The Phylogenetic Gibbs Centroid Sampler for *Cis*-Regulatory Element Discovery

Also see:
Newberg *et al.*, *Bioinformatics* 2007
Thompson *et al.*, *Nuc Acids Res* 2007
PubMed 17,488,758 & 17,483,517

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

1. The Wadsworth Center, New York State Department of Health, Albany NY USA
2. Department of Computer Science, Rensselaer Polytechnic Institute, Troy NY USA
3. Center for Computational Molecular Biology, Brown University, Providence RI USA
4. Pacific Northwest Laboratory, Richland WA USA

Why Use a Centroid?

For many tasks, it is better to focus on the region of solution space with the most posterior probability than it is to focus on the single solution that is most probable.

MLE & MAP are flawed

Maximum Likelihood Estimates and Maximum A Posteriori estimates are often problematic. Consider: The most likely single configuration of air molecules in this room would have the heaviest molecules at the bottom, lighter molecules above them, and vacuum at the top. Imagine if you tried to determine thermodynamic properties of the air assuming only that configuration!

A centroid is "typical"

A centroid is a configuration that is most representative in the following sense. For a suitably defined distance function that measures the extent to which two configurations differ, a centroid *C* is a configuration that minimizes

$$\sum_{\text{all } S} \text{distance}(C, S)\ \Pr(S).$$

That is, a centroid is least different from the probability-weighted ensemble of possible configurations.

The distance function

We choose a distance function to measure the differences we care about. In our case, the distance between two proposed sets of *cis*-regulatory element locations is the number of element locations about which they disagree.

The Algorithm

We take a random walk through the space of possible configurations, remembering those we visit.

The inputs

We need the sequence data for genomic regions of interest. These are usually promoters or other non-genic regions.

Cross-species, multisequence alignments can be quite helpful (but are not required).

We need a phylogenetic tree to evolutionarily relate sequences that have been globally pre-aligned. We use the Halpern & Bruno (1998) nucleotide substitution model.

We employ user-supplied values (or use defaults) for element hints / prior information indicating:

- number of motif models (types)
- frequency (by promoter, genome)
- size (allowed widths)
- shape (*e.g.*, palindromic?)

Gibbs sampling (MCMC)

Starting with a guess (motif models, element locations, *etc.*) we iterate:

- discard some state variable values
- re-sample values with probability proportional to their likelihood given the data and retained state.
- after a burn-in period, record the state at the end of each iteration.

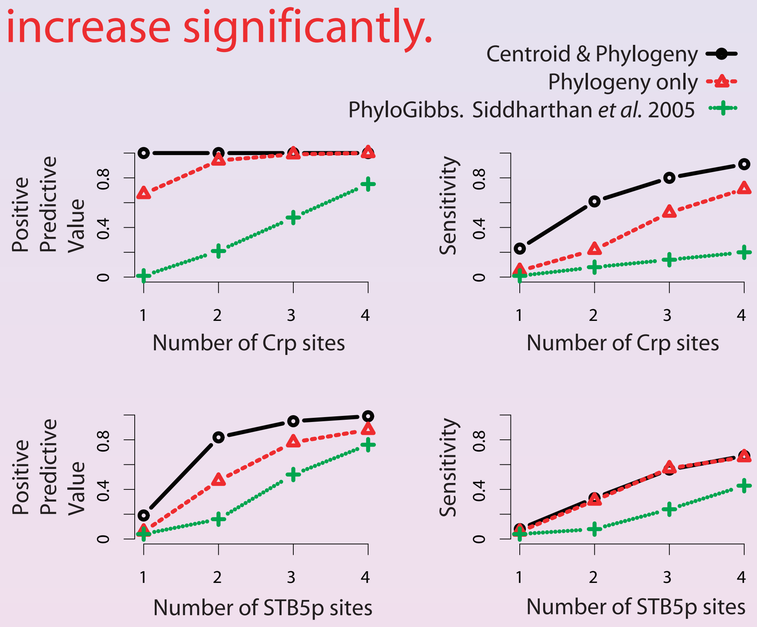
We compute a centroid from the visited states via a dynamic program.

The Results

Synthetic Data

Many simulations of orthologous data for a single gene with 1-4 planted sites. Either *E. coli* Crp sites (width 22, palindromic, 500bp regions, 8 species, 6 aligned) or *S. cerevisiae* STB5p sites (width 10, 1000bp regions, 5 aligned species).

Sensitivity (the fraction of planted sites that were discovered) and positive predictive value (the fraction of discoveries that were real) increase significantly.



Real Data

Each analyzed in isolation: 72 sets of orthologous promoters, across eight γ-proteobacterial species, each set having at least one experimentally validated binding site in *E. coli*.

sensitivity = 47.33/103 = 46.0%
positive predictive value = 81.6%

Not bad for looking at one promoter at a time. Numbers improve when looking simultaneously at co-regulated genes.

Together, centroids and the phylogenetic approach significantly reduce the number of false positives and false negatives.