Understanding Gene Regulation in *Mycobacterium tuberculosis*: The Phylogenetic Gibbs Centroid Sampler for *Cis*-Regulatory Element Discovery

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

Also see Newberg et al. (2007), *Bioinformatics*, **23**(14):1718-1727.

The Motivation

One-third of the world's population now carries *Mycobacterium tuberculosis*, the bacterium that causes tuberculosis, and new infections occur at a rate of one per second. Tuberculosis has the highest morbidity of any bacterial disease, causing more than a million deaths each year.

The transcription regulatory network is key to a bacterial pathogen's survival and ability to cause disease. It is the primary means by which the pathogen senses and responds to its environment. Consequently, the network's regulatory proteins are promising targets for new therapeutics. However, a thorough understanding of the transcriptional regulatory network of *M. tuberculosis* remains elusive.

We are developing the computational tools necessary for discovering an organism's transcriptional regulatory network via the analysis of genome sequence data of the organism and other closely related species. We expect to delineate the transcription regulatory of *M. tuberculosis* network by the computational identification of DNA regulatory sequences, short sequences from the genome that play a crucial role in gene regulation, and by the classification of genes into groups that are regulated together.

The Approach

We detect the functional DNA elements by locating short sequences that are mostly conserved across species and among co-regulated genes. However, there are two major impediments: (1) the genomes under examination are from species that are closely related, thus even junk DNA may appear to be mostly conserved from genome to genome and (2) genomes are very long, thus some nonsense sequences will be overabundant, just by chance.

To address the first impediment, we employ an evolutionary-tree-based "phylogenetic" statistical model for nucleotide mutation that precisely quantifies the extent to which cross-species similarity is noteworthy.

To address the second impediment, instead of seeking maximum-likelihood solutions, we seek "centroid" solutions. A centroid solution represents the region of solution space with the most posterior probability rather than the single solution that is most probable -- and that is where the real solutions are.

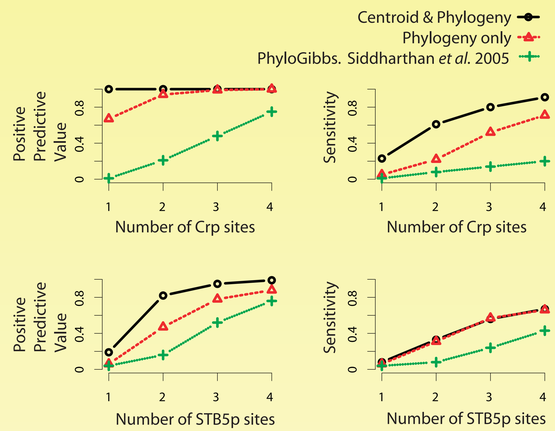
The Preliminary Results

We have tested our software tools in two settings: (1) synthetic sequence data and (2) real sequence data from *Escherichia coli* and related species.

Synthetic Data Tests

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 rediscovered) and positive predictive value (the fraction of predictions that were planted sites) increase significantly:



Real Data Tests

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 = 46%
Positive Predictive Value = 82%

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

The Bottom Line

We have tested the tools on synthetic and *E. coli* data with success. The ultimate goal of this research is to facilitate the identification of putative drug targets and vaccine candidates for tuberculosis and to provide tools that can subsequently be applied to other pathogens.

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