Multi-Species, Multi-Gene Co-Regulation: Finding DNA *Cis*-Regulatory Elements for Biofuel Pathways in Alpha-Proteobacteria

Lee Ann McCue[1]
leeann.mccue÷pnnl÷gov
FWP 55426

Lee A. Newberg[2,3*]
leen÷cs÷rpi÷edu
DE-FG02-09ER64756

William A. Thompson[4]
william_thompson_1÷brown÷edu

Charles E. Lawrence[4]
charles_lawrence÷brown÷edu
DE-FG02-09ER64757

[1] Fundamental & Computational Sciences
Pacific Northwest National Laboratory
Richland, Washington

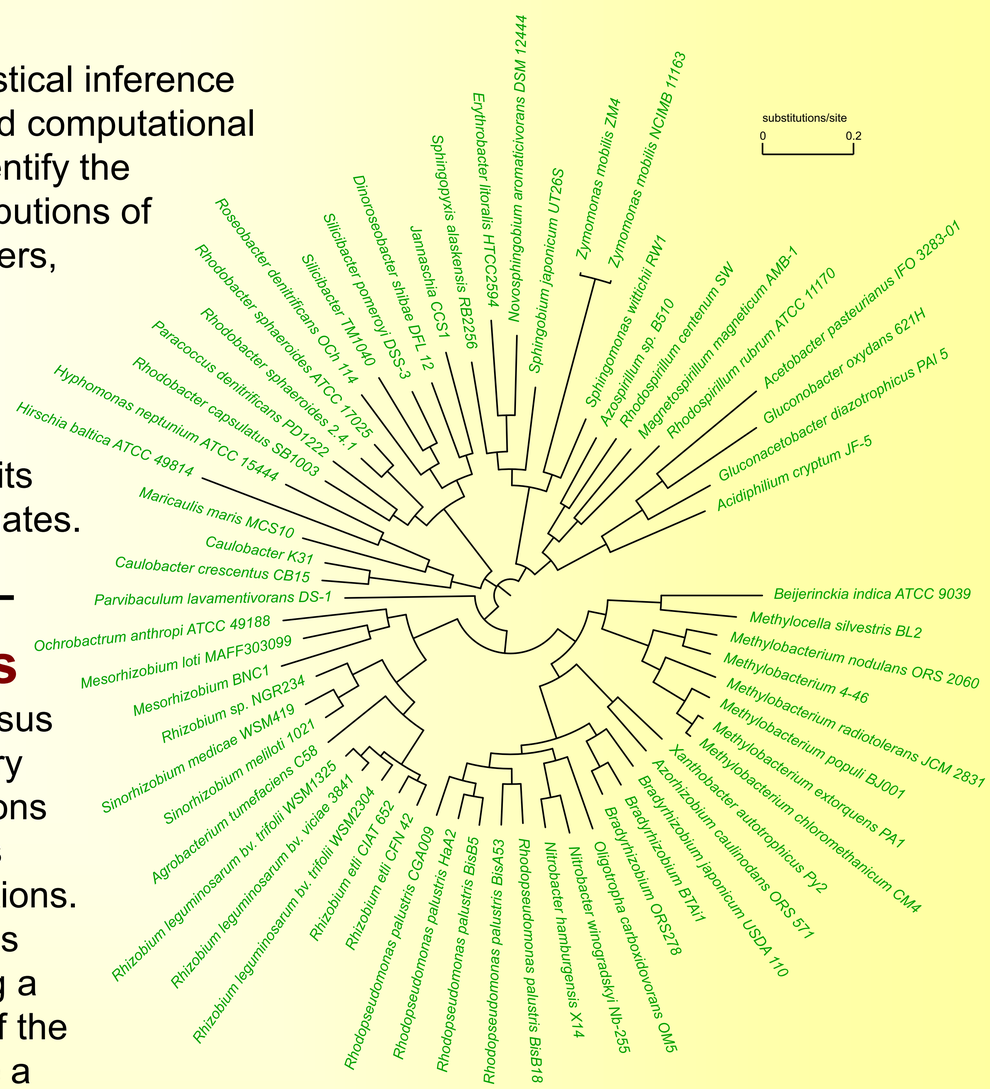
[2] Wadsworth Center
New York State Department of Health
Albany, New York

[3] Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

[4] Division of Applied Mathematics
Brown University
Providence, Rhode Island

Project goals

Decreasing America's dependence on foreign energy sources and reducing the emission of greenhouse gases are important national priorities. We are undertaking research into the *metabolic and regulatory networks responsible for biohydrogen and bioethanol production.* In particular, among the hundreds of alpha-proteobacterial species with sequenced genomes are several species with metabolic capabilities of interest. The first long term goal of this research is to *identify the ensemble of solutions that have been explored by the alpha-proteobacteria* to regulate the metabolic processes key to biofuel production. The second long term goal of this project is to *develop probabilistic models to represent these multiscale processes,* through Bayesian statistical inference procedures and computational methods to identify the posterior distributions of these parameters, efficient point estimates of their values, and Bayesian confidence limits for these estimates.



Modeling unknowns

Multiple sequence alignments often contain considerable uncertainty; rather than using only one alignment of multiple nucleotide sequences or leaving the sequences unaligned in searches for *cis*-regulatory elements, we are using a Bayesian approach that searches through the joint parameter space of *cis*-regulatory elements and multiple sequence alignments. This permits information about plausible multiple sequence alignments to aid in the locating of elements and also allows the plausible locations of elements to aid in the determination of the relative quality of multiple sequence alignments. *A Bayesian sampling of unknowns provides more realism than assigning values,* even maximum likelihood values.

Gibbs sampling

We employ Gibbs sampling, a Markov chain Monte Carlo technique, to explore the posterior probability space of solutions. At any given stage of our exploration we have a current potential solution including proposals for:

- ancestral species' sequences,
- locations of *cis*-regulatory elements in sequences,
- tree sequence alignments, and
- motif model logos.

To continue the exploration, we fix all but one part of the solution and then re-sample the one part, with a probability distribution conditioned upon the parts that are fixed and upon the known sequences. MCMC theory proves that *this random, iterative process visits solutions according to their posterior probabilities* given the known sequences.

Centroids

We take a census of *cis*-regulatory element locations from the Gibbs sampling iterations. Those locations present among a high number of the iterations have a corresponding high posterior probability. *These "centroid" locations are robust indicators of biologically meaningful sites,* more so than the locations from a maximum likelihood solution.

Algorithm summary

Realistic computational models
+ Gibbs sampling
+ Centroids
→ Robust predictions

Current status

Software for Gibbs sampling of ancestral sequences, element locations, sequence alignments, and regulatory element motif models (logos) is designed, implemented, debugged, and verified on small examples.

We are currently optimizing the code for speed so that it can run on larger data sets.