Increasing our ability to identify candidate functional elements in noncoding DNA

Jeffrey A. Thompson[1], David J. Gagne[1] and Clare Bates Congdon[1]
[1]Department of Computer Science, University of Southern Maine, Portland, ME 04104

Analysis of genomic data has revealed the existence of functional units in noncoding regions of DNA that regulate the transcription of genes, typically by forming binding sites for proteins that affect transcription. Computational approaches can help search vast amounts of genetic data to identify candidate functional elements in noncoding DNA for study in the lab. In this work, we expand our computational abilities to infer candidate elements in the noncoding regions of co-expressed genes and in identifying *cis*-regulatory modules.

GAMI[3] (Genetic Algorithms for Motif Inference) uses a Genetic Algorithms (GA) search to identify candidate functional elements in noncoding DNA. The system was designed to identify putative functional elements following the notion that elements that have been conserved across evolution are more likely to be functional; therefore, GAMI seeks to find highly conserved patterns in the data, which are called motifs. In previous work, GAMI has been shown to be adept at finding highly conserved elements in long sequence lengths (*e.g.,* 100kb and longer) and across several dozen sequences. While an interest in studying the cystic fibrosis transmembrane conductance regulator (CFTR) gene continues to motivate this work, we proceed here with data that allow us to evaluate the efficacy of our newly designed approaches.

Co-expressed genes frequently share regulatory elements[1,5], and therefore, it would be desirable to use GAMI for the inference of motifs in co-expressed genes. We have extended GAMI to allow some of the sequences in the data to drop out of the search process to enable this, and have evaluated this new capacity on several datasets, including pregnancy gene data[6] and NF-kappa B transcription factor motif data[4], and demonstrated the system's ability to identify functional elements beyond the scope of the intentions of the original software design.

Regulatory information in the non-coding DNA of higher organisms is often organized into modular units, known as *cis*-regulatory modules (CRMs)[2]. We have developed additional computational tools to filter the motifs found by GAMI to identify those that are members of candidate CRMs. Initial work on this system has been conducted with artificial data. Initial results show that implanted (known) modules are identified in most situations evaluated, although with extremely long sequence lengths (1 million bp long) and with high degradation of the implanted motifs (over 20%) results are less than 100%.

This project was supported by NSF CAREER award 953495 (CBC), the National Center for Research Resources (5 P20 RR024475-02) and the National Institute of General Medical Sciences (8 P20 GM103534-02) from the National Institutes of Health (CBC; PI: Dr. Jason Moore, Dartmouth Medical School), NSF Cooperative Agreement HRD-0833567 (USM), and Maine Economic Incentive Funds (CBC).

1. **Allocco, DJ, Kohane, IS and Butte, AJ.** Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5:18, 2004.
2. **Berman, BP, Nibu, Y, Pfeiffer, BD, Tomancak, P, Celniker, SE, Levine, M, Rubin, GM and Eisen, MB.** Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the drosophila genome. *PNAS* 99:757-762, 2002.
3. **Congdon, CB, Aman, JC, Nava, GM, Gaskins, HR and Mattingly, CJ.** An evaluation of information content as a metric for the inference of putative conserved noncoding regions in DNA sequences using a genetic algorithms approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 5:1-14, 2008.
4. **Fogel, GB, Weekes, DG, Varga, G, Dow, ER, Harlow, HB, Onyia, JE and Su, C.** Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucl. Acids Res.* 32: 3826-3835, 2004.
5. **Monsieurs, P, Thijs, G, Fadda, AA, De Keersmaecker, SCJ, Vanderleyden, J, De Moor, B, and Marchal, K.** More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics*, 7:160, 2006.
6. **Thompson, HG, Harris, JW, Wold, BJ, Quake, SR and Brody, JP.** Identification and confirmation of a module of coexpressed genes. *Genome Res*. 12:1517-1522, 2002.