A Genetic Algorithm Approach to Regulatory Motif Discovery

Mattingly, C¹, Aman, J¹ and Congdon, CB¹.².³

¹Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672
²Department of Computer Science, Colby College, Waterville, ME 04901
³University of Southern Maine, Portland, ME 04103

The ATP-binding cassette (ABC) superfamily is comprised of proteins that play significant roles in multidrug resistance and diseases such as cystic fibrosis. Despite their clinical importance, regulation of expression of ABC genes is not well understood. Recent studies demonstrated that cross-species genomic sequence comparisons greatly enhance identification of functionally conserved regulatory regions. However, most motif identification tools fail to account for phylogeny and rely on information about known transcription factor binding sites (TFBSs). We are developing GAMI (Genetic Algorithms for Motif Inference), a genetic algorithm approach to motif inference, and using it in combination with other sequence analysis tools to enable de-novo motif discovery through comparison of evolutionarily divergent sequences.

A comprehensive control study was conducted to develop and evaluate GAMI’s ability to identify known motifs in randomized non-coding sequences. Parameters evaluated included sequence length (1K, 10K, 50K), number of sequences (6, 12), motif length (12 bases, 20 bases) and degree of motif degeneracy (0%, 10%, 20%, 30%). Studies were conducted in parallel with two other motif-finding programs, MEME and Gibbs. Results indicate that these programs perform comparably in 1K and 10K data sets; 20 base motifs are found more reliably than 12 base motifs; motifs with more than ~20% degeneracy are not found reliably; and inclusion of more sequences (12 vs 6) improves the ability to find implanted motifs. These results informed the construction of “real” data sets consisting of non-coding sequences for several ABC genes from divergent organisms. These data sets were analyzed using GAMI, MEME and Gibbs. Results comprise both known and possibly novel regulatory motifs for these genes. This presentation reports results from these early experiments and discusses future development plans for GAMI.