Gibbs Sampling for Gene *Cis*-Regulatory Elements

Lee A. Newberg @GE 10/21/2011

Acknowledgments

Team:

- Sean P. Conlan (NIH)
- Travis J. Desell (UND)
- Charles E. Lawrence (Brown)
- Lee Ann McCue (Pacific Northwest NL)
- Thomas M. Smith (MIT Lincoln Laboratory)
- William A. Thompson (Brown)

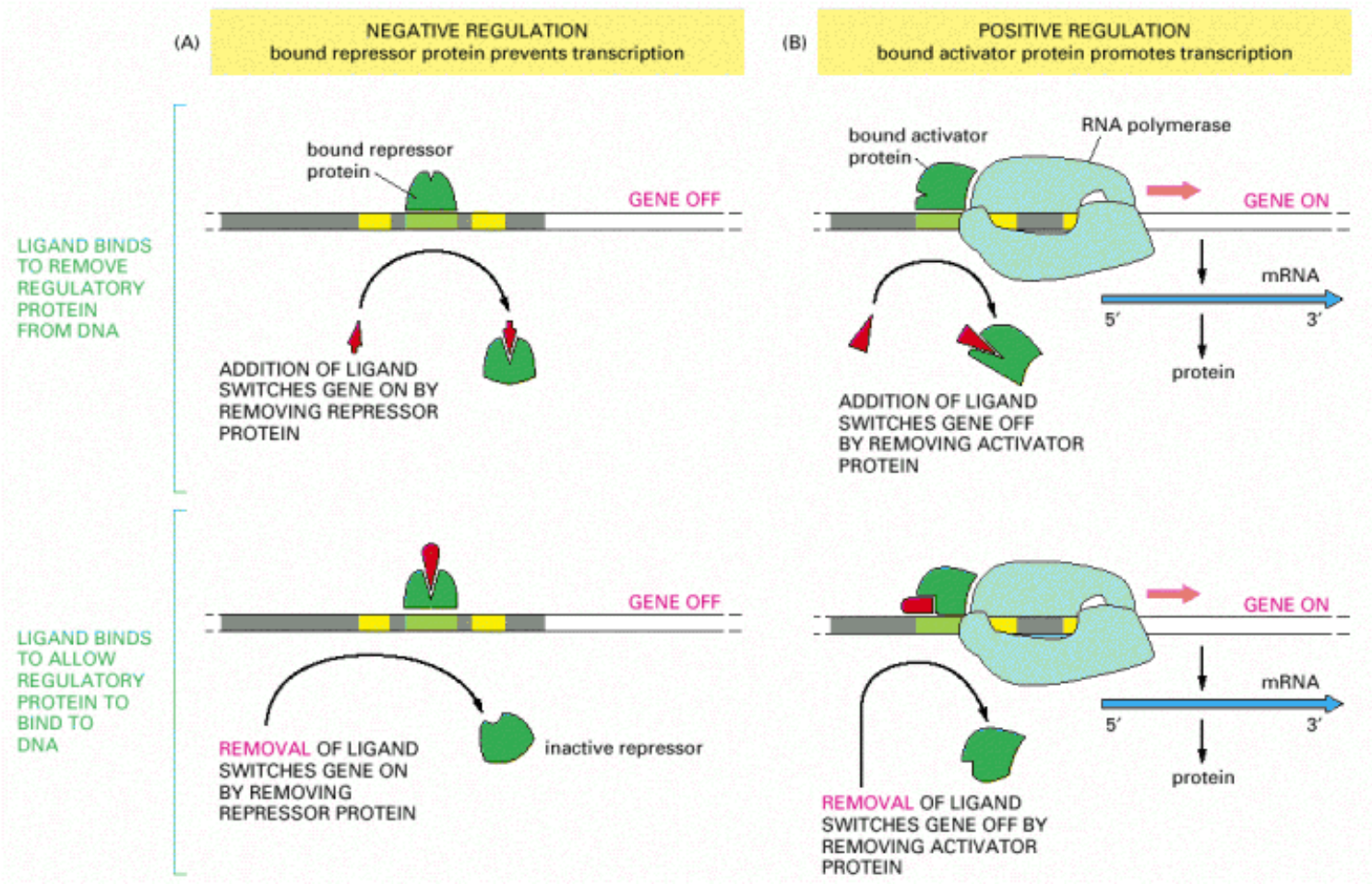
Resources:

- Wadsworth Center (including LCSB Core Facility)
- Rensselaer Polytechnic Institute
- Brown University (including CCMB)
- NIH/NHGRI: K25 Mentored Career Award "Quantitative Cross-Species Approaches to Gene Regulation" (LAN)
- DOE: "Bayesian computational approaches for gene regulation studies of bioethanol and biohydrogen production" (CEL, LAM, LAN)

Gibbs sampling for gene regulation

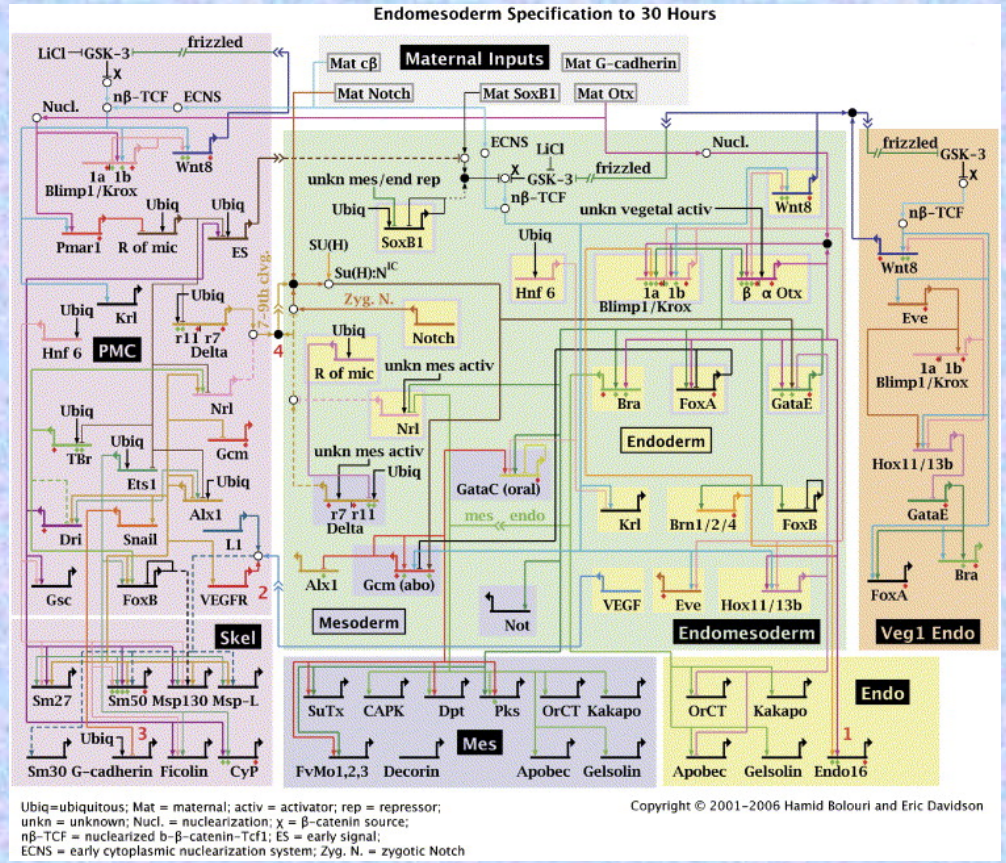
- DNA *cis*-regulatory elements
- Importance
- Computational Prediction: Inputs, Outputs
- Results
- Algorithm
  - Statistical model, Gibbs sampling, & centroids

DNA *cis*-regulatory elements



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

Importance of DNA *cis*-regulatory elements



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

Important

… 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

Typically vary species to species, but not individual to individual, except pathologically.

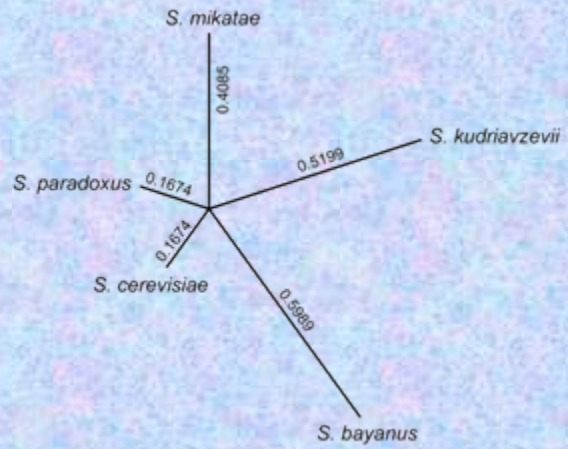
Outputs

- Element sites
- Motif (pattern) description

STB5p

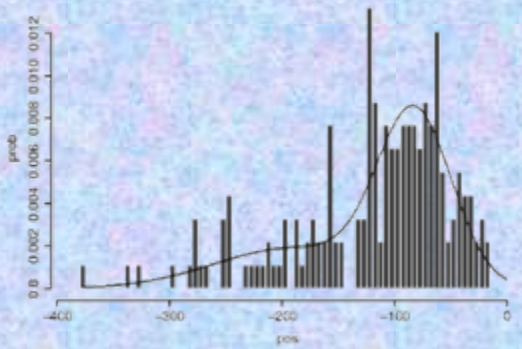
Input data

- Promoter sequences: aligned when feasible (*e.g.*, 20 sequences × 5 species)
- Phylogenetic tree and model: or *ad hoc* substitute

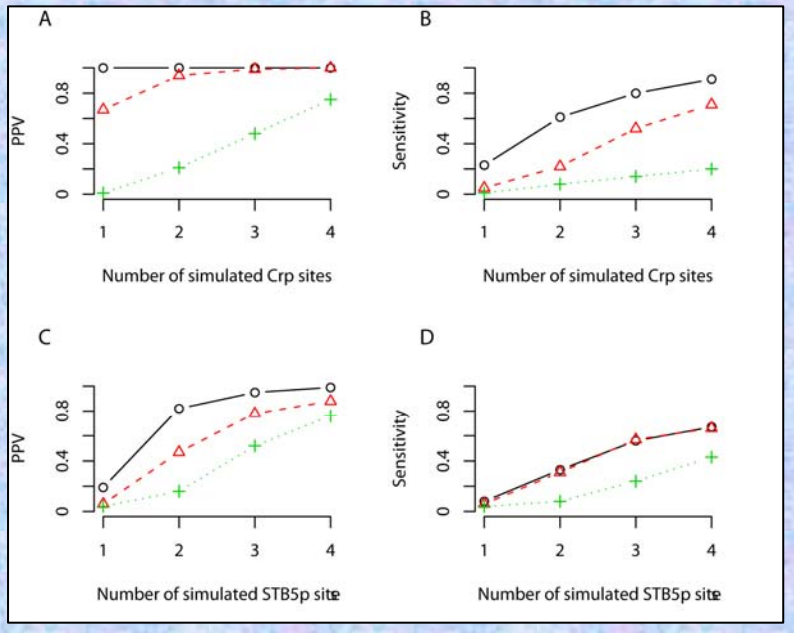


Input parameters

- Motif model type: consensus w/ deviations vs. probabilistic
- Motif size: fixed, varying. (6-36 nts.)
- Motif shapes: palindromic, off positions
- Site frequency: per promoter, genome
- Site positioning: nucleosomes, relative to +1



Effectiveness in simulations



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

Verifying known results

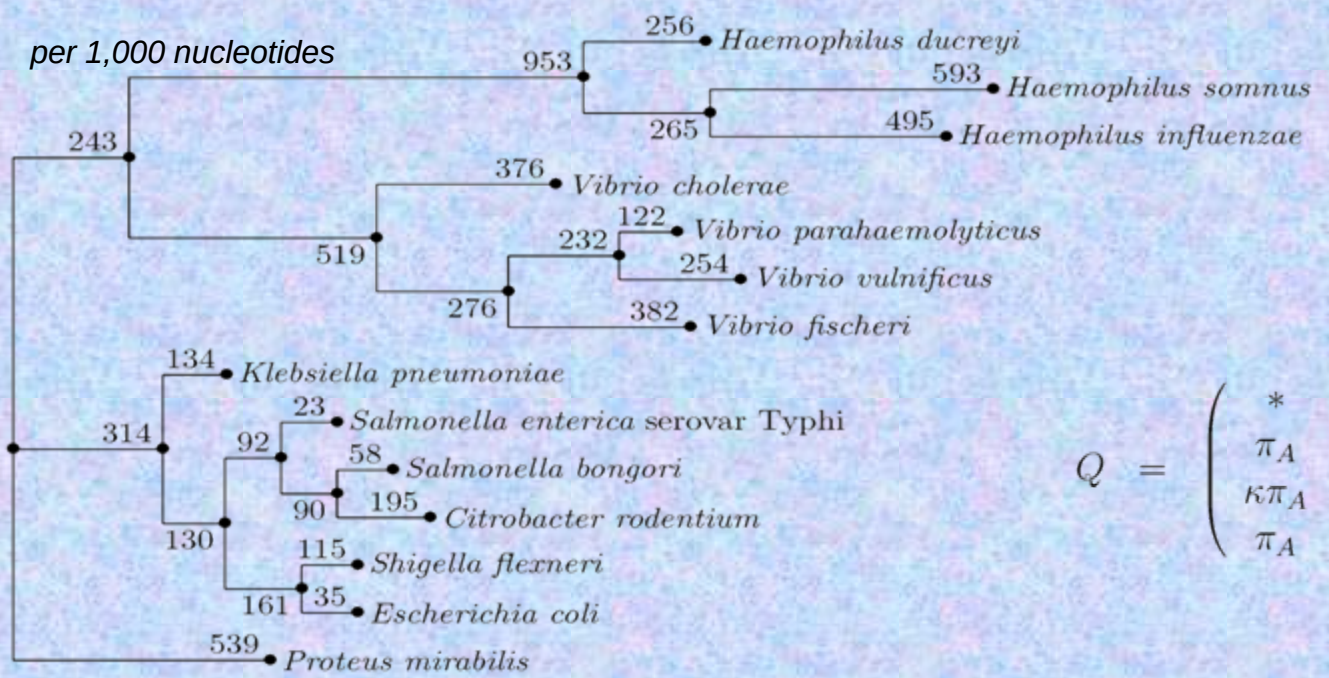
Motif Models	Possible Sites[b]	Total Predictions	True Positives	False Positives	False Negatives	PPV	Sensitivity	Mean Distance[c]
1	103	57.7	47.3	10.3	55.7	0.82	0.46	0.20
2	128	74.3	57.3	17.0	53.7	0.77	0.45	0.25
3	132	79.3	61.3	18.0	70.7	0.77	0.46	0.29

Novel predictions (by others)

E.g.: Driscoll *et al.* (2007) Carbon utilization pathway in Shewanella

What we do, part 1 of 3: Statistical Modeling

- Plausible, workable, statistical model *components*:
  - Halpern & Bruno (1998), *Mol Biol Evol*, nucleotide evolution
  - Position weight matrices for motif models, *etc.*



$$Q = \begin{pmatrix} * & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & * \end{pmatrix}$$

What we do, part 2 of 3: Gibbs sampling (MCMC)

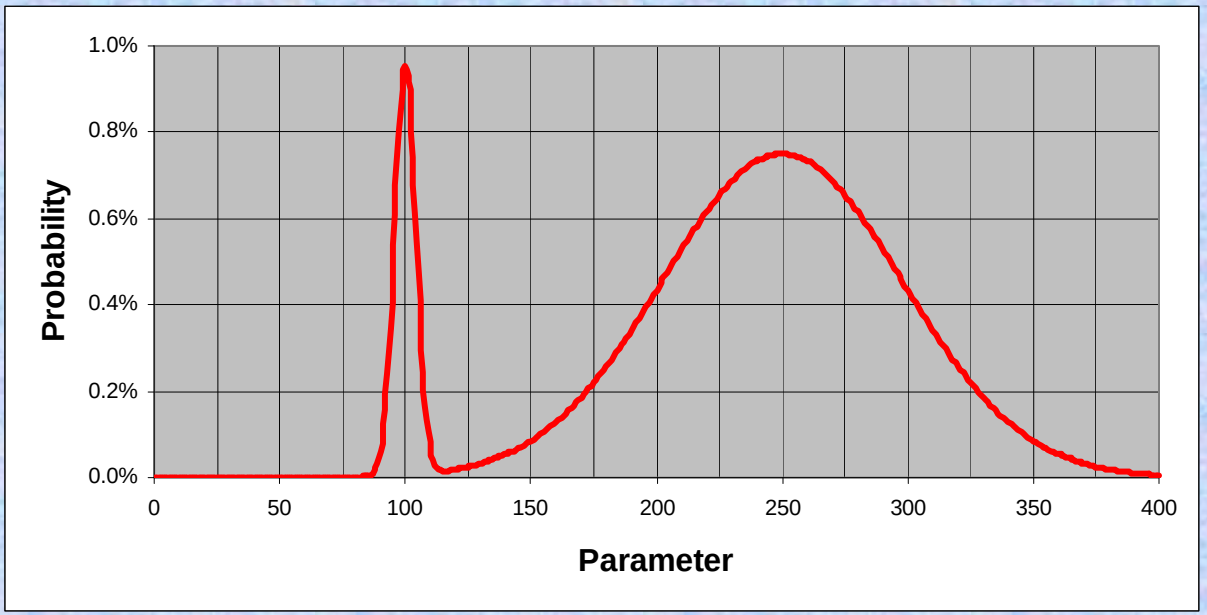
Random walk through posterior probability space.

Possible because: we resample part of a solution, conditioned on the rest.

- Element sites
- Motif model logos
- Tree sequence alignment, *etc.*

Record key features as we walk

What we do, part 3 of 3: Centroid solution



We focus on the region of solution space containing the most posterior probability, rather than on the single solution that is most probable.

→ Build centroid from marginal probability of relevant features

C/C++ UNIX.

Single CPU / Small Cluster / DNA@home.

Payoff

Plausible statistical model components
+ Gibbs sampling
+ Centroid

→ Robust predictions

- Theory: Newberg *et al.* (2007), *Bioinformatics*, Pubmed17488758
- Use: Thompson *et al.* (2007), *Nucleic Acids Research*, Pubmed17483517