BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 5 2012, pages 750–751
doi:10.1093/bioinformatics/bts016

Databases and ontologies

Advance Access publication January 11, 2012

BESC knowledgebase public portal†

Mustafa H. Syed[1,2,*], Tatiana V. Karpinets[1,2,3], Morey Parang[1,2], Michael R. Leuze[1,4], Byung H. Park[1,4], Doug Hyatt[1,2], Steven D. Brown[1,2], Steve Moulton[5], Michael D. Galloway[5] and Edward C. Uberbacher[1,2]

[1]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN, [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, [3]Department of Plant Sciences, University of Tennessee, Knoxville, TN, [4]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA and [5]Information Technology Services Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Associate Editor: Janet Kelso

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: syedmh@ornl.gov

Received on October 5, 2011; revised on December 20, 2011; accepted on January 5, 2012

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; Gnansounou 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 for cellulolytic degradation by enzymes and organisms, and to develop multitalented microbes for converting plant biomass into biofuels in a single step (Lynd *et al.*, 2005). Researchers at BESC are focused on comprehensive and system level understanding of the process of biomass and plant cell wall formation, degradation and biofuel production. As such, the center is generating large 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.

2 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.

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

*To whom the correspondence should be addressed.

†This article has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article, or allow others to do so, for United States Government purposes.

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 BeoCyc/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.

Cazymes Analysis Toolkit: the BESC Knowledgebase also hosts the Cazymes 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

Boyle,J. (2005) Gene-Expression Omnibus integration and clustering tools in SeqExpress. *Bioinformatics*, **21**, 2550–2551.

Brown,S.D. *et al.* (2011) Mutant alcohol dehydrogenase leads to improved ethanol tolerance in Clostridium thermocellum. *Proc. Natl Acad. Sci. USA*, **108**, 13752–13757.

Farrell,A.E. (2006) Ethanol can contribute to energy and environmental goals (vol 311, pg 506, 2006). *Science*, **312**, 1748–1748.

Gnansounou,E. and Dauriat,A. (2010) Techno-economic analysis of lignocellulosic ethanol: a review. *Bioresource Technol.*, **101**, 4980–4991.

Karp,P.D. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **11**, 40–79.

Kikuchi,T. *et al.* (2011) Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus. *PLoS Pathog.*, **7**, e1002219.

Lynd,L.R. *et al.* (2005) Consolidated bioprocessing of cellulosic biomass: an update. *Curr. Opin. Biotechnol.*, **16**, 577–583.

Mao,F. *et al.* (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–D463.

Moriya,Y. *et al.* (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.

Park,B.H. *et al.* (2010) CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology*, **20**, 1574–1584.

Rubin,E.M. (2008) Genomics of cellulosic biofuels. *Nature*, **454**, 841–845.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Stephanopoulos,G. (2007) Challenges in engineering microbes for biofuels production. *Science*, **315**, 801–804.

Zacchi,G. *et al.* (2006) Bio-ethanol - the fuel of tomorrow from the residues of today. *Trends Biotechnol.*, **24**, 549–556.