RECONSTRUCTING TRANSCRIPTIONAL REGULATORY NETWORKS VIA THE INTEGRATION AND OPTIMISATION OF MULTIPLE BINDING SITE PREDICTION ALGORITHMS

Alistair G. Rust [1], Stephen A. Ramsey [1], Mark Robinson [2] & Hamid Bolouri [1]
1. Institute for Systems Biology, Seattle, WA 98103, USA
2. Biocomputation Group, STRC, University of Hertfordshire, AL10 9AB, UK

{arust,sramsey,hbolouri}@systemsbiology.org || m.1.robinson@herts.ac.uk

Reconstructing transcriptional regulatory networks from experimental data is one of the corner-stones of systems biology. ChIP2chip experiments for example, allow the elucidation of regulatory network structure based upon protein-dna interactions. Such data is however highly variable and taken alone, results in many false-positive connections in the predicted networks. To increase confidence in predictions, by confirming or rejecting connections, additional sources of evidence should be included. Protein-dna interactions derived from computationally-predicted transcription factor binding sites are one such source.

A large variety of computational tools already exists to predict binding sites or motifs, where each algorithm delivers its own specific insights. We advocate that through the integration of multiple prediction algorithms, more reliable and robust predictions of binding sites can be obtained. Consensus predictions from multiple algorithms are reinforced and mismatches rejected. Therefore, we have implemented a software framework that integrates 12 prediction algorithms into an automated motif analysis pipeline. The system is called Mogul (*gaelic*; mesh of a net) and it incorporates key classes of algorithms:

- **Motif Scanners**. These algorithms analyze intergenic sequences using matrices derived from experimentally verified transcription factor binding sites.
- **Single Scan**. Included in this category are algorithms that effectively perform *ab-initio* scans of single sequences.
- **Co-regulated**. This class incorporates algorithms that analyze multiple sequences, whose genes share similar expression profiles. The algorithms use the hypothesis that the analyzed set of intergenic sequences share common binding sites.
- **Comparative**. These algorithms seek to identify conserved blocks of intergenic regions for pairs of sequences.
- **Evolutionary**. Using the increasingly large number of sequenced genomes, these algorithms incorporate phylogenetic information to make their predictions.

Each class of prediction algorithm is represented in the pipeline, including a number of the most widely-used motif prediction algorithms such as AlignACE [1], meme [2] and FootPrinter [3]. The choice of algorithms to run is highly-configurable such that the user can, for example, choose to run just a single class of algorithms or even an individual algorithm alone. An interface to the Mogul pipeline is shown in Figure 1(a).

Consensus motif predictions are output in gff format. This enables results to be viewed in a variety of ways, such as in simple text format or via visualization tools that can interpret gff format. For example, predicted motifs alignments can be viewed using the Apollo genome browser [4] (see Figure 1(b)) and also in postscript/pdf using the gff2ps utility [5]. Aggregated results of predictions from multiple intergenic regions can be exported and viewed in the network visualization tool Cytoscape [6] (see Figure 1(c)).

Applying the pipeline to the analysis of sequences from different species will require the parameter settings of the algorithms to be tuned to gain the best predictions. We are therefore using optimization methods to improve upon the default parameters of the included algorithms.

The system is being systematically enhanced with the inclusion of additional prediction algorithms, species-specific tailored search options and alternate modes for viewing results. The pipeline is currently being used to elucidate networks from different species, in particular yeast and sea urchin. Results will be presented in full at the conference.

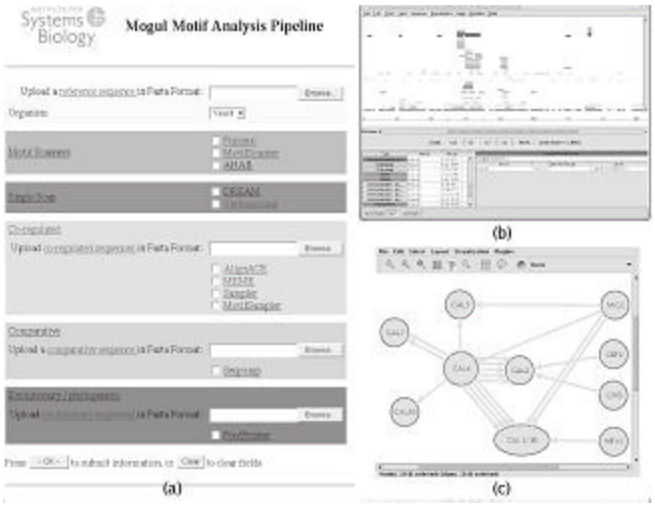

Figure 1. (a) The web interface to the Mogul motif analysis pipeline. (b) An example output of predicted motifs displayed in the Apollo genome browser tool. (c) The integration of predicted motifs within the Cytoscape network visualization package.

References:

[1] Roth, FR, Hughes, JD, Estep, PE and Church GM, Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation, *Nature Biotechnology,* 16(10):939-45, 1998

[2] Bailey, TL and Elkan, C, Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers, *Procs. of the 2nd Int. Conf. on Intelligent Systems for Molecular Biology*, 28-36, 1994.

[3] Blanchette, M and Tompa, M, FootPrinter: A Program designed for Phylogentic Footprinting, *Nucleic Acids Research*, 31(13):3840-3842, 2003

[4] Lewis, S.E. et al, Apollo: A Sequence Annotation Editor, *Genome Biology,* 3(12):research0082, 2002.

[5] Abril, JF and Guigó, R, gff2ps: Visualizing Genomic Annotations, *Bioinformatics*, 16(8):743-744, 2000.

[6] Cytoscape Homepage: http://www.cytoscape.org