Initial results for GAMI: A genetic algorithms approach to motif inference

Clare Bates Congdon[1], Charles W. Fizer[1], Noah W. Smith[1], H. Rex Gaskins[2], Joseph Aman[3], Gerardo M. Nava[2], and Carolyn Mattingly[3]

[1]Department of Computer Science, Colby College, Waterville, ME 04901
[2]University of Illinois at Urbana-Champaign, Urbana, IL 61801
[3]Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672

We have developed GAMI[1], an approach to motif inference that uses a genetic algorithms (GA) search. Specifically, we are looking for putative conserved regulatory regions in non-coding sequence; several studies suggest that comparative analysis of evolutionarily diverse organisms will help to predict functionally important non-coding regions[2]. GAMI searches the space of possible motifs to find those that are most strongly represented across the sequences of interest.

We have been working with several genes studied by MDIBL researchers. The preliminary results reported here show motifs found in the 5' upstream and 3' downstream regions of ABCC7, the cystic fibrosis transmembrane conductance regulator (CFTR), using human, mouse, pig, chicken, and Fugu as representative divergent species. In this work, we looked for the strongest 20mers in the data, which went upstream or downstream from the CFTR gene until the next known gene (up to 50kb). We also show preliminary results from the upstream and downstream regions of GST Mu1, using human, mouse, dog, Xenopus, zebrafish, Fugu, and Ciona; in this study, we looked for the strongest 10mers in the data, which went upstream or downstream 4kb for all species. Example motifs found are illustrated in Figures 1 and 2.

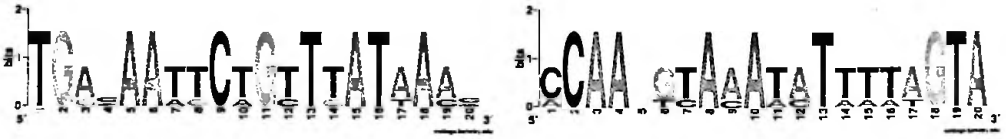

Fig. 1. Example motifs found for CFTR data; 5' upstream on the left and 3' downstream on the right.

Fig. 2. Example motifs found for GST Mu1 data; 5' upstream on the left and 3' downstream on the right.

To date, we have demonstrated GAMI to be an effective tool for searching large datasets (long sequence lengths and possibly many sequences) of divergent species. The system has been validated for small problems, finding known transcription factor binding sites (TFBS) referenced in other published work and finding the best motifs identified by exhaustive search. We have also compared the CFTR motifs found by GAMI to the full human genome and to known TFBSs. These motifs are non-promiscious; some of these motifs represent known TFBS for other genes while some of these motifs may represent novel discoveries.

Supported by NIH Grant Number 2-P20-RR-016463-04 from the INBRE Program of the National Center for Research Resources and with funds from MDIBL's NIEHS Center for Membrane Toxicity Studies, Grant Number ES03828-19.

1. **Congdon C. B., C. W. Fizer, N. W. Smith, H. R. Gaskins, J. Aman, G. Nava, and C. Mattingly.** Preliminary Results for GAMI: A Genetic Algorithms Approach to Motif Inference. *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*: 97-104, 2005.
2. **Dubchak I., M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer.** Active conservation of non-coding sequences revealed by three-way species comparisons. *Genome Res.* 10: 1304-1306, 2000.