OrthoGibbs & PhyloScan

A comparative genomics approach to locating transcription factor binding sites

Lee Newberg, Wadsworth Center, 6/19/2007

Acknowledgments

Team:

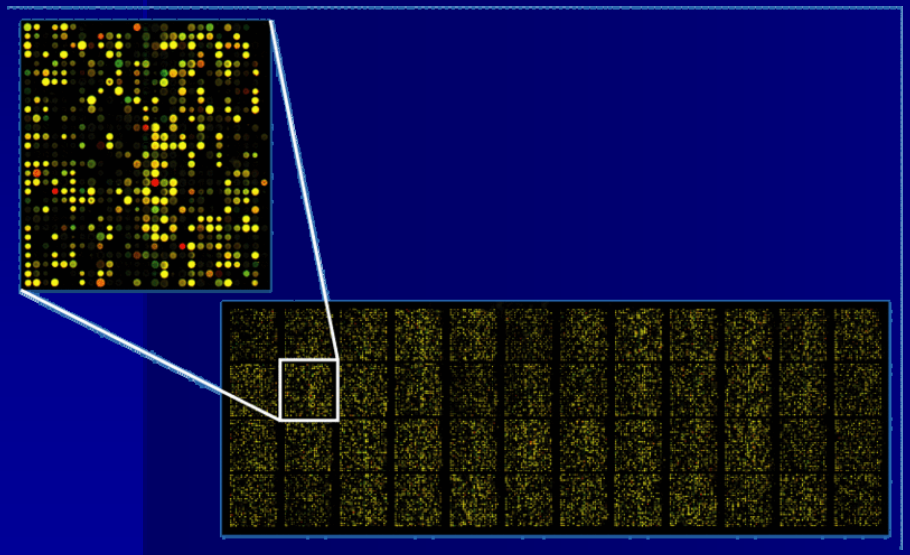
- C. Steven Carmack (Wadsworth)
- Sean P. Conlan (Columbia)
- Charles E. Lawrence (Brown)
- Lee Ann McCue (Pacific Northwest NL)
- Thomas M. Smith (MIT Lincoln Laboratory)
- William A. Thompson (Brown)

Resources:

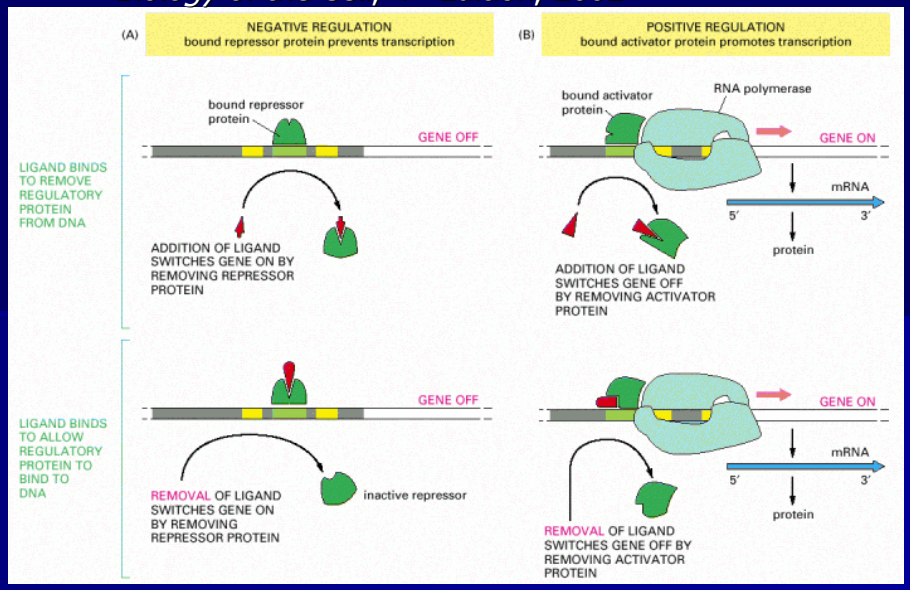
- Wadsworth Center (& use of the CMBS Core Facility)
- Rensselaer Polytechnic Institute
- NIH/NHGRI: K25 "Mentored Career Award" (LAN), R01 (CEL)
- Department of Energy awards (CEL, LAM)

Relevance

The identification and characterization of functional, non-coding DNA sequence elements is important



Alberts, Johnson, Lewis, Raff, Roberts, & Walter, *Molecular Biology of the Cell*, 4th Edition, 2002



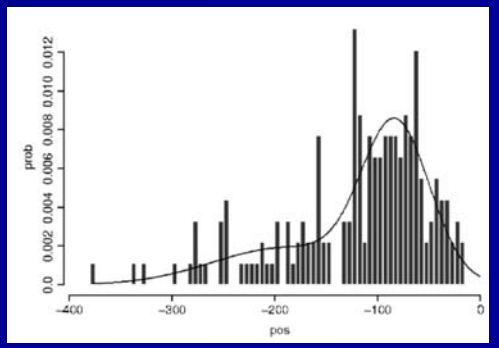
… for the understanding of cell function, differentiation, and pathology

… because the elements affect both the products of genes and when and to what extent the genes are expressed

Previous Approaches

Look for overrepresented DNA patterns (motifs). Additional biases:

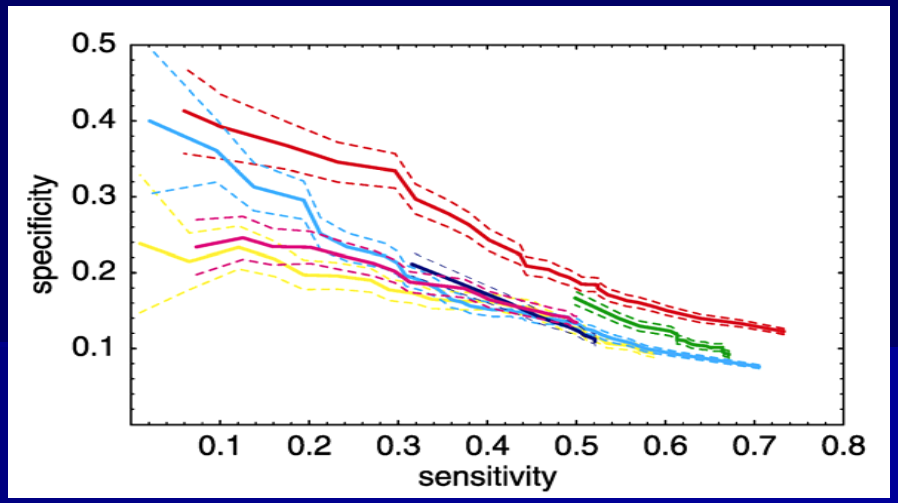
- Frequency (per promoter, genome)
- Positioning (relative to +1, TFBSs)
- Size
- Shape (palindromic, "off"-positions)
- *Ad hoc* evolutionary models



Gibbs Recursive Sampler. Thompson, Rouchka, & Lawrence, *Nucleic Acids Res*, 2003

These are Crp & PurR motifs. Cameron & Redfield, *Nucleic Acids Res,* 2006

However …



Siddharthan, Siggia & van Nimwegen, *PLoS Comput Biol*, 2005

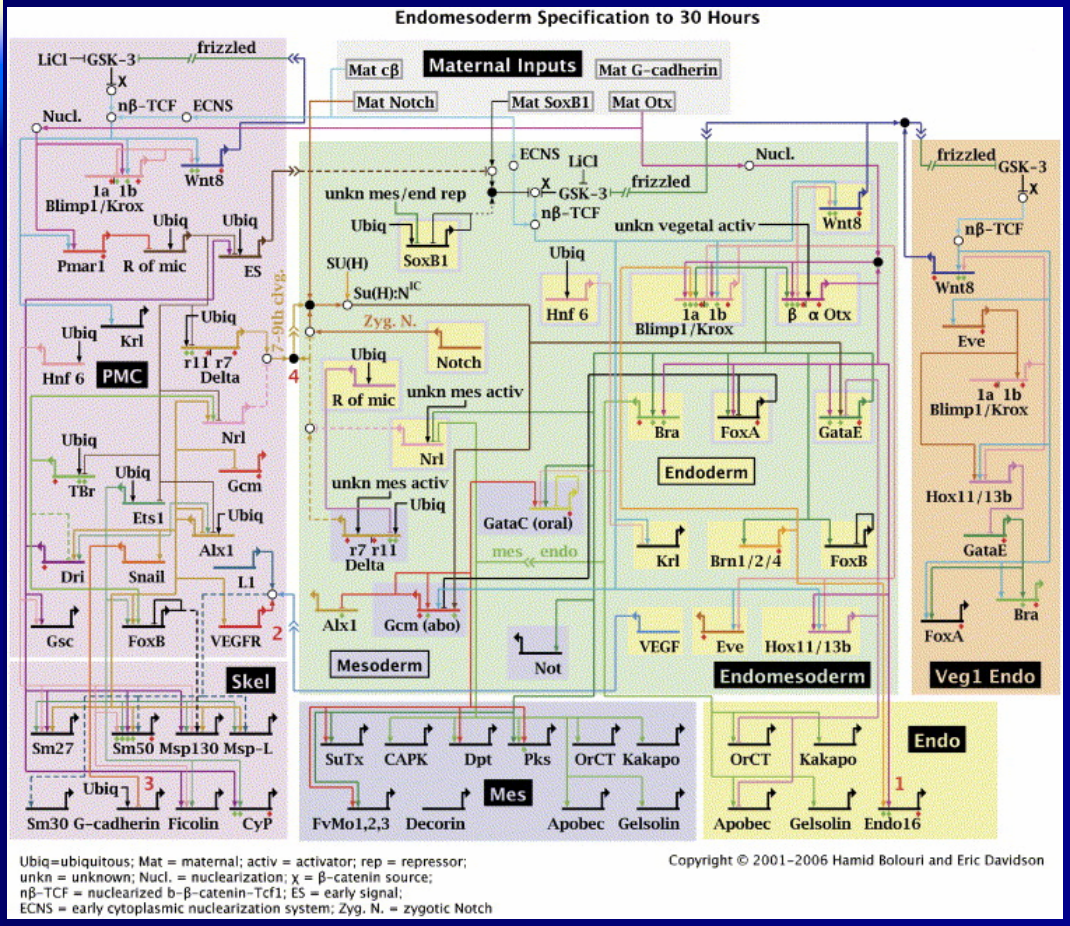
In the results:

- Too many omissions – Low sensitivity
- Too many false discoveries – Low positive predictive value

Furthermore: new sequencing technologies → progress on deciphering gene regulation will lag further behind the production of the experimental results that harbor its understanding

Research Objective

Howard-Ashby, Materna, Brown, Tu, Oliveri, Cameron, & Davidson, *Dev Biol*, 2006



Significantly decrease false positives and false negatives as a fraction of actual sites

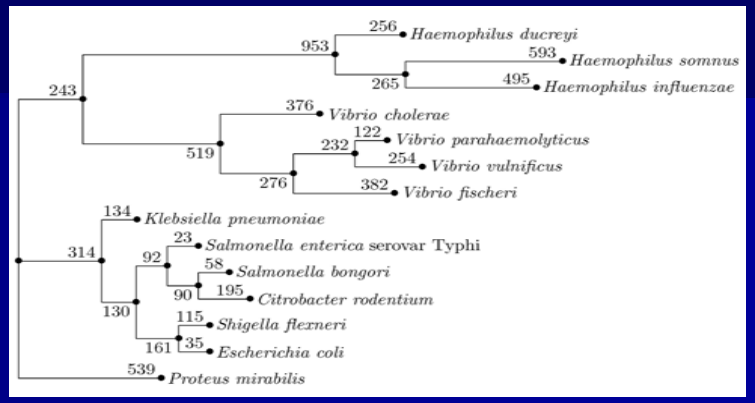
… to greatly ease the task of decoding gene regulatory circuits

Our Approach, 1 of 2

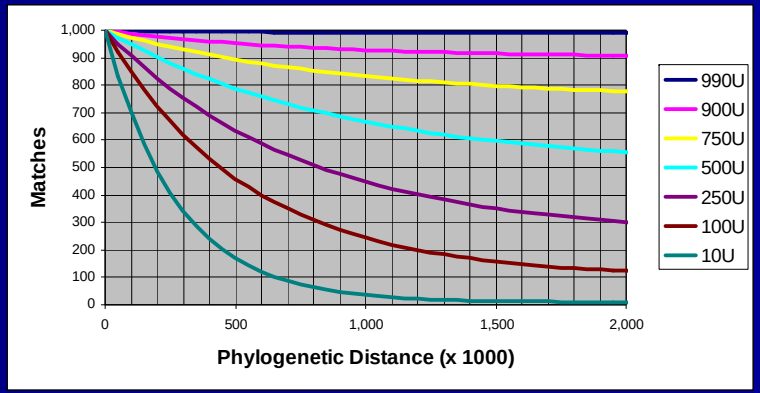
Rigorously model evolution

- Organize species into clades, with multisequence alignments
- Phylogenetic tree for each clade
- Model of selection pressures

Seek binding sites that are consistent with the phylogenetic model – in addition to criteria for over-representation, positioning, size, & shape




```
...1234567890... ... ...1234567890...
GGCCGGTGCTATTACG ... GCACGGAGTTATGCGA  S. cerevisiae
GGTCGGTGCTATCACG ... TCGCGGAGGTATAGGA  S. paradoxus
GGCCTGTGTTATTTCG ... GCGCGGTGTTATACGA  S. mikatae
AACCGGTGTTATTACA ... GCGCGGAGTTATAAAG  S. kudriavzevii
AGACGGTGTTATGGCA ... ACGCGGAGGTATGCGG  S. bayanus
```

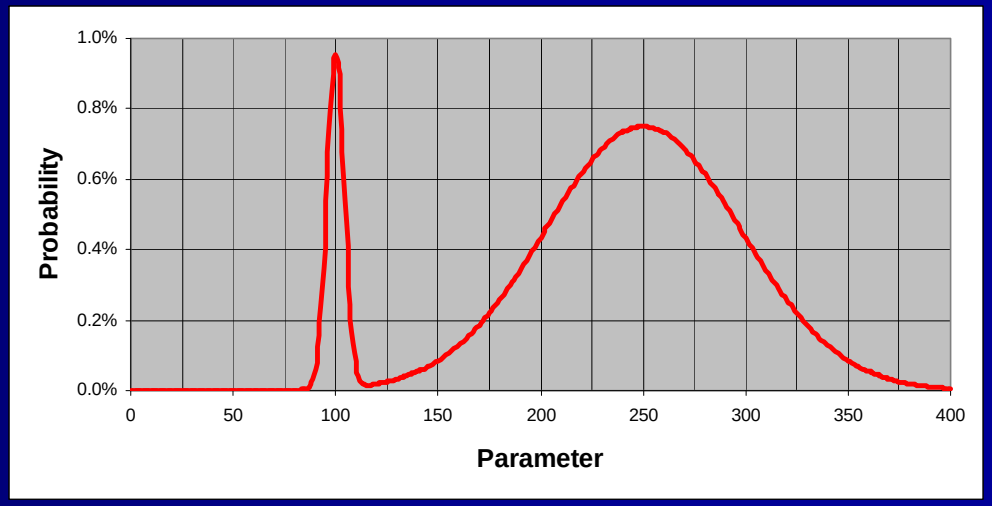


Our Approach, 2 of 2

Employ Centroids

… seek the region of solution space containing the most posterior probability

… rather than the single solution with the most posterior probability



… faster, more accurate

Payoff

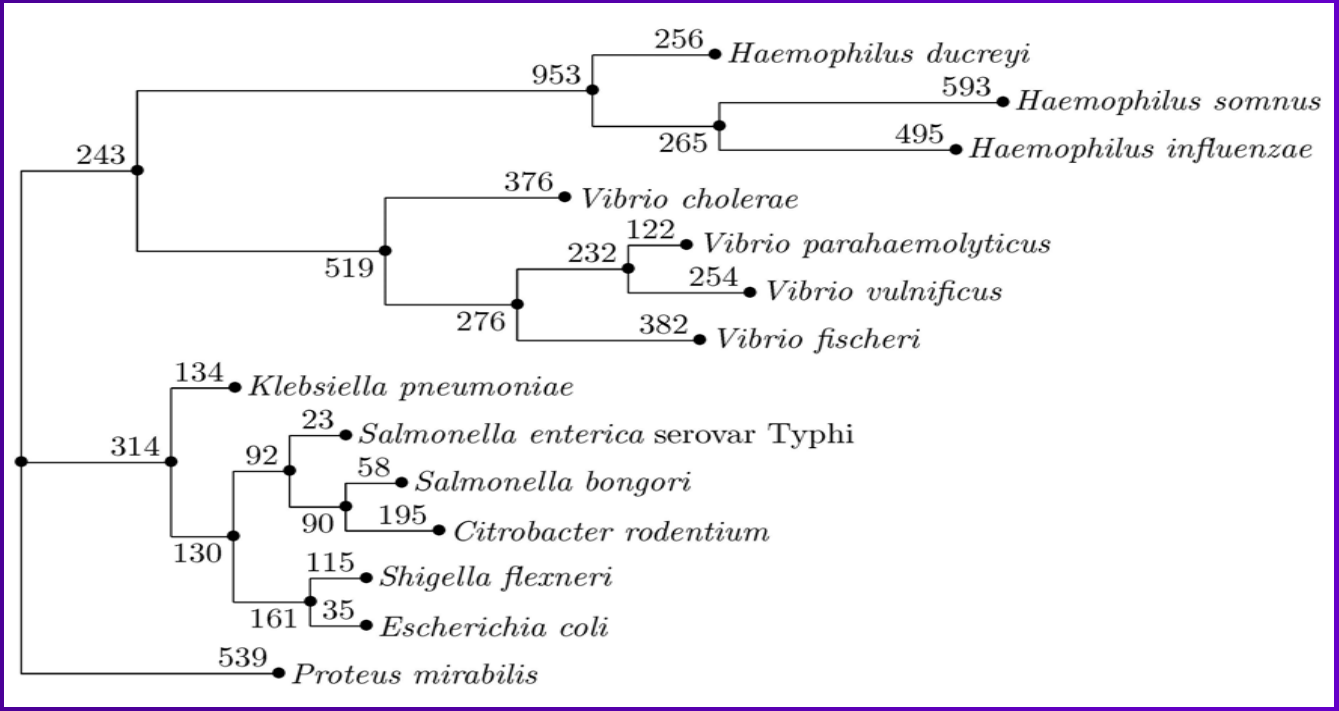
- OrthoGibbs –for *de novo* discovery
- PhyloScan – for known-pattern search

… to ease reconstruction of gene regulatory circuits

… to further our understanding of cell function, differentiation, and pathology

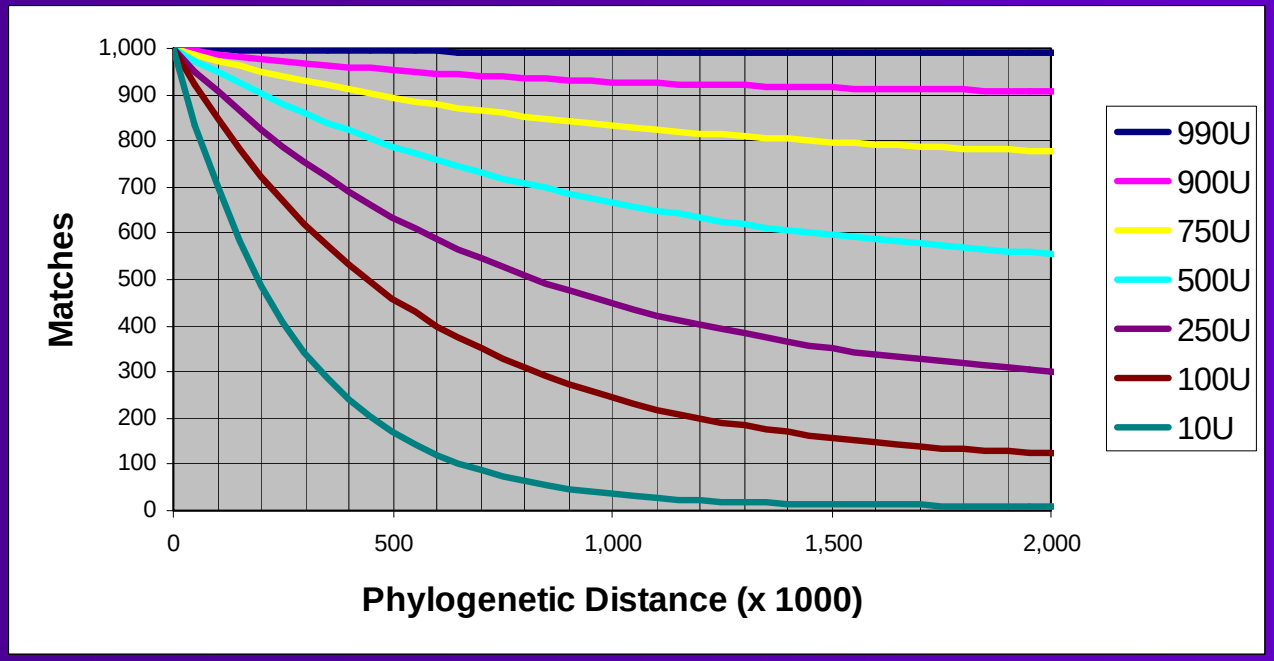
… for the betterment of human health

Our Approach: Phylogenetic Tree



This tree is approximate; built from a "stretched" 16S rRNA gene tree.
Edge length = number of mutations expected per 1,000 nucleotides of junk.
Observe clades: *Pasteurellales*, *Vibrionales*, *Enterobacteriales*, *P. mirabilis*

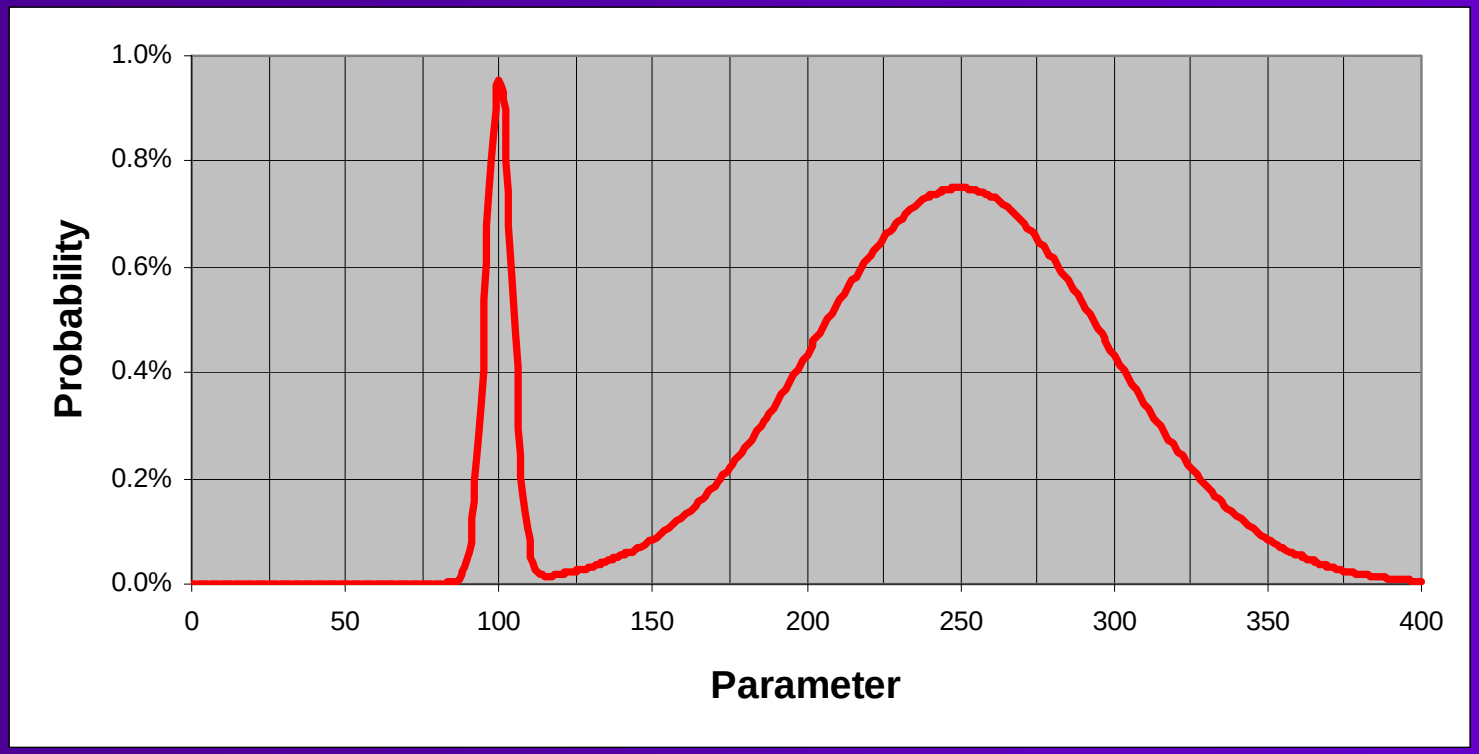
Our Approach: Selection Pressures Model



Model of Halpern & Bruno, *Mol Biol Evol,* 1998 – reverse engineers fitnesses (and hence fixation probabilities) from equilibrium

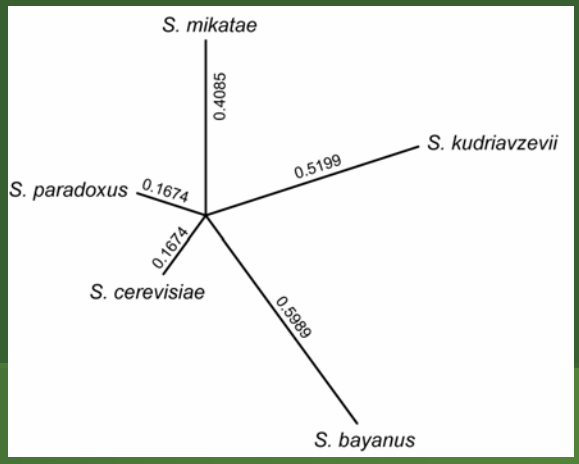
For various selection equilibria, the probability of a nucleotide remaining fixed, as a function of phylogenetic distance

Our Approach: Centroid Solution



Peak at 100 is the Maximum Likelihood Estimator (MLE) or Maximum A posteriori Probability (MAP). Peak at 250 is the Centroid Estimator

OrthoGibbs



Example (fictitious) tree tested by Siddharthan, Siggia & van Nimwegen, *PLoS Comput Biol*, 2005

Discovering transcription factor binding sites *de novo*

- Multisequence alignments helpful, but not required
- Phylogenetic tree required to relate multiply aligned sequences within each clade

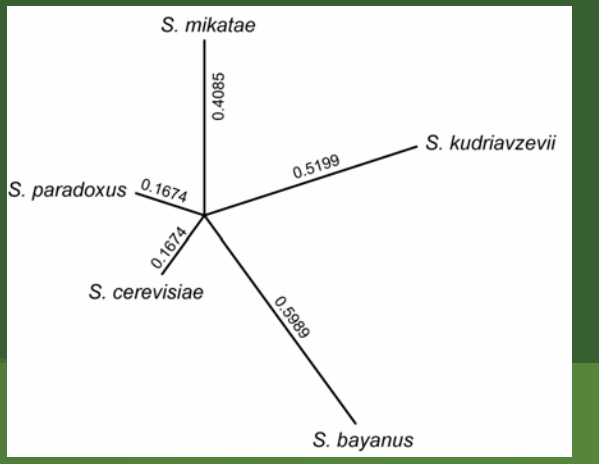
```
...1234567890... … ...1234567890...
GGCCGGTGCTATTACG … GCACGGAGTTATGCGA  S. cerevisiae
GGTCGGTGCTATCACG … TCGCGGAGGTATAGGA  S. paradoxus
GGCCTGTGTTATTTCG … GCGCGGTGTTATACGA  S. mikatae
AACCGGTGTTATTACA … GCGCGGAGTTATAAAG  S. kudriavzevii
AGACGGTGTTATGGCA … ACGCGGAGGTATGCGG  S. bayanus
```

- Multiple instances within a sequence is helpful

OrthoGibbs: Markov Chain Monte Carlo

- Start with a randomly guessed solution
- Repeat for many iterations:
  - Throw away part of the solution
  - Extend remainder to full solution
- From record of each iteration's solution, take census, compute centroid

OrthoGibbs: Each iteration



- Updates locations of proposed binding sites – via Gibbs Sampling
- Updates motifs that describe the binding sites – via approximate sampling, corrected with Metropolis-Hastings test

Achieves "detailed balance" – solutions are visited with probability proportional to their modeled likelihood

```
...1234567890... … ...1234567890...
GGCCGGTGCTATTACG … GCACGGAGTTATGCGA  S. cerevisiae
GGTCGGTGCTATCACG … TCGCGGAGGTATAGGA  S. paradoxus
GGCCTGTGTTATTTCG … GCGCGGTGTTATACGA  S. mikatae
AACCGGTGTTATTACA … GCGCGGAGTTATAAAG  S. kudriavzevii
AGACGGTGTTATGGCA … ACGCGGAGGTATGCGG  S. bayanus
```

STB5p binding site

	1	2	3	4	5	6	7	8	9	10
A	4.9%	4.8%	6.7%	34.6%	4.9%	12.5%	1.9%	91.3%	17.3%	55.8%
C	87.4%	6.7%	1.0%	8.7%	25.2%	6.7%	1.9%	1.0%	8.7%	6.7%
G	1.0%	86.5%	90.4%	9.6%	66.0%	2.9%	3.9%	1.0%	4.8%	17.3%
T	6.8%	1.9%	1.9%	47.1%	3.9%	77.9%	92.2%	6.8%	69.2%	20.2%

OrthoGibbs: Centroid

```
...1234567890... … ...1234567890...
GGCCGGTGCTATTACG … GCACGGAGTTATGCGA S. cerevisiae
GGTCGGTGCTATCACG … TCGCGGAGGTATAGGA S. paradoxus
GGCCTGTGTTATTTCG … GCGCGGTGTTATACGA S. mikatae
AACCGGTGTTATTACA … GCGCGGAGTTATAAAG S. kudriavzevii
AGACGGTGTTATGGCA … ACGCGGAGGTATGCGG S. bayanus
```

Intuitive idea …

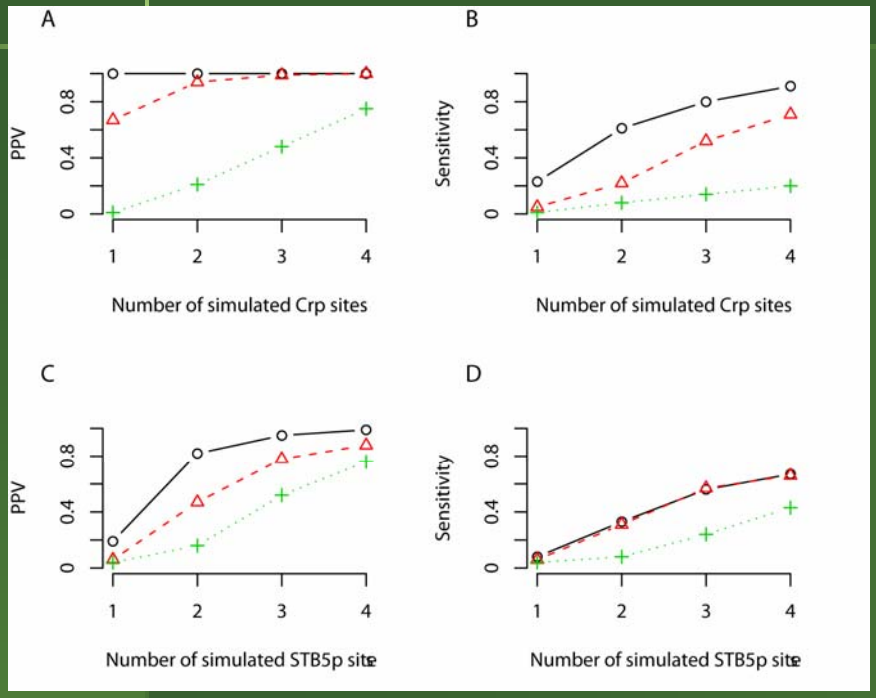
- Report all those binding sites that appear in at least half the algorithm iterations

Details & Subtleties:

- Sites need not appear together in the same iteration
- Sites need not correspond to identical motif models
- Burn-in iterations
- Sufficient overlaps
- Exclusivity of nearby sites

OrthoGibbs: Simulations & Results

Improved Sensitivity & Positive Predictive Value!



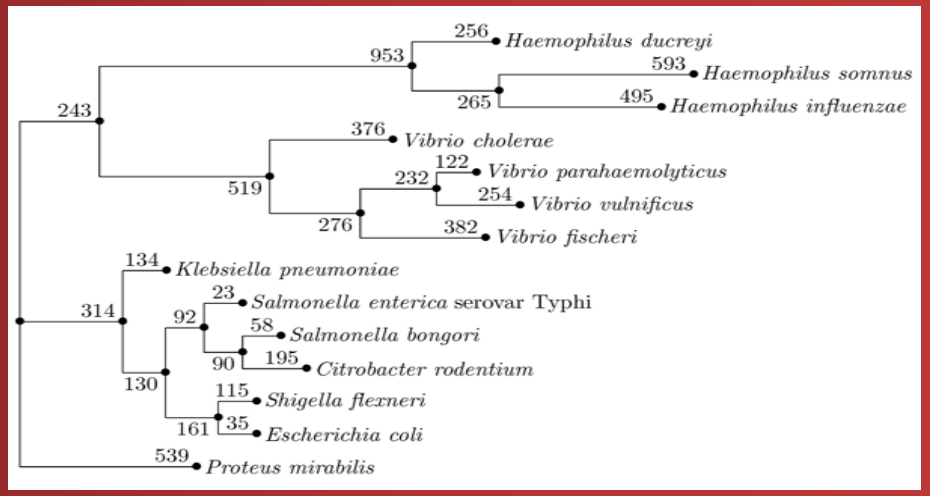
TFBS # Sites	Algorithm[a]	Total Predictions	True Positives	False Positives	False Negatives	PPV[b]	Sensitivity[b]
Crp							
1	MAP+Phylogeny	8.0	5.3	2.7	92.0	0.67	0.05
1	Centroid+Phylogeny	22.7	22.7	0.0	77.3	1.00	0.23
1	PhyloGibbs	53.3	0.7	52.7	99.3	0.01	0.01
2	MAP+Phylogeny	46.0	43.3	2.7	156.7	0.94	0.22
2	Centroid+Phylogeny	122.3	122.0	0.3	77.7	1.00	0.61
2	PhyloGibbs	79.7	16.7	63.0	183.3	0.21	0.08
3	MAP+Phylogeny	157.0	155.7	1.3	144.3	0.99	0.52
3	Centroid+Phylogeny	241.3	241.3	0.0	58.7	1.00	0.80
3	PhyloGibbs	86.7	41.7	45.0	258.3	0.48	0.14
4	MAP+Phylogeny	284.7	284.3	0.3	115.7	1.00	0.71
4	Centroid+Phylogeny	364.0	364.0	0.0	36.0	1.00	0.91
4	PhyloGibbs	109.3	81.7	27.7	318.3	0.75	0.20
STB5p							
1	MAP+Phylogeny	95.7	6.0	89.7	94.0	0.06	0.06
1	Centroid+Phylogeny	42.0	8.0	34.0	92.0	0.19	0.08
1	PhyloGibbs	106.7	4.3	102.3	95.7	0.04	0.04
2	MAP+Phylogeny	131.7	62.3	69.3	137.7	0.47	0.31
2	Centroid+Phylogeny	79.7	65.3	14.3	134.7	0.82	0.33
2	PhyloGibbs	96.0	15.0	81.0	185.0	0.16	0.08
3	MAP+Phylogeny	218.3	171.3	47.0	128.7	0.78	0.57
3	Centroid+Phylogeny	177.7	169.3	8.3	130.7	0.95	0.56
3	PhyloGibbs	137.7	71.0	66.7	229.0	0.52	0.24
4	MAP+Phylogeny	302.7	265.7	37.0	134.3	0.88	0.66
4	Centroid+Phylogeny	273.0	269.3	3.7	130.7	0.99	0.67
4	PhyloGibbs	228.0	172.3	55.7	227.7	0.76	0.43

Motif Models	Possible Sites[b]	Total Predictions	True Positives	False Positives	False Negatives	PPV	Sensitivity	Mean Distance[c]
1	103	57.7	47.3	10.3	55.7	0.82	0.46	0.20
2	128	74.3	57.3	17.0	53.7	0.77	0.45	0.25
3	132	79.3	61.3	18.0	70.7	0.77	0.46	0.29

Newberg, Thompson, Conlan, Smith, McCue, & Lawrence, *Bioinformatics*, 2007

PhyloScan

E. coli Crp motif from Carmack, McCue, Newberg, & Lawrence, *Algorithms Mol Biol*, 2007



For finding (additional) binding sites that match a known pattern

- Multisequence alignments helpful
- Phylogenetic tree required for each clade
- Multiple instances within a sequence is helpful

- Prior work: MONKEY by Moses, Chiang, Pollard, Iyer, & Eisen, *Genome Biol,* 2004

PhyloScan: Deterministic Scanning

Asking how promising is …

1. a binding site possibility in an multisequence-alignment intergenic region
2. the best site in such an intergenic region given the length of the region. Likewise for 2nd best, …
3. an intergenic region, as indicated by its best sites. If good enough…
4. an intergenic region, with additional evidence from the orthologous regions in other clades
5. a set of orthologous intergenic regions, given the number of sets examined

The nitty gritty:

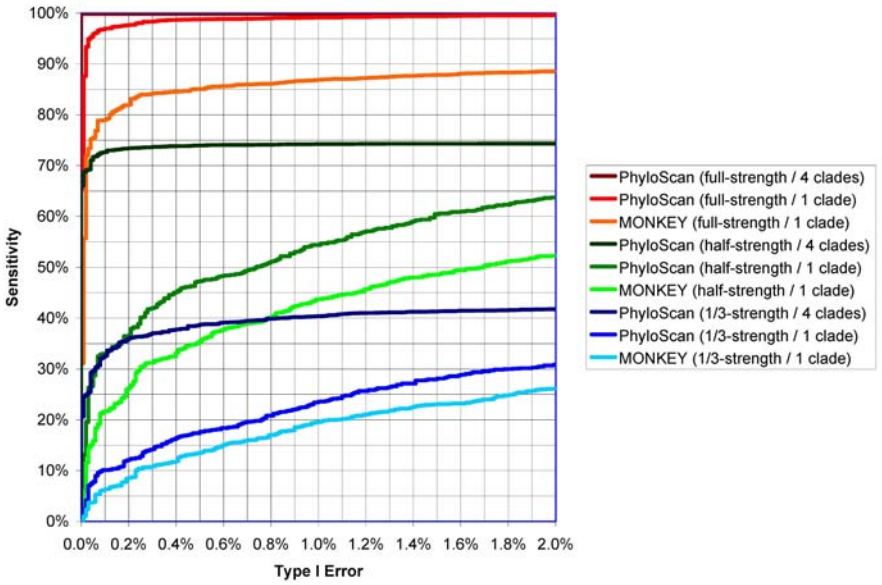
1. Compute exact p-value. Phylogenetic adaptation of Staden, *Comput Appl Biosci*, 1998
2. Compute p-value given a site's order statistic
3. Combine computed p-values of best sites in an intergenic region – Neuwald & Green, *J Mol Biol*, 1994
   If good enough…
4. Combine across clades – Bailey & Gribskov, *J Comput Biol*, 1998
5. Convert to q-values. Storey, *J Royal Stat Soc B*, 2002

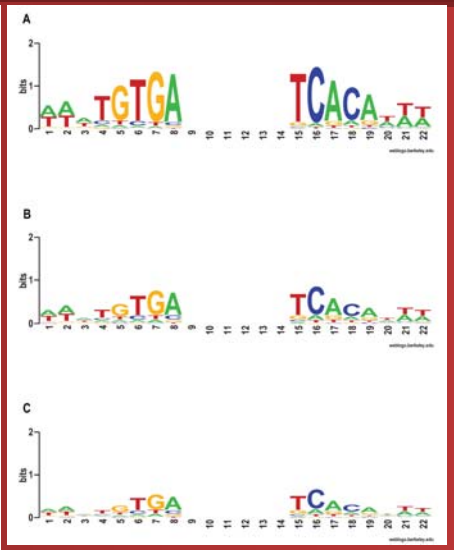
PhyloScan: Tuning Parameters

- Specify weights for combining best sites in an intergenic region — $1^{st} = 0.9$, $2^{nd} = 0.1$
- Designate threshold to prevent "rescue" of dubious intergenic regions — $p \leq 0.05$
- Designate report threshold — $q \leq 0.001$

PhyloScan: Simulations & Results

Improved Sensitivity & Specificity!





Crp motifs:
- full-strength
- 1/2-strength
- 1/3-strength

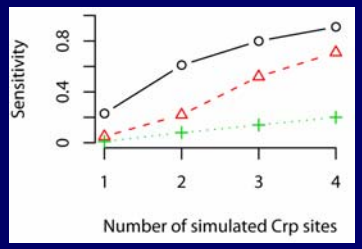
A pair of planted sites

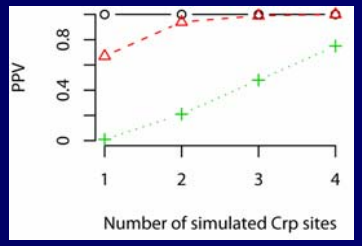
Carmack, McCue, Newberg, & Lawrence, *Algorithms Mol Biol*, 2007

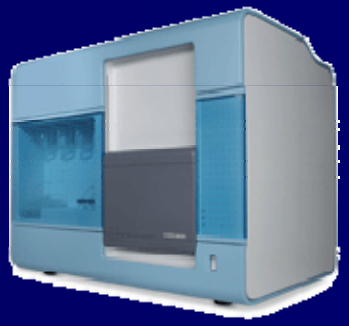
	C1	C2	C3	C4	C5	C6
E. coli Sequence Data	Full[a]	Full[a]	Red.[b]	Red.[b]	Red. & Aligned[c]	Red. & Aligned[c]
Indep. Species	No	Yes	No	Yes	No	Yes
Crp Known[d]	1(2)	7(10)	1(2)	8(12)	4(6)	11(16)
Crp Novel[d]	0(0)	16(20)	0(0)	16(18)	6(7)	18(21)
PurR Known[d]	1(1)	9(9)	1(1)	11(11)	9(9)	12(12)
PurR Novel[d]	0(0)	4(5)	0(0)	4(5)	3(4)	6(7)

PurR motif

Conclusions







- We reduced omissions and false discoveries as a fraction of actual sites
  - OrthoGibbs for *de novo* detection Newberg, Thompson, Conlan, Smith, McCue, Lawrence, *Bioinformatics*, 2007
  - PhyloScan for additional sites Carmack, McCue, Newberg, & Lawrence, *Algorithms Mol Biol*, 2007