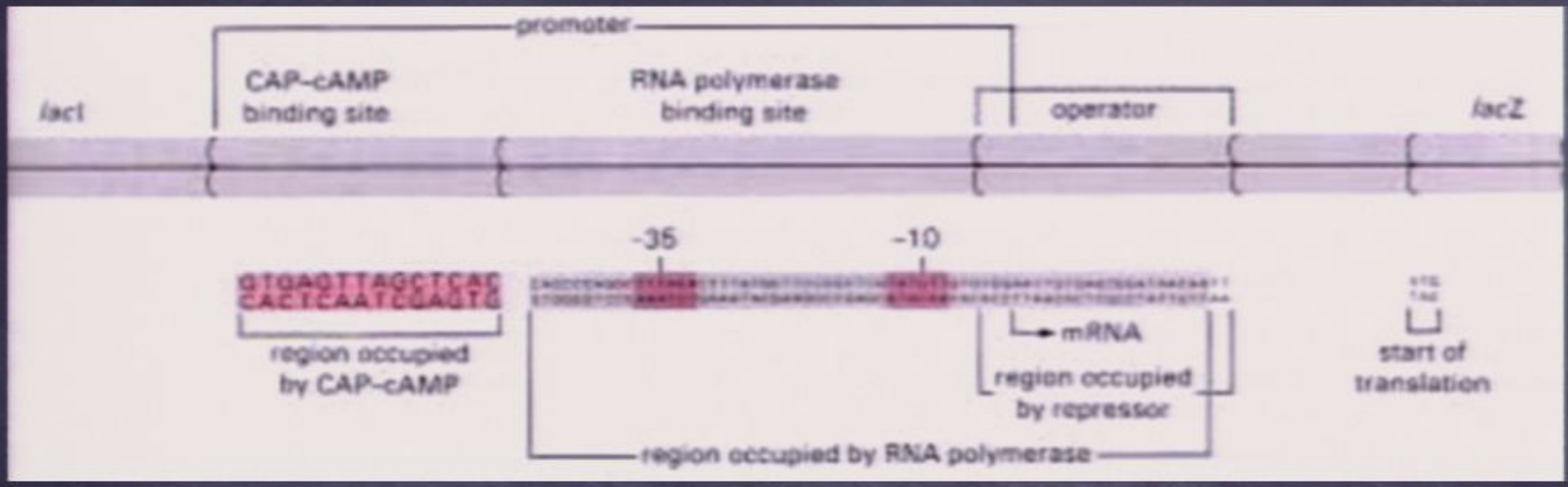
Some 180 bacterial genomes are completely sequenced. The genome and lots of other information is available from www.ncbi.nih.gov

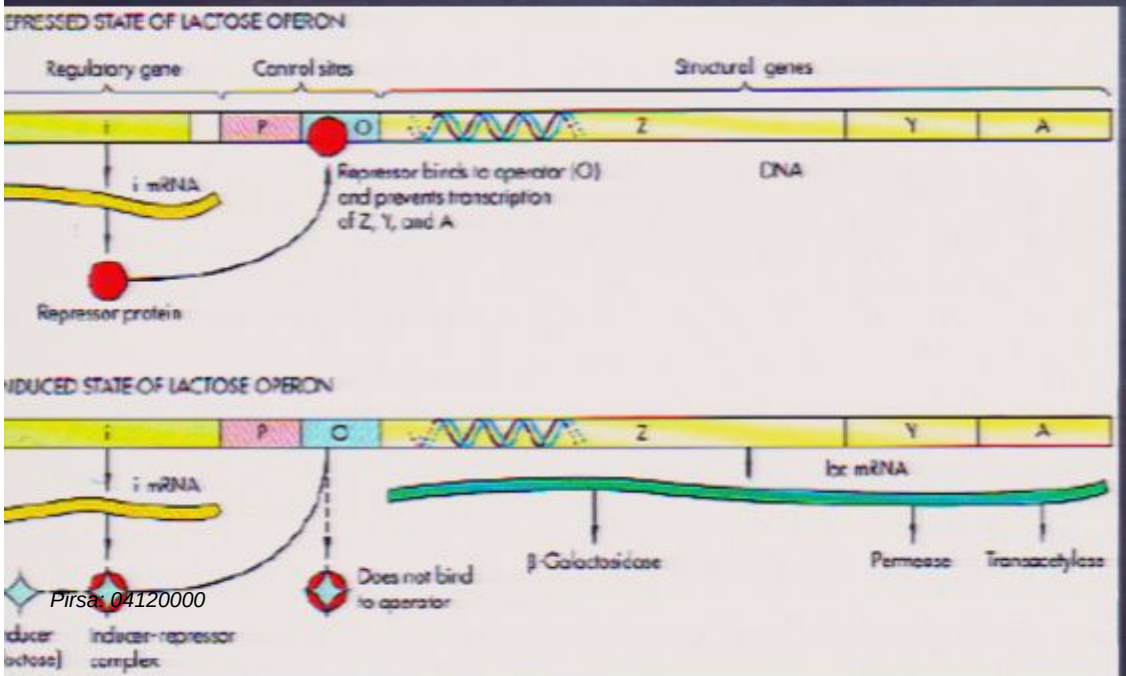
E. coli has 3400 genes. Online protein tables tell you the gene coordinates, name and function. Most common annotation: *unknown*

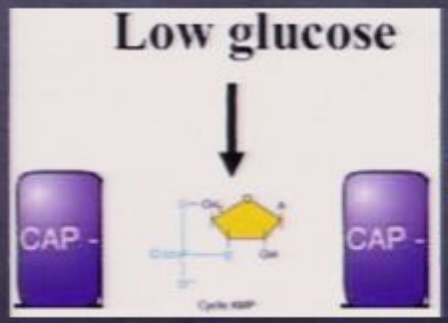
Non-coding regions can be derived from these tables. Cover relatively little real estate, but most TF binding sites lie there (for obvious reasons).

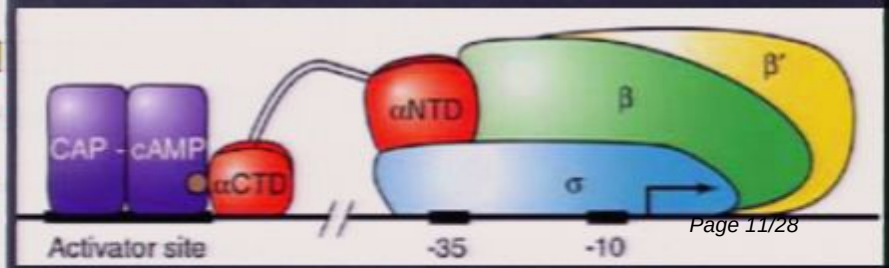
Genomic data is highly non-random: intelligent statistical analysis needed to unravel gene expression network

Gene Regulation by LacI and Crp (or CAP)









Transcription Factor Binding Site Statistics

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

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

Matrix Model for Sequence-Dependent Binding

Introduced in mid-80s by Berg + von Hippel (still the main contender)

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

PWM: 4xL matrix $e_i(b_a)$ contains all info about sequence specificity of binding.
Compressed rep'n of complex physics!

Different TFs will have different PWMs. The elements of the PWM can be estimated by *in vitro* biochemical experiments (Stormo *etal*), but this is really hard work. B+vH proposed an algorithm for estimating *energy* from *statistics* of the known binding sites (evolution as statistical mechanics):

If $N_i(b)$ is the number of occurrences of base b at position i:

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

Normalization: most common base is assigned e=0 by convention
Consensus site: all sub-energies = 0; may not exist in the actual genome
Pseudocount: rational approach to $N_i(b)=0$ observation (blowup issue)

Energy from Sequence by Optimization

Special sites $s^1, \ldots, s^K$ with known relative affinities ρ_k

Linear site energy function (PWM): $E(s_1 s_2 \ldots s_L) = \Sigma_{a=1}^{L} \epsilon_a(s_a)$

Probability of finding TF bound to site r: $p(r) = e^{-E(r)} / \Sigma_{u \in G}\, e^{-E(u)}$

Probability to fish out N copies of G with TF bound to n_1 times to site s_1 (etc.) with n_i proportional to the relative affinities is $Prob_N = \hat{p}(s^1)^{n_1} \hat{p}(s^2)^{n_2} \ldots \hat{p}(s^K)^{n_K}$

Maximize that probability by varying the elements of the PWM:

Unlikelihood:

$$\begin{aligned} U = & -\log(Prob_N)/N = -\Sigma_{k=1}^{K} \tfrac{n_k}{N} \log(p(s^k)) \\ & = \Sigma_{k=1}^{K} \rho_k E(s^k) + \log(\Sigma_{u \in G}\, e^{-E(u)}) \end{aligned}$$

Minimize by varying energy parameters:

$$\frac{\partial U}{\partial \epsilon_a(b)} = \Sigma_{k=1}^{K} \rho_k t_b^a(s^k) - \frac{\Sigma_{u \in G}\, t_b^a(u) \exp^{-E(u)}}{\Sigma_{u \in G} \exp^{-E(u)}} = 0$$

$t_b^a(s) = 1, 0$ depending on whether s has base b at position a or not

Minimum identifies `best' energy parameters given the data. Using random genome yields the B+vH formula of the previous slide!

Matrix Model for Sequence-Dependent Binding

Introduced in mid-80s by Berg + von Hippel (still the main contender)

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

PWM: 4xL matrix $e_i(b_a)$ contains all info about sequence specificity of binding.
Compressed rep'n of complex physics!

Different TFs will have different PWMs. The elements of the PWM can be estimated by *in vitro* biochemical experiments (Stormo *etal*), but this is really hard work. B+vH proposed an algorithm for estimating *energy* from *statistics* of the known binding sites (evolution as statistical mechanics):

If $N_i(b)$ is the number of occurrences of base b at position i:

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

Normalization: most common base is assigned e=0 by convention
Consensus site: all sub-energies = 0; may not exist in the actual genome
Pseudocount: rational approach to $N_i(b)=0$ observation (blowup issue)

Energy from Sequence by Optimization

Special sites $s^1, \ldots, s^K$ with known relative affinities ρ_k

Linear site energy function (PWM): $E(s_1 s_2 \ldots s_L) = \Sigma_{a=1}^{L} \epsilon_a(s_a)$

Probability of finding TF bound to site r: $p(r) = e^{-E(r)} / \Sigma_{u \in G}\, e^{-E(u)}$

Probability to fish out N copies of G with TF bound to n_1 times to site s_1 (etc.) with n_i proportional to the relative affinities is $Prob_N = \hat{p}(s^1)^{n_1} \hat{p}(s^2)^{n_2} \ldots \hat{p}(s^K)^{n_K}$

Maximize that probability by varying the elements of the PWM:

Unlikelihood:

$$\begin{aligned} U = & -\log(Prob_N)/N = -\Sigma_{k=1}^{K} \tfrac{n_k}{N} \log(p(s^k)) \\ & = \Sigma_{k=1}^{K} \rho_k E(s^k) + \log(\Sigma_{u \in G}\, e^{-E(u)}) \end{aligned}$$

Minimize by varying energy parameters:

$$\frac{\partial U}{\partial \epsilon_a(b)} = \Sigma_{k=1}^{K} \rho_k t_b^a(s^k) - \frac{\Sigma_{u \in G}\, t_b^a(u) \exp^{-E(u)}}{\Sigma_{u \in G} \exp^{-E(u)}} = 0$$

$t_b^a(s) = 1, 0$ depending on whether s has base b at position a or not

Minimum identifies `best' energy parameters given the data. Using random genome yields the B+vH formula of the previous slide!

Transcription factor binding statistics: LacI

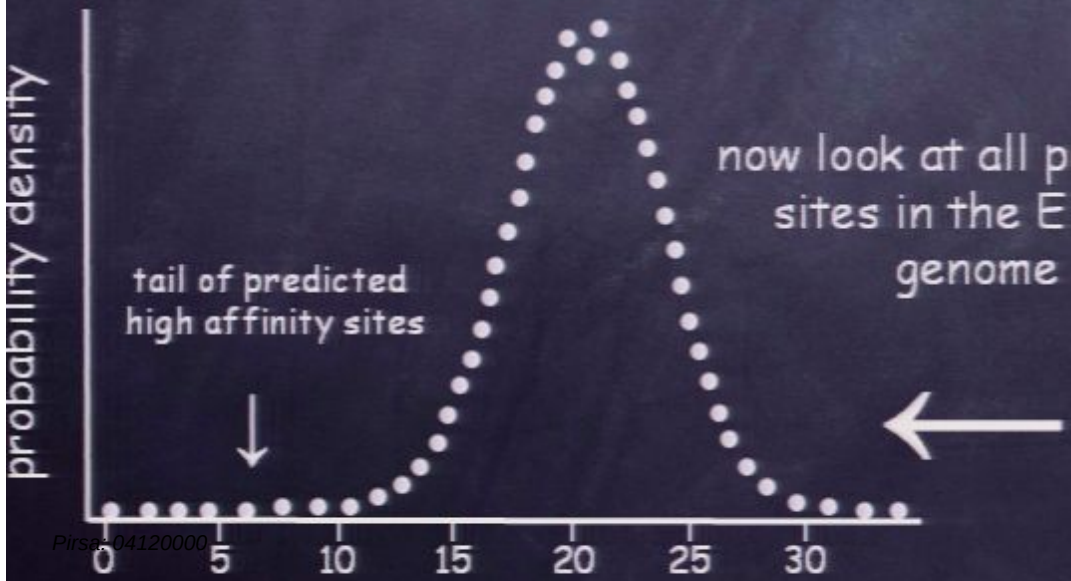
Given sequences for the strongest sites, construct the sequence-dependent energy function; run it over the genome to find site binding energy distribution; ....

sequences of the three LacI sites (21 bp)

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

B+vH rule assigns entries in PWM to match observed frequencies

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



now look at all possible sites in the E Coli genome

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

Energy from Sequence by Optimization

Special sites $s^1, \ldots, s^K$ with known relative affinities ρ_k

Linear site energy function (PWM): $E(s_1 s_2 \ldots s_L) = \Sigma_{a=1}^{L} \epsilon_a(s_a)$

Probability of finding TF bound to site r: $p(r) = e^{-E(r)} / \Sigma_{u \in G}\, e^{-E(u)}$

Probability to fish out N copies of G with TF bound to n_1 times to site s_1 (etc.) with n_i proportional to the relative affinities is $Prob_N = \hat{p}(s^1)^{n_1} \hat{p}(s^2)^{n_2} \ldots \hat{p}(s^K)^{n_K}$

Maximize that probability by varying the elements of the PWM:

Unlikelihood:

$$\begin{aligned} U = & -\log(Prob_N)/N = -\Sigma_{k=1}^{K} \tfrac{n_k}{N} \log(p(s^k)) \\ & = \Sigma_{k=1}^{K} \rho_k E(s^k) + \log(\Sigma_{u \in G}\, e^{-E(u)}) \end{aligned}$$

Minimize by varying energy parameters:

$$\frac{\partial U}{\partial \epsilon_a(b)} = \Sigma_{k=1}^{K} \rho_k t_b^a(s^k) - \frac{\Sigma_{u \in G}\, t_b^a(u) \exp^{-E(u)}}{\Sigma_{u \in G} \exp^{-E(u)}} = 0$$

$t_b^a(s) = 1, 0$ depending on whether s has base b at position a or not

Minimum identifies `best' energy parameters given the data. Using random genome yields the B+vH formula of the previous slide!

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	0.809	AATTGTGAGCGGATAACAATT
365446	0.799	AATTGTTATCCGCTCACAATT
365145	4.068	AAATGTGAGCGAGTAACAACC
365145	4.058	GGTTGTTACTCGCTCACATTT
365638	6.449	GGCAGTGAGCGCAACGCAATT
365638	6.439	AATTGCGTTGCGCTCACTGCC

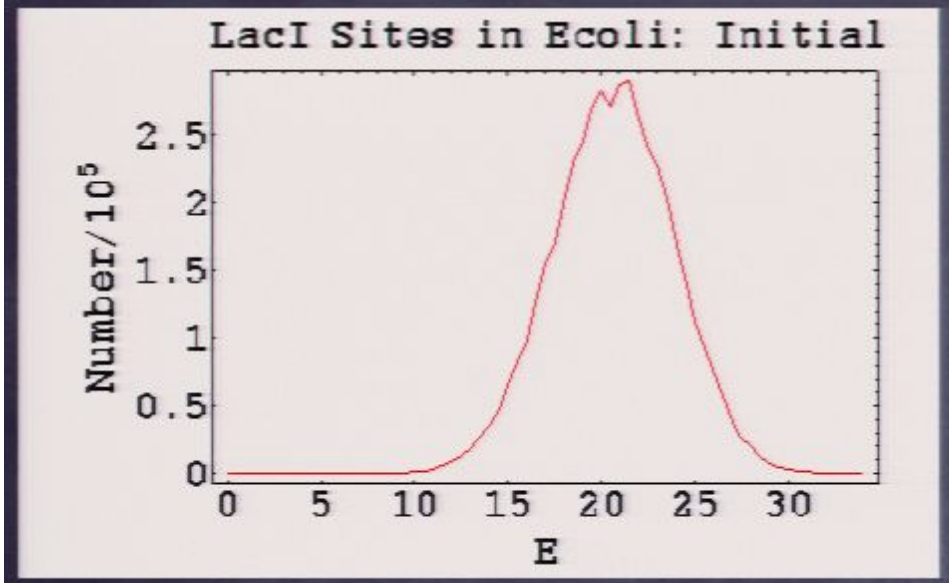
B+vH rule assigns entries in PWM to match observed frequencies

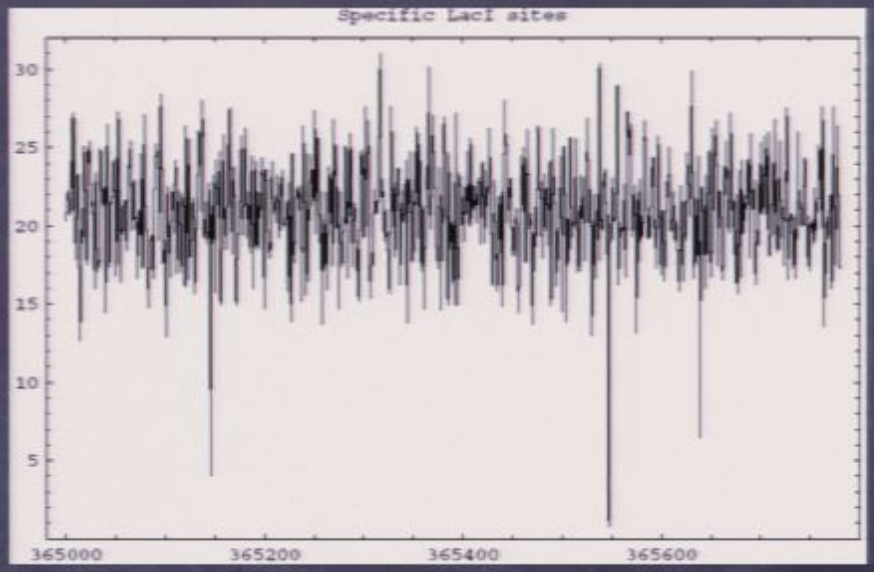
A	C	G	T
0.000	1.609	0.511	1.609
0.000	1.609	0.511	1.609
0.916	0.916	1.609	0.000
1.099	1.792	1.792	0.000
1.946	1.946	0.000	1.946
1.792	1.099	1.792	0.000
1.609	1.609	0.000	0.511
0.000	1.792	1.792	1.099
1.386	0.693	0.000	0.288
1.609	0.000	0.916	0.916
1.386	0.000	0.000	1.386
0.916	0.916	0.000	1.609
0.288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

now look at all possible sites in the E Coli genome

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

Side Remark: Energy Landscape Issue





Sites close to zero (strongest affinity) are functional: they are extreme outliers, rare.

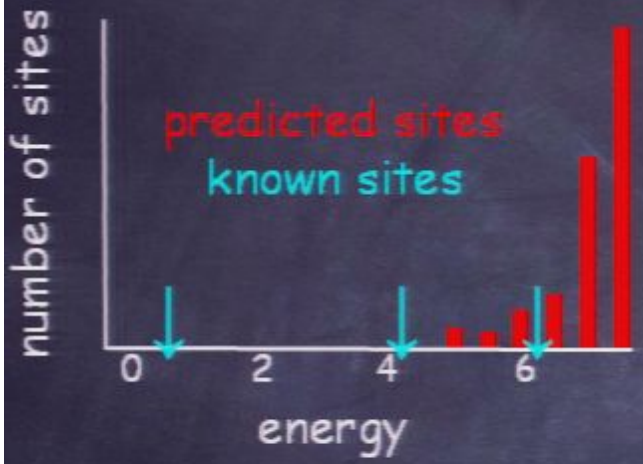
Sites with less extreme energy are numerous, not functional per se, but affect rate of TF diffusion on the DNA: whole spectrum determines TF response time.

There seems to be a conflict between specificity (strong binding to few sites) and known speed of response (transcription turns on in minutes or less).

Physics of diffusion leads Mirny/Slutsky to propose 2-state picture of TF binding to DNA (non-specific vs specific). Structural studies quite neatly confirm this story.

A problem for LacI and its "solution" by optimization

(expand the tail from previous plot)



Main site that governs LacI transcription is well separated from the rest of the genome.

But 2 subsidiary sites are not: they compete with ~ 10 other sites elsewhere in the genome.

There are only a few LacI molecules in the cell, so how do they find their true active sites?

A "gap" in the energy spectrum would be good!

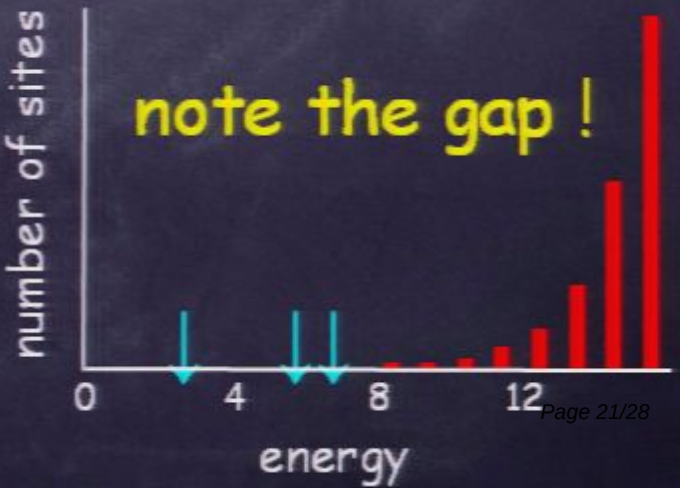
nitial PWM was a rough guess. Can we do better?

eed strong binding to known sites and weak nding to the rest of the genome.

eturn to our optimization problem for (3x21)/2 dependent entries of the PWM, but do it for the tual genome, not a random approximation.

olve by relaxation/MC to find best parameters! odest computer exercise, even for large genome.

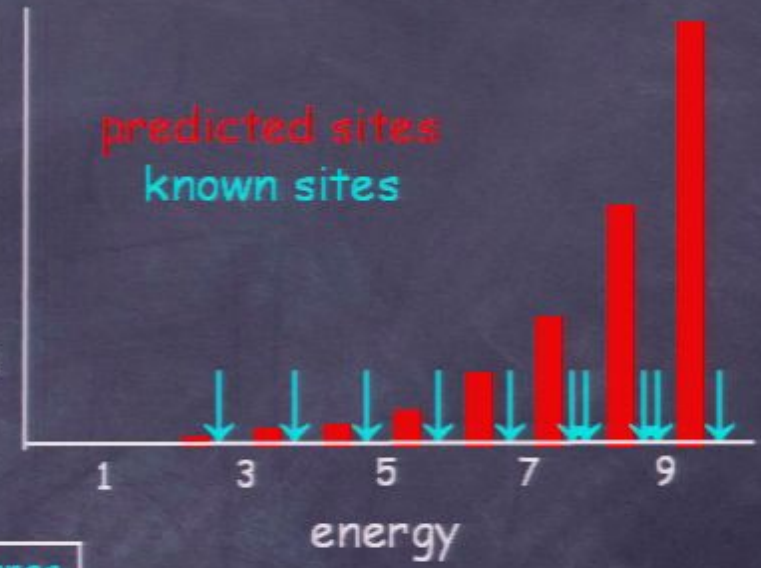
(also, statistics of non-functional sites may be related to kinetics of finding functional sites)



Broad regulator Crp is different ...

Even with optimization, known sites are buried in a dense background of predicted sites ... known sites also seem to have a very broad range of affinities

Crp is a broad metabolic regulator, known to regulate many genes, unlike LacI which regulates one operon. Some of the 48 known sites:



location	"energy"	sequence	flanking genes
70158	6.188	AAGTGTGACGCCGTGCAAATAA	araB araC
431345	6.357	AACTGTGAAACGAAACATATTT	tsx yajI
431384	9.873	GTGTGTAAACGTGAACGCAATC	tsx yajI
702991	6.714	TTTTGTGAGTTTTGTCACCAAA	nagB nagE
791335	6.900	AAGTGTGACATGGAATAAATTA	galE modF
1019443	7.764	ATGCCTGACGGAGTTCACACTT	ompA sulA
1236678	5.007	AGATGTGAGCCAGCTCACCATA	ycgB dadA
2229736	6.836	ATTTGCGATGCGTCGCGCATTT	yohK cdd
2229786	4.218	TAATGAGATTCAGATCACATAT	yohK cdd
2350502	4.464	ATGTGTGCGGCAATTCACATTT	glpT glpA
2350552	11.720	AAACGTGATTTCATGCGTCATT	glpT glpA

Has our simple model failed?

Without simple models, how will we get to the network level?

Might the myriad predicted sites be functional?

Strong tendency for low-E sites to be in non-coding regions ...

Bringing in Evidence from Evolution

"Spurious" sites could in fact be functional: they lie in non-coding regions. If so, they should have clear orthologs in nearby organisms. Experimental test?

Strategy: take ecoli and salmonella; find all orthologous intergenic regions; align them (ClustalW); assemble population of predicted intergenic ecoli sites for Crp; they align to 22bp sequences in salmonella; defines a population of sites in salmonella; ask if mutation pattern is nonrandom.

Some data: ~3500 intergenic regions in both genomes. Call them orthologous if flanked by same genes (by name). ~1500 orthologous intergenic regions! Mean intergenic mutation rate (after alignment) is 25% (quite a lot!).

Ecoli sites are selected using their Crp PWM energies. Salmonella sites are generated purely by alignment, have many mutations: no a priori need to be strong binders. N.B. Ecoli and salmonella Crp are virtually identical (1 aa).

Key points: We don't expect (don't see) strict sequence conservation Between orthologous sites. Binding energy, not sequence is conserved. Also, the useful tests are population-based.

Orthology and Alignment of Genomes + Sites

Example of 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

      xxxxxxxxxxxxxxxxxxxxx
----TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTA
AGTAAAAAATGTGATGTTCTGCAAACTTTACTGCTAATTGGCTGTTTTTGAACTACTGTA
     ***  **** **    ** ******* ***** **  ********    * ****

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

ATCATCCTCGCACCAGTCGACGACGGTTTACGCTTTACGTATAGTGGCGACAATTTTTTT
ATCCTCTTTTCGTCAGTCGACGAAAGATTGCGCTTTACGTATAGTGGCGGCAATTTTTTT
*** ** *  *  **********  * ** ******************** **********

T--ATCGGGAAATC-TCA
TGTATCAGGAAATAATTA
*  *** ******  * *
```

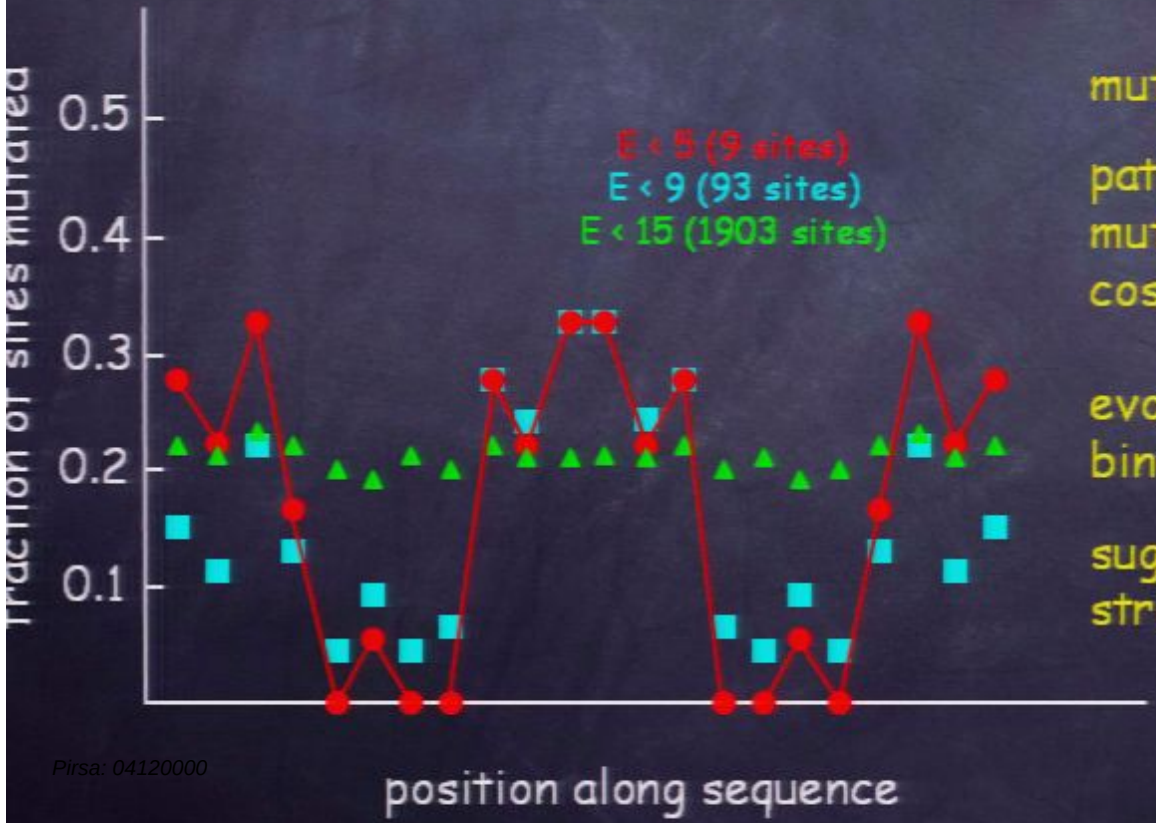
Alignment of related sequences amounts to finding the most parsimonious way of mutating one into the other (including possibility of creating gaps). Powerful software readily available: ClustalW used here. Can also search for "most likely" ancestor sequence of the descendants (well-studied subject in comp-bio).

Evolution reveals function of new Crp sites

Compare E Coli to Salmonella.

Take a predicted strong binding site in E Coli and find the string to which it aligns in Salmonella ... the sequence is not (quite) the same; note locations of mutations.

Collect data on a population of such sites ...



mutations are highly non-random

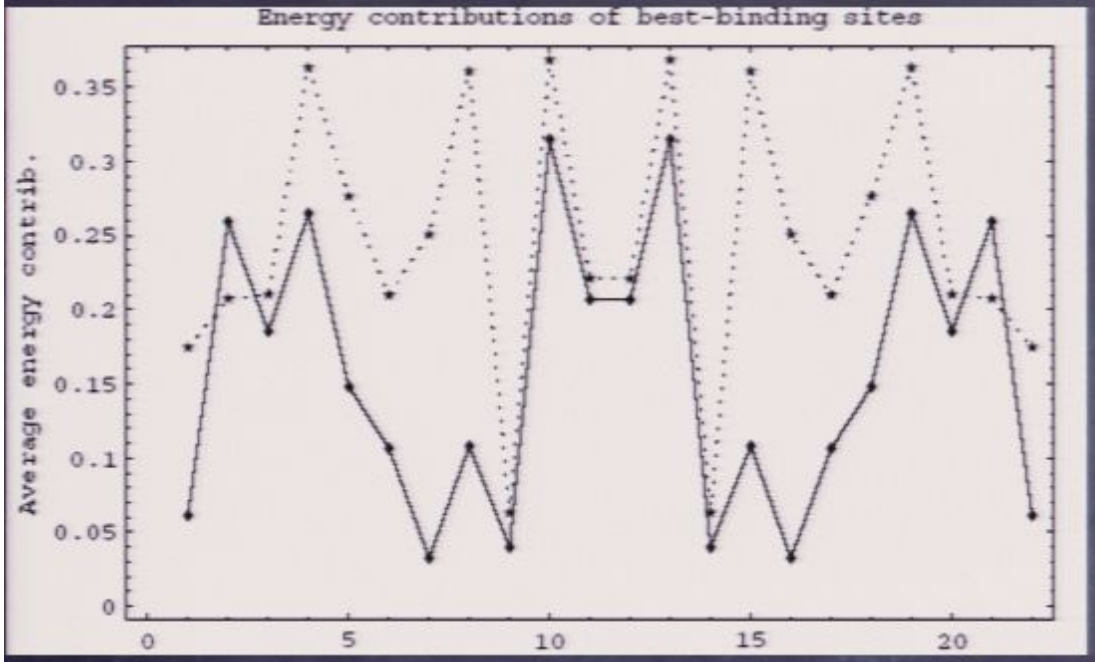
pattern matches Crp weight matrix: mutations rare where wrong base costs large energy

evolution conserves the (theoretical) binding energy - more than sequence

suggests that predicted sites with strong binding are functional!

(CT Brown & C Callan, PNAS 2004)

Binding Energy, Not Sequence, is Conserved



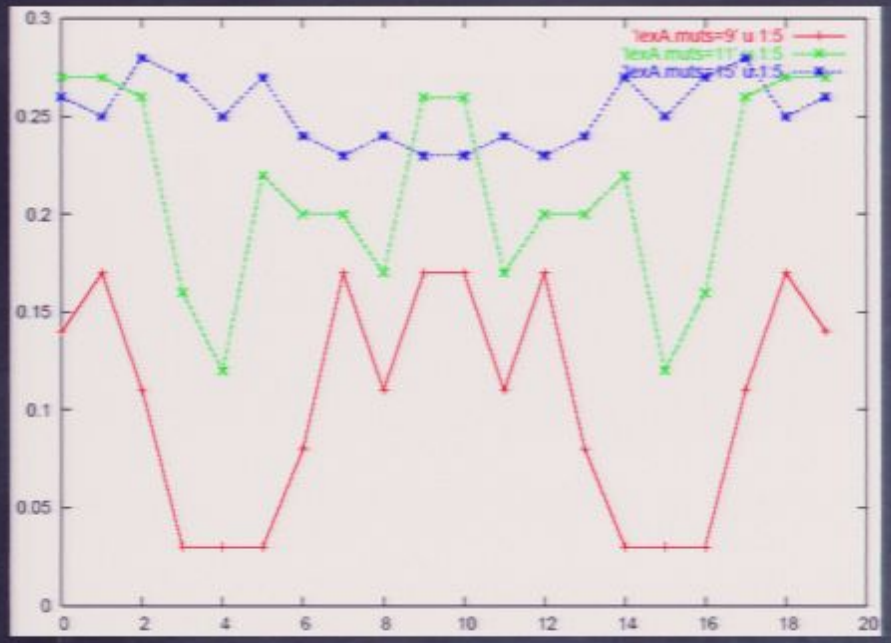
Contribution of different site positions to Crp energy averaged over top 100 (non-coding) sites in genome

Comparison with mutation profile is instructive: mutation is least likely in Positions contributing most strongly to binding energy.

Solid line shows average over 100 best-binding sites in non-coding regions

- Binding energies correlate between two species (not just sequence conservation).
- Positional mutation profile is a strong function of predicted binding energy.
- Underlying cause must be primarily conservation of site binding energy.
- Suggests that PWM binding energy is a reasonable surrogate for the real thing.

Other Transcription Factors, Other Genomes



lexA in ecoli:
regulator for SOS(lexA) regulon

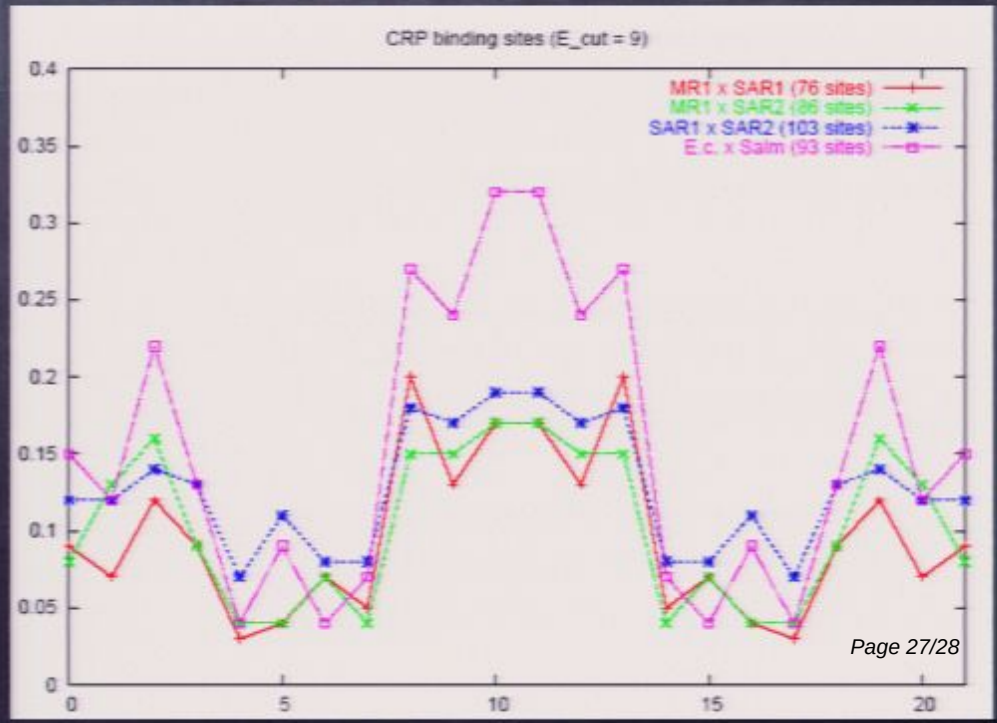
cutoff	total	coding	noncoding	known
1.00	2	0%	100%	1/19
3.00	7	0%	100%	4/19
5.00	24	0%	100%	10/19
7.00	46	11%	89%	16/19
9.00	111	44%	56%	19/19

Nice 20bp dimeric TF with core region (4-6,15-17)

Crp in Sargasso Sea Shotgun Genomes!

Two (uncultured) strains of Shewanella sequenced from sea water! Use ecoli Crp PWM to scan for E<9 binding sites. Familiar mutation profiles emerge!

strain	total	in genes	intergenic	known
S. oneid.	342	41%	59%	27/48
S. SAR1	284	20%	80%	27/48
S. SAR2	355	25%	75%	27/48



Fluctuations, Noise and Genetic Switch Stabilty

- Many cellular processes depend on presence (absence) of a small number of actors (TFs, signaling molecules, photons, ..)
- The associated fluctuations and noise have critical influence on how things work (not always fully appreciated):
  - Sensing chemical gradients (chemotaxis)
  - Stability of genetic switches (phage lysis/lyosogeny)
  - Can a gene do more than just be on or off?
- Remarkably little is known about this: the stochastic properties of cellular events are just beginning to be explored quantitatively (its not just Poisson).