27. Computation Component of ENIGMA: from data to predictive models

Pavel S. Novichkov[1*](psnovichkov@lbl.gov), Serdar Turkarslan[2], Alexey E. Kazakov[1], Sarah P. Preheim[3], Marcin P. Joachimiak[1], Max Shatsky[1], Michael S. Samoilov[4], John-Marc Chandonia[1], Inna Dubchak[1], Adam P. Arkin[1], Eric J. Alm[3], **Nitin S. Baliga[2]** and **Paul D. Adams[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Institue for Systems Biology, Seattle, WA; [3]Massachusetts Institute of Technology, Cambridge, MA; [4]Department of Bioengineering, University of California Berkeley, Berkeley, CA

http://enigma.lbl.gov

Project Goals: The overarching goal of the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) project is to understand environmentally- relevant microbial communities and processes through an integrated field-to-laboratory approach.

The Computation Component facilitates the analysis, management, and dissemination of data generated as part of ENIGMA. These activities have a reciprocal relationship with the other Core Components – They support the scientific objectives of the cores, and in turn the scientific needs drive the development of new, innovative computational tools. ENIGMA has put a strong focus on all three aspects of computation. Analysis is a strong point of the ENIGMA team, but roughly equal emphasis has put on data management, and visualization/sharing of data.

ENIGMA has developed innovative tools to make high-quality **prediction of associations between the content of microbial community and various geochemistry parameters**. These include a novel Distribution-based OTUs algorithm for grouping of individual reads into operational taxonomic units (OTUs) and the SparCC algorithm for identifying correlations within "compositional data", in which proportions rather than actual numbers are counted.

Molecular network inference continues to be a strength of ENIGMA. We have advanced and integrated the regulatory network inference algorithms into an ensemble modeling framework. This framework takes as input just gene expression data from carefully designed experiments and genomic sequence to (1) simultaneously discover environment-dependent membership of genes within co-regulated modules, (2) predict transcriptional changes within each module in new environments, and (3) predict cis- and trans-acting regulators for each module. Further, we have developed a suite of algorithms for automated reconstruction of regulons through comparative genomics

Data standardization is key to a large scale project like ENIGMA with multiple data types. The Computation Component recently completed a survey to identify and prioritize data exchange modalities for standardization. To facilitate this effort, we released a Data Management Guide that includes standards for data types (esp when produced by several labs), links to data resources, guidance for experiment planning, sample identifiers, etc.

Data dissemination both within and outside of ENIGMA continues to be a focus of our efforts. ENIGMA maintains several widely used websites, including MicrobesOnline, Network Portal, and RegPrecise, and has begun to incorporate ENIGMA data sets and algorithms into DOE Systems Biology Knowledgebase (KBase). Currently four ENIGMA computational tools are implemented as KBase services: cMonkey, Inferelator, MAK, and BAMBI. Within ENIGMA, data dissemination is critical to engage multiple off-site investigators in meaningful scientific dialog.

This work conducted by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory, was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02- 05CH11231.