Computational Resources for Biofuel Feedstock Species

Final Progress Report for DE-FG02-08ER64631

(08/15/2008-08/14/2012)

ABSTRACT

While current production of ethanol as a biofuel relies on starch and sugar inputs, it is anticipated that sustainable production of ethanol for biofuel use will utilize lignocellulosic feedstocks. Candidate plant species to be used for lignocellulosic ethanol production include a large number of species within the Grass, Pine and Birch plant families. For these biofuel feedstock species, there are variable amounts of genome sequence resources available, ranging from complete genome sequences (e.g. sorghum, poplar) to transcriptome data sets (e.g. switchgrass, pine). These data sets are not only dispersed in location but also disparate in content. It will be essential to leverage and improve these genomic data sets for the improvement of biofuel feedstock production. The objectives of this project were to provide computational tools and resources for data-mining genome sequence/annotation and large-scale functional genomic datasets available for biofuel feedstock species. We have created a Bioenergy Feedstock Genomics Resource that provides a web-based portal or “clearing house” for genomic data for plant species relevant to biofuel feedstock production. Sequence data from a total of 54 plant species are included in the Bioenergy Feedstock Genomics Resource including model plant species that permit leveraging of knowledge across taxa to biofuel feedstock species.We have generated additional computational analyses of these data, including uniform annotation, to facilitate genomic approaches to improved biofuel feedstock production. These data have been centralized in the publicly available Bioenergy Feedstock Genomics Resource (http://bfgr.plantbiology.msu.edu/).

REPORT:

This project was jointly funded with USDA for its first two years and the objectives of the USDA and DOE portions are closely intertwined and this progress report includes activities for both components of the project.

The primary objective of this project was the creation of the Bioenergy Feedstock Genomics Resource (BFGR; http://bfgr.plantbiology.msu.edu/) which provides a web-based portal or “clearing house” for genome sequence/annotation (structural, functional, and comparative), genetic data, germplasm data, and large-scale functional genomic datasets such as expression, metabolite, and proteomic profiles for plant species relevant to biofuel feedstock production. A total of three releases of the database were made with release 3 containing sequence data from 54 species, including model organisms with complete genomes, which are uniformly annotated and available through the project website. All data generated by the project are accessible through the website via webforms, genome browsers, or direct download. Search tools include 1) sequence-based searches through a BLAST server, 2) text-based functional annotation (i.e., gene name assignment), 3) InterPro domain and motif identification, 4) Gene ontology, 5) KEGG biochemical pathway membership, 6) Genetic markers (Simple sequence repeats, Single nucleotide polymorphisms, and 7) Sequence identifiers. A total of seven genome browsers, which are fully integrated with data in the BFGR, are available on the BFGR site thereby permitting comparative analyses among model organisms and biofuel feedstock species. While the web interface facilitates use by many biologists, there is a subset of users that wish to perform large-scale analyses and all data within the BFGR is available from the site's FTP server (ftp://ftp.plantbiology.msu.edu/pub/data/BFGR/).

We developed a centralized annotation report page that provides a uniform view of the annotation that is available for all sequences in the BFGR. Each annotation report page provides a functional description of the sequence, a link to view the transcript and protein sequence of the gene or transcript assembly, a link to the sequence annotation page at the website of the project from which the sequence was obtained, links to the sequence within the BFGR genome browsers, tables of alignments between the sequence and UniRef Proteins, InterPro domains/motifs, a list of significant alignments to KEGG database proteins, proteins from model and finished genomes within the BFGR, gene trees displaying orthologous/paralogous sequences from model and related species, a link to any available expression data, a table of computationally identified genetic markers within the sequence.

Robust documentation and documentation of the BFGR can be found via the main page (http://bfgr.plantbiology.msu.edu/index.shtml) which provides a description of the project and the type of data that can be accessed at the site. Several pages provide users with detailed answers to the methods (http://bfgr.plantbiology.msu.edu/FAQ.shtml) that have been used to generate data at the site and describe the latest updates that have been made to the website and its data (http://bfgr.plantbiology.msu.edu/new.shtml). In addition, the sequence summary page (http://bfgr.plantbiology.msu.edu/sequence_summary.php) allows users to quickly view the numbers and types of sequences that are available at NCBI for each of the species that are covered by BFGR. A full description of the BFGR database was published in 2012 in the journal Database by Childs et al. (doi: 10.1093/database/bar061).

To support the BFGR, we constructed Postgres-based chado databases (one per species) which contain the bulk of the data, and these databases are kept on a dedicated Postgres server while a separate custom Postgres database provides fast query-responses for the various search functions. All sequence searches are conducted on a dedicated server that supports all blast requests submitted by the public while the genome browsers are supported by MySQL databases that are hosted on a dedicated MySQL database server. As a consequence, the BFGR website provides rapid responses to user interactions.