Integrated Computational Biology Capabilities at BESC

Jeremy Smith,[1,2] Ed Uberbacher[2]* (uberbacherec@ornl.gov), Eberhard Voit,[3] **Ying Xu**,[4] Brian Davison,[2] and **Paul Gilna**[2]

[1]University of Tennessee, Knoxville; [2]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Georgia Institute of Technology, Atlanta, Georgia; and [4]University of Georgia, Athens

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). Addressing the roadblock of biomass recalcitrance will require a greater understanding of plant cell walls from synthesis to deconstruction. This understanding would generate models, theories and finally processes that will be used to understand and overcome biomass recalcitrance. This grand challenge calls for an integrated research approach, which is illustrated by the project described below. BESC research is multidisciplinary by design and multi-institutional in composition.

The Computational Biology Team of the BESC Center has been developing a suite of computational biology and bioinformatics tools in support of the needs of researchers at BESC and beyond. These tools address *omics* data analysis and interpretation, genome assembly and annotation, protein and assembly structure/functional prediction and analysis, phylogenetic tree reconstruction and analysis, as well as inference, analysis and modeling of regulatory and metabolic pathway systems. In addition, a number of databases have been developed for organizing data generated to support BESC research. These databases include the DOOR operon database, genome Barcode server cBAR, pDAWG for plant cell wall genes, and the GASdb database for glycosyl hydrolysis systems. Most of these tools and databases are in the process of being integrated into the BESC Knowledge Base (KBase).

Currently the core KB for plants consists of 21 plant genomes including six algal genomes, along with a rich set of annotated data and computed information for: (a) gene structures; (b) protein products; (c) homology-based functional prediction; (d) domain structures; (e) ortholog and paralog prediction; (f) gene ontology; and (g) metabolic and enzymatic pathways. Currently, the KB's reference plant data consist of over 500,000 coding genes from which nearly 400,000 protein coding genes with function prediction have been identified. The Plant KB maintains available gene model variations, alternative gene models (including alternative splicing), and historical versioning. The CompBio team has also collected rich set of omics data from external resources like NCBI GEO and EMBL ArrayExpress that are related to the assembly of the cell wall pectic matrix, cellulose synthesis and cell elongation, cellulose synthase mutants, cell wall stress, primary to secondary stem development and many other experiments for *Arabidopsis*, Rice and Popular.

The core KB for microbes consists of information regarding 37 microbes, including biomass degraders, fuel producers, endophytes and model organisms. Data collected on these organisms include genome annotation, biochemical data including enzymes, ligands and pathway annotation generated using the Pathway Tools software and provided in KEGG, carbohydrate active enzyme data generated by CAT and provided by CAZy, operon predictions from BeoCyc and DOOR,

protein functional domain predictions from CDD, Pfam, COGS, TigrFam, SMART. The KB also contains omics data on the effects of alcohol on cells, fermentation time courses, growth on model substrates found in lignocelluloses, growth on simple and complex sugars, biomass deconstruction, and many other experiments from several *Clostridium* and other microbes relevant in bioenergy research.

Using these capabilities, the CompBio team has carried out a variety of computational studies in support of the BESC mission. Examples include: (1) a molecular dynamics simulation of lignocellulosic biomass; (2) simulation of enzyme reaction mechanisms using combined quantum mechanical and molecular mechanical models; (3) genome annotation of *Caldicellulosiruptor bescii;* (4) detailed functional inference and phylogeny analyses of a number of gene families associated with plant cell walls, such as GT8 genes, the cellulose synthase superfamily, UDP-4-keto-pentose/UDP-xylose synthase, nucleotide sugar interconversion enzymes and syringyl lignin biosynthesis genes; (5) modeling of lignin biosynthesis in *Populus* and *Medicago*; and (6) systems biology studies of *C. thermo*.

Our ultimate goal in developing these tools and the KBase is to provide one unified interface and computational environment to facilitate integrated analyses of plant cell-wall related genes, pathways and enzymes, and microbes capable of degrading lignocellulosic materials.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.