The United States Department of Energy (DOE) initiative to increase volumes of diverse data including genome sequences, many types of omics data and various assay results related to biomass properties, structure and composition. Besides managing data generated by BESC researchers, integration of key reference data such as genome, gene annotations and metabolic annotation for bioenergy relevant organisms is a central component, which provides a context for understanding experiments. In this article, we describe some important computational tools and data available from the BESC Knowledgebase.

1 INTRODUCTION

The United States Department of Energy (DOE) initiative to increase the production of renewable biofuels such as cellulosic ethanol has led researchers to explore various biomass feedstocks and microbes for alternative fuel solutions (Farrell, 2006; Gnanasundram and Dauriat, 2010; Rubin, 2008; Stephanopoulos, 2007; Zacchi et al., 2006). The BioEnergy Science Center (BESC) was established by DOE in 2007 as a multi-institutional partnership driving scientific efforts in this direction. BESC is undertaking large experimental campaigns to understand and mitigate the recalcitrance of biomass feedstocks and to develop biofuel solutions. BESC researchers, integration of key reference data such as genome, gene annotations and metabolic annotation for bioenergy relevant organisms is a central component, which provides a context for understanding experiments. In this article, we describe some important computational tools and data available from the BESC Knowledgebase.

BESC KNOWLEDGEBASE DATA AND TOOLS

Reference genomic data: the microbial domain of the knowledgebase currently contains 37 microbial genomes. Currently, the KB’s reference microbial data consists of over 134 000 protein coding genes. We continually add new organisms of interest, genome annotations for organisms including BLAST hits, domain annotations, protein localization, transcription unit data, enzyme and pathway annotations. The microbial index page provides a set of search interfaces to query the annotations and to download search results. Reference genomic data are also available for each gene through a unified interface, a gene card. The plant domain consists of 21 plant and algal genomes along with a rich set of their annotations including gene structures, protein products, homology-based functional prediction, domain structures, ortholog and paralog prediction, gene ontology, and metabolic and enzymatic pathways. Currently, the KB’s reference plant data consists of over 500 000 coding genes from which nearly 400 000 protein coding genes with function prediction have been identified. The database keeps track of available gene model variations and alternative splicing variants.

Phenotype comparison toolkit (CBP): this toolkit provides tools to compare microbes across different phenotypes or genetic traits at various levels of organization, such as whole genome, a biological process, enzyme family, domain or sequence level. Comparisons can be made between microbes with different phenotypes such as aerobes and anaerobes, thermophiles and mesophiles. For each of the phenotypes, users can select a group of organisms and compare the groups in terms of their metabolic pathways, enzyme profiles, protein family domains and orthologs.

BioCyc: we have a collection of Pathway/Genome DataBases (PGDBs) for Bioenergy relevant organisms, which we call ‘BioCyc’. We have reconstructed metabolic pathways using the Pathway Tools software (Karp et al., 2010). Although the PGDB

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

The BioEnergy Science Center (BESC) is undertaking large experimental campaigns to understand the biosynthesis and biodegradation of biomass and to develop biofuel solutions. BESC is generating large volumes of diverse data, including genome sequences, omics data and assay results. The purpose of the BESC Knowledgebase is to serve as a centralized repository for experimentally generated data and to provide an integrated, interactive and user-friendly analysis framework. The Portal makes available tools for visualization, integration and analysis of data either produced by BESC or obtained from external resources.

Availability: http://besckb.ornl.gov
Contact: syedmht@ornl.gov
Received on October 5, 2011; revised on December 20, 2011; accepted on January 5, 2012
is generated automatically by the Pathologic program from the Pathway Tools software, annotation of the genome with EC numbers was improved before the reconstruction by using enzyme prediction tools, like KAAS (Moriya et al., 2007), and by searching for orthologous genes in model organisms.

Pathway Tools software provides a diverse set of options to query, visualize information in the databases, overlay experimental data on metabolic maps and pathways and perform comparative analysis.

Gbrowse: we have also configured Gbrowse—a popular genome browser (Stein et al., 2002) for all complete microbial genomes in the BESC knowledgebase. Besides genes, we have operon predictions from BesCyc/Pathway Tools software (Karp et al., 2010) and a database for prokaryotic operons (Mao et al., 2009) available from the browser. SNP/indels from resequencing of an ethanol tolerant Clostridium thermocellum strain (Brown et al., 2011) is also made available from the genome browser. Users can also upload their own data, such as SNPs/indels from some new strain, and compare it with data from BESC by drawing tracks parallel to default tracks available from the browser.

Integration and analysis of omics data from BESC, GEO and ArrayExpress: the BESC Knowledgebase provides tools that allow the user to search for experiments in external resources such as NCBi GEO (Boyle, 2005) using keywords, bring them into the local analysis environment and integrate datasets with genomic data in the BESC Knowledgebase. Additional tools have been developed in the framework for statistical analysis, such as generating interactive scatterplots, heatmaps and mapping experimental data onto pathways.

Cazyymes Analysis Toolkit: the BESC Knowledgebase also hosts the Cazyymes Analysis Toolkit—CAT (Park et al., 2010), developed and published earlier. This toolkit provides methods to search and annotate CAZymes. This tool has already been used outside BESC to annotate genomes (Kikuchi et al., 2011).

BESC KB Genome Resequencing toolkit: the target for resequencing is usually the genome of a mutant strain with a practically important phenotype evolved by an adaptation of a wild-type strain. The toolkit provides a way to identify genomic modifications underlying the specific phenotype of the mutant and to understand potential biological effects of the mutation. The tool kit consists of the following tools that can be applied either as a pipeline or independently. ‘SNP-indel caller’ finds changes in the genome of the mutant strain and their location given a file with high confidence 454 reads as the input. ‘Mutant protein fasta’ generates a FASTA file of proteins for the mutant strain. ‘Mutant protein CDD’ annotates the proteins with protein family domains using the CDD pipeline for the mutant strain. ‘Mutation Mapper’ annotates all identified changes in the genomic sequence of the mutant strain, its position in the genome, location of the change within an intergenic region or gene, the type of change (synonymous or non-synonymous, multiple changes, insertions or deletions), and amino acids and codons in the M and WT strains. ‘Regulation change predictor’ produces the list of genes with mutational changes, either SNPs or indels, in the upstream intergenic regions. Entire sequence span between coding sequences of two adjacent genes are used without predicting transcription factor binding sites. ‘Function change predictor’ predicts potential changes in the protein function, like its gain or loss. Detail documentation of the toolkit is available from BESCKB website (http://cricket.ornl.gov/html/download/resequencing/Resequencing-ToolkitDocumentation_16Dec2011.pdf). The tools were applied to analyze the resequencing data for the ethanol adapted strain of C.thermocellum ATCC 27405 (Brown et al., 2011).

3 CONCLUSIONS

We have made available suite of tools and diverse types of bioenergy relevant data through the BESC public portal. The tools may be especially helpful in integrating variety of data such as genomic, phenotypic, metabolic and experimental data, and gaining comprehensive, system level understanding of cellular processes involved in plant biomass formation, degradation and biofuel production.

ACKNOWLEDGEMENTS

Authors would like to thank Dr Paul Gilna for providing valuable comments and suggestions for this article.

Funding: Office of Biological and Environmental Research in the Department Of Energy Office of Science through the BioEnergy Science Center, a Department Of Energy Bioenergy Research Center. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the Department Of Energy under Contract DE-AC05-00OR22725.

Conflict of Interest: none declared.

REFERENCES


