INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Analysis of Storage Technologies of Biological Data

H. S. Shashidhara
Professor, Department of ISE, MSRIT, Bangalore, India
Krishangee Bora
Department of ISE, MSRIT, Bangalore, India

Abstract:
CottonGen is curate and integrated web based relational database which makes use of resources of available genetic, genomic useful for breeding data of cotton. It's features of genome sequences, EST's, markers, trait loci, genes, taxonomy marked with easy access through genome pages, search tools, G-Browse, search tools like Batch Blast make it highly useful. This paper does an in-depth study of how this database can help in the genetic improvement and overall higher production of cotton.
The second part of Diamondback moth database is an online repository for storing data of storing this genomic type diamondback moth database. Also search tools ease of access, downloadable datasets further help scientists to study comparative genomics, biological interpretation and gene annotation of the insect pest. The main objective is to reduce the harm by DBM on agriculture and thereby prevent losses. This paper presents DBM-DB coordinating genomic resources available for insect.

1. Introduction

1.1. Cottongene

Cottongene is the world's leading natural textile fibre crop and significant contributor of oilseed.50 species with different levels of plocdy,gossypium has answered biological questions on genome evolution, plant development, polyplodization, crop production which includes breeding information etc. on 49 genetic maps,24000 markers greater than 1000 quantitative trait loci(QTL) greater than 30 agronomical important traits greater than 15,000 germplasm accessions etc. expression data in the form of microarrays and RNA-sequence from high-throughput sequencing. This provides a major source of candidate genes which could contribute to the genetic improvement of cotton quality and productivity. Henceforth a database in this area would help in fuller utilization of resources.

The chronology of a database started as follows [1]:- Three online databases traditionally hosted much of the available genomic and genetic cotton data prior to 2012.CottonDB (founded in 1995) was used for all agricultural commodities. Then using a hybrid database system, the genomic, genetic, taxonomic and bibliographic data were stored in an object-oriented Ace DB database. However the genetic maps and genome sequences were maintained in a MySQL relational database. The third TropGeneCotton was a larger project to manage genetic, molecular and phenotypic data on tropical crop species using Java. ICGI is a non-profit organization created in 2000 to increase knowledge of the structure and function of the cotton genome for the benefit of the global community. Cotton DB changed to cottongene by the following changes as given [1]: - -> the Gossypium Raimondi whole genome assemblies annotated unigene for the Gossypium genus

- Extensive esgenetic and QTL maps, markers and traits
- Trait evaluation of data.
- Enhanced user interfaces including various search tools with downloadable results and
- Resources to support the various cotton activities and enhance communication between the many cotton researchers.

Cotton Gen includes the first fully sequenced cotton species, Gossypium Raimondi whereby, these assemblies are titled the 'Gossypium Raimondi (D5) genome JGI[1]. Also the predicted genes from these assemblies have been further annotated by the team to include homology to proteins in other well annotated or closely related species, and in silico annotation of InterPro protein domains, GO terms and Kyoto Encyclopedia of Genes and Genomes database (KEGG) pathway terms etc. , providing information on probable pathways and traits. Additional annotation by the CottonGen team includes the alignment of cotton. Single nucleotide polymorphisms (SNPs) between the diploid genomes of A and D and those between the tetraploid of AT and DT whereby T represents tetraploid) were also aligned to the JGI version of the G. Raimondi. In the case of annotated EST unigene CottonGene contains all Gossypium

EST's publicly from the NCBI. To avoid redundancy in EST's CottonGen v1.0 unigene came up. Routine processing involving sequence filtering for contamination against the NCBI UniVec database and species-specific chloroplast, mitochondrial, tRNA and rRNA sequences using the BLAST algorithm with NCBI UniVec-recommended parameters etc. [2]. In the domain of NCBI genes all Gossypium sequences from the NCBI nucleotide database were downloaded, parsed for gene, mRNA,CDS, 50 UTR and 30 UTR features and imported to CottonGen. Further sequence parsed from NCBI annotated by homology to genes in other species, InterPro protein domains, GO terms and KEGG pathway terms. The distinct gene names in Gossypium are stored in a database to form a community of its own. There can be one question that arises is how can we search for the data in such huge databases, the solution is by map marker and QTL data where all markers can be searched by marker source, map information or nearby loci. The various analysis tools [2] used are mainly online like NCBI's BLAST tool (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/wwwblast/) and a custom Batch BLAST tool where users can perform pair-wise BLAST alignments. CottonGen will be updated as new data become available and new or improved functionality is added to the site which can be done by the addition of GBrowse-syn, a GBrowse-based synteny browser to view multiple sequence alignment.

Diamondback moth genome database

The diamondback moth is one of the most destructive pests of cruciferous crops and costs the USA agriculture close to $2 billion. Henceforth a database exclusively enabling better division of these pesticides and helping in prevention includes the following features [3]-

- Co-ordinates genomic sequence of Fuzhou-S and related genomic and transcriptomic sequential data.
- provides a centralized database
- uses a simple and intuitive interface
- Details missing/misannoted genes.

1.2. Working

The DBM-DB is an extensive online database that catalogues DBM genomic data, it's user-friendly and web-based mode, operating by the analysis tools like Search, Overview, BLAST and GBrowse, which are interlinked with the Gene Information [3]. MySQL database language was used as a tool to manage and store the datasets of DBM-D. The gene location is linked to GBrowse, which provides gene structure visualization .Uniprot, GO, KEGG and InterPro databases accession numbers are also given for further assistance. The underlying point is how to search:-we can retrieve gene information of interested by inputting specific codes and/or keywords accordingly (Five types of search terms are available: Gene ID ranging from Px000001 to final OGSv1), annotation keyword, GO ID or term, KEGG ID or keyword and InterPro ID or domain name.

1.3. Architecture

The architectural description is as follows [4]:- DBM-DB was developed under the Linux system using several common software packages including PHP, Apache web server, MySQL database management and Perl Fast CGI .Also several custom PHP scripts were developed to make the database flexible, interactive and intuitive which enabled users to readily access and obtain the information they need either for molecular analysis or practical application. Additionally as mentioned earlier, the generic Genome Browser (GBrowse) package, a component of the Generic Model Organism Project (GMOD), was used for genome data visualization, which allows users to obtain the information on gene structures based on the DBM genome assembly.

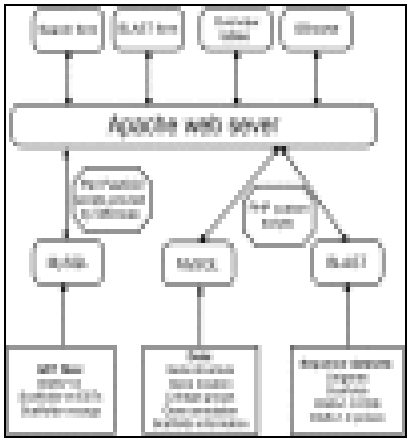

Figure 1: Illustrating Architecture of DBM-DB

2. Future Scope

DBM-DB provides a large-scale set of the genomic data and also provides convenient tool for further research on genomics, genetics and molecular biology of *P.xylostella* and other species of insects [3]. The main advantage of this is that this database was designed with the room to accommodate and house future data that will be generated, to regularly update and upgrade the data resources. However other add-ons are inclusion of digital gene expression profiling of different developmental stages or tissues, data supporting

microRNAs expression and the meta-genomics of DBM midguts. Further, it's been a while trying to integrate DBM sequences from NCBI database as well as DBM-related publications into DBM-DB. New web tools are being developed to allow more efficient and effective use of DBM-DB.

3. Conclusion

CottonGen is the base for the cotton community in terms of breeding of cotton. The major advantage being that it contributes to a comprehensive, integrated database which by Tripal genome database and Dribal, Chado results in storage of genomic and genetic data.

DBM-DB on the other hand provides large scale tool for further research on genomics, genetics, molecular biology of P.xylostella and other species of insects and in this way helps in the prevention of destruction of pests on crops.

Henceforth both these databases have been invaluable and sought after in their respective field's one in the rearing of cotton and another in the protection of crops against pests.

4. References

1. CottonGen: a genomics, genetics and breeding database for cotton research by Jing Yu, Sook Jung, Chun-Huai Cheng, Stephen P. Ficklin, Taein Lee, Ping Zheng, Don Jones , Richard G. Percy and Dorrie Main.(Nucleic Acids Research, 2014, Vol. 42, Database issue)
2. Brubaker,C.L. and Wendel,J.F. (1994) Reevaluating the origin of domesticated cotton (Gossypium hirsutum; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs).), 'Cotton: Origin, History, Technology and Production'. Wiley, New York, pp. 3–31.
3. Diamondback moth gene database by Weiqi Tang, Liying Yu, Guang Yang, Fushi Ke, Simon W. Baxter, Shijun You Carl J. Douglas and Minsheng You (NCBI, January.)
4. Talekar NS, Shelton AM. Biology, ecology, and management of the diamondback moth. Annu. Rev. Entomol. 1993; 38:275–301.y, 2014)
5. Figure (1) source- Diamondback moth gene database by Weiqi Tang, Liying Yu, Guang Yang, Fushi Ke, Simon W. Baxter, Shijun You Carl J. Douglas and Minsheng You (NCBI, January.)