Comparative Genomics with cMonkey: Integrative Biclustering of Multiple Species

Alex Greenfield[2], Thadeous Kacmarcyzk[1], Peter Waltman[2], Richard Bonneau[1,3]

cMonkey is an algorithm that identifies gene modules using multiple data types. The input is expression data, upstream sequence information, and association networks. The output is an ensemble of modules, or biclusters. Each bicluster contains a subset of genes and a subset of conditions under which these genes are co-regulated. Here, we extend the original cMonkey algorithm to simultaneously bicluster the genomes of multiple species. Initial results indicate that this approach yields evolutionary insights into the formation and conservation of regulatory modules.

***Keywords* — clustering, integrative biclustering**

ORGANISMS respond to their changing environment via complex regulatory interaction networks, relaying information to target genes and cellular processes. Extensive data from high-throughput experimental technologies challenge systems-biologists to create new methods to clarify these inherently complex regulatory networks.

Integrative biclustering: Complete and accurate models of complex biological systems benefit from the integration of multiple forms of evidence derived from measuring these systems on different information levels (e.g. interaction networks, sequence motifs, protein and RNA expression, etc.). Biclustering, the simultaneous clustering of both genes and experiments, has emerged as an effective approach for the analysis of multiple systems biology data-types. Recently, we have introduced cMonkey[1], an algorithm that allows one to integrate diverse systems biology data-types to form optimal biclusters[2].

Comparative Integrative Biclustering: Here, we have extended the original cMonkey algorithm to simultaneously bicluster the genomes of multiple species. This method provides a framework that allows the insights from a well-studied organism to aid in the analysis of related but less-studied organisms. By leveraging the power of comparative analysis, we identify both conserved modules of orthologous genes, as well as those that have diverged, yielding evolutionary insights into the formation and conservation of regulatory modules. We present initial results from the integrative biclustering of two prokaryotic species *Bacillus subtilis* and *Bacillus anthracis*.

[1] Center for Genomics and Systems Biology, New York University
tjk229@nyu.edu, bonneau@nyu.edu
[2] Computational Biology Program, New York University
ag1868@nyu.edu, waltman@cs.nyu.edu
[3] Courant Institute of Mathematical Sciences, New York University
bonneau@nyu.edu

REFERENCES

[1] Bonneau R, et.al. (2006) Integrated Biclustering of Heterogeneous Genome-wide Datasets for the Inference of Global Regulatory Networks. *BMC Bioinformatics* 7:280

[2] Bonneau R, et. al. (2007) A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. Cell **131**, 7:135