A Phylogenetic Gibbs Recursive Sampler for Locating Transcription Factor Binding Sites

Sean P. Conlan[1] Lee Ann McCue[2] Lee A. Newberg[1,3]
Thomas M. Smith[3] William Thompson[4] Charles E. Lawrence[4]

[1]Wadsworth Center, New York State Department of Health

[2]Pacific Northwest National Laboratory

[3]Department of Computer Science, Rensselaer Polytechnic Institute

[4]Department of Applied Mathematics, Brown University

International Conference in Phylogenomics 2006

Conclusions

Take-Home Points

- Phylogenetic modeling (with full Felsenstein's Algorithm) helps
- Use of ensemble centroids helps
- New nucleotide selection model is intriguing

Goal: Finding *Cis*-Regulatory Elements *De Novo*

What We're Looking For

- Seeking elements that are short: 6–30 bp
- Only partial conserved
- Isolated elements or multiple elements per module
- Single or multiple intergenic regions per genome
- Alignable and unalignable sequence data across genomes

Measures of Success

- Sensitivity — minimize false negatives
- Selectivity — minimize false positives

Previous Work

Non-Phylogenetic Algorithms

Many good algorithms including

- Gibbs Recursive Sampler (Thompson *et al.*, 2003)

But need to be better when analyzing closely related species.

Phylogenetic Algorithms

Several good algorithms

- Non-statistical and/or two-species only
- PhyloGibbs (Siddharthan *et al.*, 2005). Uses successive star-toplogy approximations, maximum likelihood

But improvement is possible with full Felsenstein's Algorithm and with ensemble centroids

Gibbs Sampling

Gibbs Sampling Overview

Move from proposed solution to proposed solution via Gibbs Sampling.

- From any proposed set of sites
  - Re-choose sites in one multi-sequence[a], with probability conditioned on sites in remaining multi-sequences
- Iterate to explore parameter space.
  - Explores each proposed set of sites with probability proportional to its likelihood.

[a] An unalignable sequence or a set of aligned sequences

Probability Conditioned on Remaining Sites

An slight oversimplification . . .

Probability Calculation

- Current Iteration has a *position-weight matrix*, which gives current motif descriptions, and is built from counts & pseudocounts
- A position's weights parameterize a Dirichlet distribution, which is used to draw an equilibrium distribution
- The equilibrium is used to parameterize a nucleotide substitution model (*e.g.*, HKY85).
- The substitution model is used to evaluate all positions posited to belong

Ensemble Centroid

Computing the Ensemble Centroid

With each sample from the Gibbs Sampler (after "burn in" iterations)

- For each sequence position record a "1" if it is part of a *cis*-element, record "0" otherwise.
- The vector of 0's and 1's is the corner of a hypercube

Ensemble centroid = corner nearest to the center of mass of the collected samples

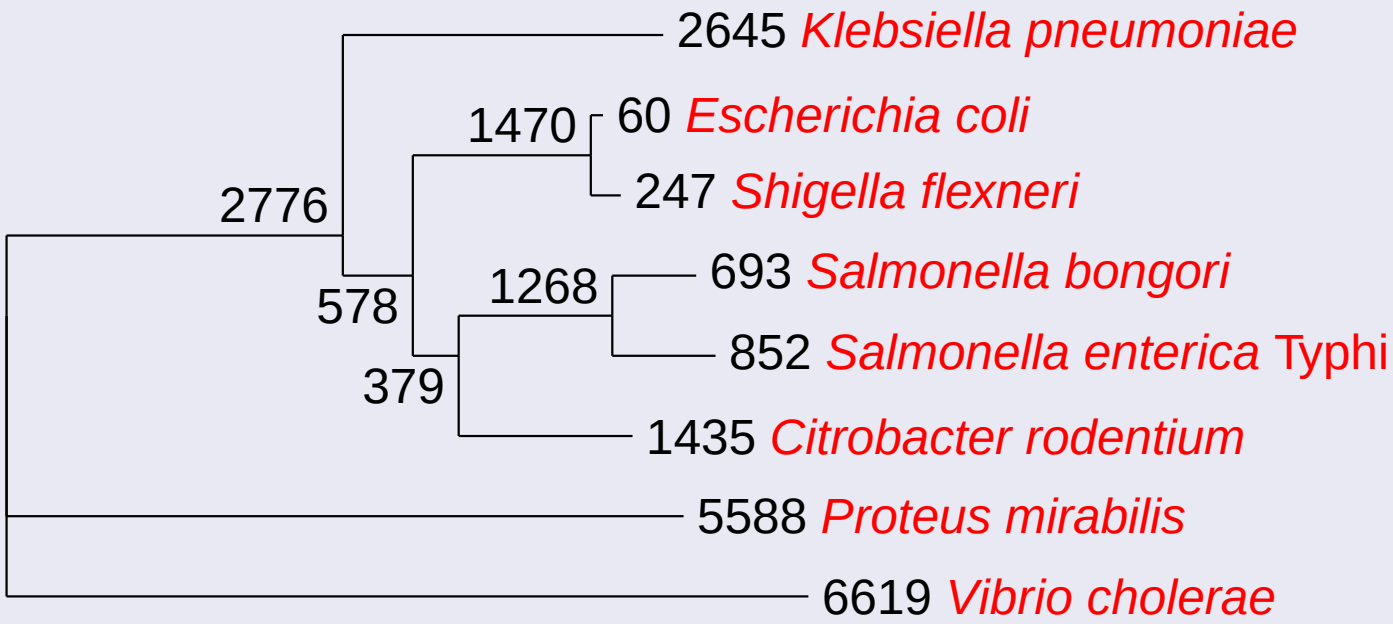
Advantages

Advantages of Ensemble Centroid

- Expensive *a posteriori* probability calculation not needed
  - Star-topology approximation unnecessary
- Gives "entropic" solutions their due

Synthesizing the Data

A data set: Eight species' 500 bp sequences



Expected number of substitutions $\times 10^4$

- Gapless sequence data generated according to tree
- *P. mirabilis* and *V. cholerae* subsequently treated as not alignable

Synthesizing the Data

Five Collections of Data Sets

- Five collections of data sets: $k \in \{0, 1, 2, 3, 4\}$
- 100 data sets in each collection
- A data set is 8 sequences
  - one for each species
  - each of length 500 bp
  - each with k planted *Escherichia coli* Crp binding sites
  - related by phylogenetic tree

Data Analysis

- Each data set run separately — 500 runs total
- Accumulate results across data sets in each collection.

Sensitivity & Selectivity

E.g., second entry shows 116 *E. coli* sites found across 61 data sets.

Our Algorithm (top) and PhyloGibbs (bottom)

Data Collection	#0	#1	#2	#3	#4
Sites Found		17/17	116/61	154/82	176/93
(True Positives)		0/0	13/8	54/26	75/35
False Sites	3/3	5/4	2/2	0/0	0/0
(False Positives)	47/46	60/51	63/44	40/30	30/24
Sites Missed		83/83	84/45	146/100	224/100
(False Negatives)		100/100	187/95	246/89	325/97

"BRASS" implementation of our algorithm (Smith, 2006), configured to find up to two sites per multi-sequence

Modeling How Nucleotides Evolve

Existing Models

- Arbitrary equilibria
- Transition/transversion rate ratio
- Mutation rate variation within a genome
- Selection effects via scaled fixation rates (Halpern & Bruno, 1998)
- Context sensitive: Di- and tri-nucleotide models
- Indel support, though difficult with Felsenstein's Algorithm

A New Model for Selection Effects

Newberg (2005) allows that SNPs are not improbable. (*I.e.*, without the specious fixation on species fixation.)

Traditional Nucleotide Substitution Model

Traditional Mutation (without Selection)

$$M_x = \begin{pmatrix} \Pr[A|A] & \Pr[C|A] & \Pr[G|A] & \Pr[T|A] \\ \Pr[A|C] & \Pr[C|C] & \Pr[G|C] & \Pr[T|C] \\ \Pr[A|G] & \Pr[C|G] & \Pr[G|G] & \Pr[T|G] \\ \Pr[A|T] & \Pr[C|T] & \Pr[G|T] & \Pr[T|T] \end{pmatrix}$$

$$= \begin{pmatrix} 0.96 & 0.01 & 0.02 & 0.01 \\ 0.01 & 0.96 & 0.01 & 0.02 \\ 0.02 & 0.01 & 0.96 & 0.01 \\ 0.01 & 0.02 & 0.01 & 0.96 \end{pmatrix}$$

Each row sums to 1.0.

A New Nucleotide Substitution Model

Population Model for Selection (without Mutation)

$$M_x = \begin{pmatrix} \Pr[A|A] & \Pr[C|A] & \Pr[G|A] & \Pr[T|A] \\ \Pr[A|C] & \Pr[C|C] & \Pr[G|C] & \Pr[T|C] \\ \Pr[A|G] & \Pr[C|G] & \Pr[G|G] & \Pr[T|G] \\ \Pr[A|T] & \Pr[C|T] & \Pr[G|T] & \Pr[T|T] \end{pmatrix}$$

$$= \begin{pmatrix} 1.1 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 1.0 \end{pmatrix}$$

Each row no longer sums to 1.0 but . . .

A New Nucleotide Substitution Model

Starting with 100 organisms of each type . . .

Population Model for Selection (without Mutation)

$$M_x = \begin{pmatrix} 1.1 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 1.0 \end{pmatrix}$$

$$(100, 100, 100, 100)M_x = (110, 100, 100, 100)$$

$$\frac{(110, 100, 100, 100)}{410} = (0.268, 0.244, 0.244, 0.244)$$

A New Nucleotide Substitution Model

Combining the two . . .

Mutation *and* Selection

$$M_x = \begin{pmatrix} \Pr[A|A] & \Pr[C|A] & \Pr[G|A] & \Pr[T|A] \\ \Pr[A|C] & \Pr[C|C] & \Pr[G|C] & \Pr[T|C] \\ \Pr[A|G] & \Pr[C|G] & \Pr[G|G] & \Pr[T|G] \\ \Pr[A|T] & \Pr[C|T] & \Pr[G|T] & \Pr[T|T] \end{pmatrix}$$

$$= \begin{pmatrix} 1.056 & 0.01 & 0.02 & 0.01 \\ 0.011 & 0.96 & 0.01 & 0.02 \\ 0.022 & 0.01 & 0.96 & 0.01 \\ 0.011 & 0.02 & 0.01 & 0.96 \end{pmatrix}$$

A New Nucleotide Substitution Model

Some Details

- Each generation: mutation at DNA replication; selection between replications.
- Instantaneous rate formalism $M_x = \exp(xR)$ still applies, so generation length need not be known.
- 2x invocations of Felsenstein's Algorithm, because each row no longer sums to 1.0.
- Easily computed, one-to-one correspondence between nucleotide equilibria $\vec{\theta}$ and diagonal selection matrix[a]

[a] assuming, *e.g.*, asymptotic population stability

Conclusions

Take-Home Points

- Phylogenetic modeling (with full Felsenstein's Algorithm) helps
- Use of ensemble centroids helps
- New nucleotide selection model is intriguing

Contact Information

lnewberg©wadsworth·org www.rpi.edu/~newbel/

→ Felsenstein, J. (1981) PubMed 7288891.

→ Halpern, A. L. & Bruno, W. J. (1998) PubMed 9656490.

→ Hasegawa, M., Kishino, H. & Yano, T. (1985) PubMed 3934395.

→ Newberg, L. A. (2005).
http://www.cs.rpi.edu/research/pdf/05-08.pdf.

→
Siddharthan, R., Siggia, E. D. & van Nimwegen, E. (2005) PubMed 16477324.

→ Smith, T. M. (2006). PhD thesis, Rensselaer Polytechnic Institute
Troy, NY. In preparation.

→
Thompson, W., Rouchka, E. C. & Lawrence, C. E. (2003) PubMed 12824370.