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

Lee Ann McCue,[1] **Lee A. Newberg**,[2,3]* (lee.newberg@wadsworth.org), William A. Thompson,[4] and **Charles E. Lawrence**[4]

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]Wadsworth Center, New York State Department of Health, Albany, New York; [3]Rensselaer Polytechnic Institute, Troy, New York; and [4]Brown University, Providence, Rhode Island.

No project URL.

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.

This poster's message: 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. This combination enables us to better locate *cis*-regulatory elements.

The details: Key to understanding gene regulation in the biofuel pathways is the locating of the *cis*-regulatory elements that are recognized by regulatory transcription factors. Each regulatory factor energetically prefers a short sequence of nucleotides, typically 10-30 nucleotides in bacteria. There is much tolerance for variation in the recognized sequence and thus the computational search for a regulator's elements is a search for a set of similar short sequences that are present significantly more often than would be expected by chance. The search is through the genomic regions near genes that are likely to be co-regulated and across species that are likely to exhibit similar pathways of interest.

Quantification of the significance of abundance and similarity among *cis*-regulatory elements depends upon the hypothesized evolutionary relationships of the elements. For example, when input sequences are nearly identical the similarity of putative *cis*-regulatory elements is a less compelling indication of functional conservation than when the same amount of similarity is present in more distantly related sequences or in evolutionarily unrelated sequences; in the latter cases it is easier to conclude that the similarity is due to selective

pressures associated with functionality rather than mere evolutionary nearness. Almost always, multiple sequence alignment (or its variant, tree sequence alignment) is the way that the orthologous relationships among nucleotides are specified. However, alignments and regulatory elements are not independent; a supposition of an alignment affects the apparent quality of a set of proposed elements and, vice-versa. As a result, a motif finding algorithm that searches for *cis*-regulatory elements using a multiple alignment built using a tool that is unaware of elements, and those that search in unaligned sequences both make sacrifices in the trade-off between sensitivity and specificity.

With a Markov chain Monte Carlo, Bayesian approach, we are simultaneously searching the space of possible sequence alignments and possible regulatory elements to find combinations that work well together to enhance the identification of unreported *cis*-regulatory elements. Here we present preliminary results regarding the effectiveness of this Bayesian approach as applied to biofuel pathways in alpha-proteobacteria.

This research is supported by the Office of Biological and Environmental Research in the DOE Office of Science under a multi-institutional grant entitled "Bayesian computational approaches for gene regulation studies of bioethanol and biohydrogen production." (Grant No. DE-FG02-09ER64757 to C.E.L., DE-FG02-09ER64756 to L.A.N., and Grant No. FWP 55426 to L.A.M.)