International Journal for Computational Biology (IJCB)
Vol.3, No.2, August 2014, pp. 44-48
ISSN: 2278-8115

SMDB: Soybean Marker DataBase

Ashwani Kumar, Abhay Pratap, Urvashi, Rajinder Singh Chauhan, Tiratha Raj Singh*
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Waknaghat, Solan– 173234, H.P., India

Article Info

Article history:

Received Apr 24th, 2014
Revised Jul 11th, 2014
Accepted Jul 24th, 2014

Keyword:

Leguminous crop
Soybean
Transcription factor
Chaperons
Heat shock proteins.

ABSTRACT

Soybean Marker Database (SMDB) is a repository of important genomic information for soybean. At present several genomic databases are available for plants. Some of the important oilseeds plant databases are ATPID database, Castor Bean Genome Database, CGPDB, SoyBase, Legume Information System (LIS), Brassica database, Sinbase, etc. To gain comprehensive information from varied amount of resources, we developed this database which provides general as well as specific information at universal level. Along with this it also furnishes gene level information for various functional categories such as transcription factor, disease resistant varieties, heat shock protein, genetically modified strain of soybean. The bunch of information available to researchers today increases in tremendous manner. Hence understanding the plant genome specific databases for acquiring specific information is the demand of time for crop improvement and research programmes. SMDB is designed for the purpose of exploring potential gene differences in different plant genotypes, including genetically modified and disease resistant crops beneficial to the farmer who cultivate this crop. SMDB is publicly accessible for academic and research purpose at: http://www.bioinfoindia.org/smdb/.

Corresponding Author:

Tiratha Raj Singh
Department of Biotechnology and Bioinformatics, Jaypee University of InformationTechnology (JUIT), Waknaghat, Solan, India
Email: tiratharaj@gmail.com

How to Cite:

Ashwani kumar et. al. SMDB: Soybean Marker DataBase. IJCB. 2014; Volume 3 (Issue 2): Page 44-48.

1. INTRODUCTION

Soybean is important leguminous crop. Its seeds are rich in oil (approximately 20%) and protein (approximately 40%). In 2013, the global area, production and productivity of soybean were 90.1 million ha, 285.5 million ton and 2.44 t hac, respectively [1]. The USA, Brazil, Argentina, China and India are the major soybean-producing countries. Many biotic and abiotic stresses limit soybean production in different parts of the world. Vast research has been carried out worldwide for different aspects of soybean.The soybean cultivation first started in China. China was the world's largest soybean producer and exporter during the first half of the 20th century. Now, USA is the largest soybean-producing country in the world.

Soybean production in India is developing rapidly and the cultivated area of soybean is about the same as in China, but the yield per unit area is still relatively low [2]. The main reason for this is the scale at which the soybean cultivated by farmers is small and, therefore, advanced cultural practices needs to be adopted [3]. Along with economic developments and improvements in people's living standards, the demand for soybean in India is increasing rapidly and the domestic production of soybean cannot meet those demands.

Soybean is an dominant crop due to its nutritional value and wide utilization at household as well as industrial level. Soybean derived products are considerably cheaper than other protein-rich sources, such as fish, meat, milk and legumes. The cost of protein, when purchased as soybean, is only about 10-20% of the cost of protein from fish, meat, eggs or milk. Therefore, soybean is suitable to areas where other protein sources are unavailable or too expensive. Most of the essential amino acids are also found in soybean[4].

India has good potential for increased soybean production. The crop can be grown almost everywhere mostly in Madhya Pradesh, Uttar Pradesh, Maharashtra, Gujarat and some part of Himachal Pradesh. Soybean requires at least 500 mm of well distributed rainfall in three to four months and soil pH above 5[3]. The crop is unpopular because most of the people are ignorant. Our objective is to put efforts to provide important information to the biologists and scientific community so directly or indirectly it will help in increasing the soybean yield for farmers living in rural parts of India. Actual yield at present is about 5-10 q/ha. Several databases for soybean have been built and made publicly available, such as SoyGD[5], SoyBase[6] and SoyXpress[7]. These databases contain a variety of information, such as soybean genome sequences, bacterial artificial chromosome (BAC), expressed sequence tags (EST), and some useful tools including genome browsers, BLAST searching, and pathway searching. Opaque genomic information of soybean is there. To fill this gap, we designed SMDB database by using PHP and MySQL. This database not only contains most of the content and features already existed in earlier database of soybean but also contains some new content and features like general, national, international, statistical information along with rate of production in different part of world. SMDB database tries to fulfill the shortcomings found in earlier databases.

2. SURVEY OF OTHER SOYBEAN DATABASE

The goal of all databases is to provide comprehensive information at single window relevant to a species. The genomic information generated through wetlab experiment by researchers. Sequence information is generated by bioinformaticians which explore hidden information in sequences. We also surveyed few available database on soybean and their brief details along with their important features are available as follows.

2.1 Glycine max genome database(GmGDB)

The purpose of this resource is to provide a convenient sequence-centered genome view for Glycine max, with a narrow focus on gene structure annotation [8].

2.2 Soybean Knowledge Base (SoyKB)

It is an all-inclusive web resource for soybean translational genomics. SoyKB is designed to handle the management and integration of soybean genomics, transcriptomics, proteomics and metabolomics data along with annotation of gene function and biological pathway. It contains information on four entities, namely genes, microRNAs, metabolites and single nucleotide polymorphisms (SNPs) [9]. Other similar SNPs based plant specific databases provides good piece of information to the scientific community [10].

2.3 SoyPLEX

Plant-compliant and Plant Ontology enhanced expression resource for Soybean. PLEXdb (Plant Expression Database) is a unified gene expression resource for plants and plant pathogens. PLEXdb is a genotype to phenotype, hypothesis building information warehouse, leveraging highly parallel expression data with seamless portals to related genetic, physical, and pathway data [11].

2.4 Soybean Functional Genomics Database (SFGD)

The SFGD is a comprehensive database integrating genome and transcriptome data, and also for soybean acyl-lipid metabolism pathways. It provides useful toolboxes for biologists to improve the accuracy and robustness of soybean functional genomics analysis, further improving understanding of gene regulatory networks for effective crop improvement [12].

2.5 SoyBase

It is the USDA-ARS soybean genetic database which is a repository for professionally curated genetics, genomics and related data resources for soybean. SoyBase contains the most current genetic, physical and genomic sequence maps integrated with qualitative and quantitative traits. The quantitative trait loci (QTL) represent more than 18 years of QTL mapping of more than 90 unique traits. SoyBase also contains the well-annotated 'Williams 82' genomic sequence and associated data mining tools [13].

3. CONSTRUCTION AND CONTENTS

3.1 Database Overview

SMDB contain the annotation of genes and proteins for disease resistance, genetically modified, Heat Shock Proteins (HSP), yield and stress protein along with putative transcription factor. Users can access the main

components from the home page viz. general and specific information at gene level, protein browsing, family information, protein information, and FTP site. And also it is hyperlinked to other external public databases like NCBI, Agbioforum and National center for soybean biotechnology. It also provides feedback facility by which user can easily ask any query to the different people connected to this database [Fig. 1]. User can search the database either using Gene ID or by selecting any one of the 5 major class being defined for the genomic data such as disease resistance, genetically modified, HSP, yield and stress protein. One example search using the option disease resistance has been shown in Fig. 2.

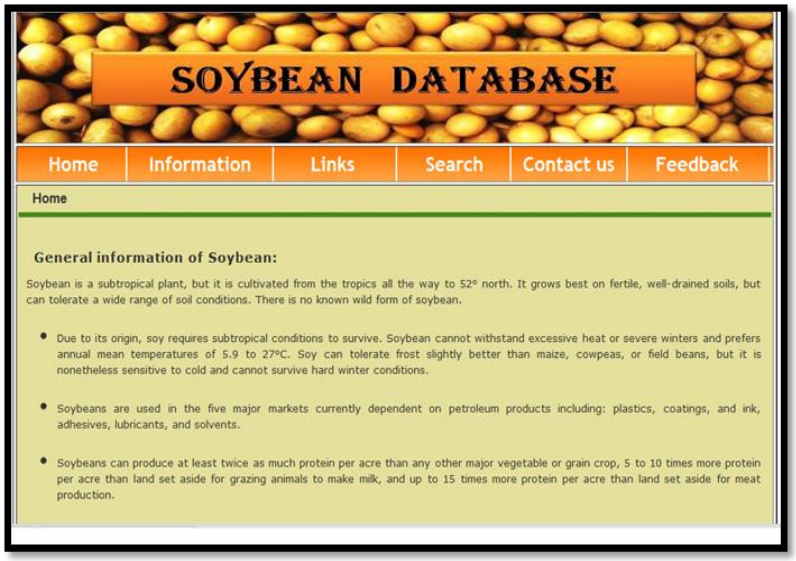


Figure1**:** Screenshot of SMDB database

3.2 Data Sources

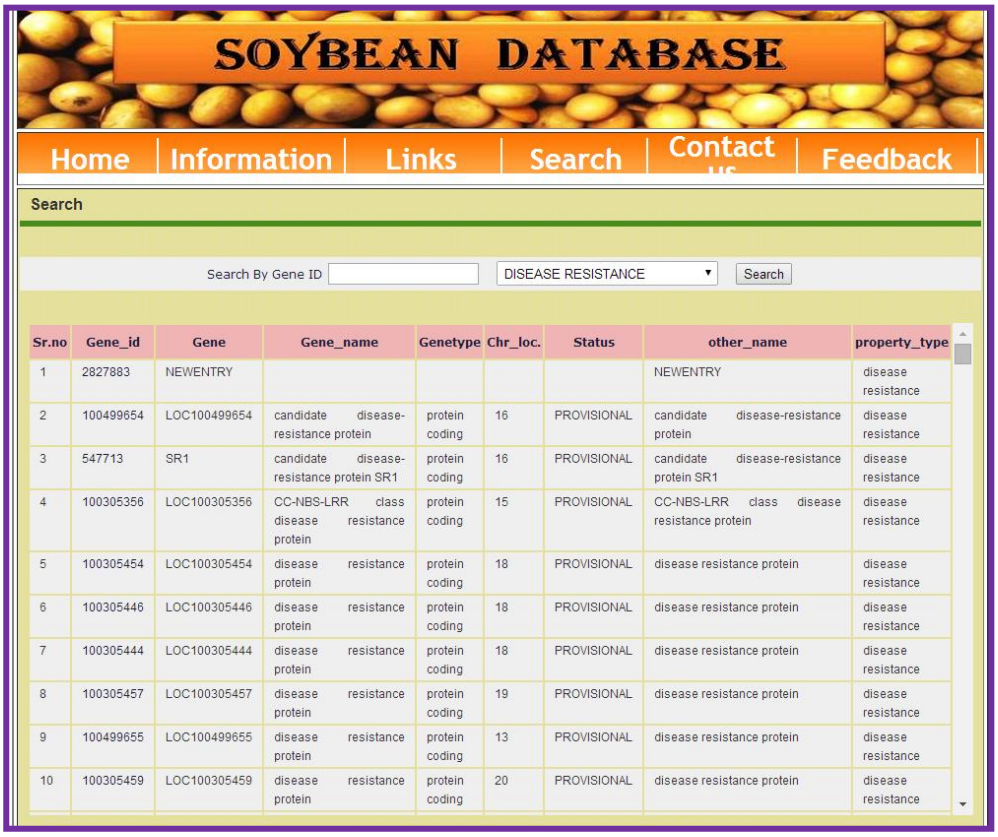
The soybean gene sequences information related to different properties used in this study were acquired from the publicly available database NCBI, biochemical pathways knowledge for role of key gene from KEGG and protein interaction information from STRING and BIND databases. For transcription factor identification, the standalone versions of InterProScan[14] is used to search all the soybean protein sequences against integrated in InerPro[15]. These databases and their corresponding scanning methods include: PROSITE (pfscan)[16], PRINTS (FingerPRINTScan)[17], Pfam (HMMPfam)[18], ProDom (ProDomBlast3i) [19] etc. InterProScan systematically searches each of these databases using their corresponding scanning methods to find domains. The proteins predicted to contain transcription factors (TF) related domain(s) were considered as putative transcription factors.

4. ANNOTATION USING BIOINFORMATICS TOOLS

The online versions of several bioinformatics tools were locally used to generate annotations related to information like disease resistant genes, transcription factors, genetically modified strain and high yield strains [Table 1]. The Protein Data Bank (PDB) was used to predict tertiary structure of each transcription factor and predicted structure is visualised through JMOL. MULTICOM generates the sequence alignments between the transcription factor and its homologous templates using PSI-BLAST.

4.1 Links to External Databases

External links greatly expand the annotation scope of SMDB providing related knowledge from various perspectives. SMDB provides a systematic view of a transcription factor – from the features of the protein itself, to the biological pathway it locates in. The external protein databases include industrial database like SOPA, Qualisoy, Indian Soybean Association and biotechnology based database like NCBI, Agbioforum and National Center for Soybean Biotechnology. The links are scheduled to be updated once every six months. This time interval can be changed if necessary.



Sr.no	Gene_id	Gene	Gene_name	Genetype	Chr_loc.	Status	other_name	property_type
1	2827883	NEWENTRY					NEWENTRY	disease resistance
2	100499654	LOC100499654	candidate disease-resistance protein	protein coding	16	PROVISIONAL	candidate disease-resistance protein	disease resistance
3	547713	SR1	candidate disease-resistance protein SR1	protein coding	16	PROVISIONAL	candidate disease-resistance protein SR1	disease resistance
4	100305356	LOC100305356	CC-NBS-LRR class disease resistance protein	protein coding	15	PROVISIONAL	CC-NBS-LRR class disease resistance protein	disease resistance
5	100305454	LOC100305454	disease resistance protein	protein coding	18	PROVISIONAL	disease resistance protein	disease resistance
6	100305446	LOC100305446	disease resistance protein	protein coding	18	PROVISIONAL	disease resistance protein	disease resistance
7	100305444	LOC100305444	disease resistance protein	protein coding	18	PROVISIONAL	disease resistance protein	disease resistance
8	100305457	LOC100305457	disease resistance protein	protein coding	19	PROVISIONAL	disease resistance protein	disease resistance
9	100499655	LOC100499655	disease resistance protein	protein coding	13	PROVISIONAL	disease resistance protein	disease resistance
10	100305459	LOC100305459	disease resistance protein	protein coding	20	PROVISIONAL	disease resistance protein	disease resistance

Figure2**:** Screenshot of SMDB database along with a search made through disease resistance option.

Table 1**:** Table illustrating the property type of soybean and their corresponding number of entries in SMDB database

Property Type	**Number of entries**
Disease Resistance	270
Genetically Modified	13
Heat Shock Protein	365
Yield Protein	39
Stress Protein	300
Total	987

4.2 Availability and Requirements

SMDB is freely available for academic and research purpose at http://www.bioinfoindia.org/smdb/. It is developed using PHP and MySQL. Based on assessment, SMDB is fully functional with three web browsers: Mozilla Firefox, Internet Explorer and Google chrome; and four operating systems: Windows XP, Windows Vista, Linux, and Mac OS. The only system requirement for SMDB is that JAVA runtime environment (JRE) which needs to be installed and set fully functional in order to make visualising and data search easier and faster.

5. CONCLUSION AND FUTURE PLAN

SMDB is a comprehensive database for soybean. It combines bioinformatics tools and various external databases to provide rich annotations, which can be browsed and retrieved through convenient web interfaces. The automated process generates annotations and creates database and website, and can be used to annotate other related species. At present, there is a limited number of entries for datatypes related to soybean but we will update our database in future by planning to link more soybean databases such as SoyBase, Soydb etc. with it and will add a human expert discussion section for each transcription factor where biologists can register, log in, and make comments on any annotation items. Also, we plan to connect the protein name, listed in each protein information page to its entry in Phytozome. By doing this, SMDB can be associated with other soybean genome annotations. Furthermore, we will identify the binding regions on the soybean DNA sequences, which can further help biologists to target the regulated regions on soybean genome.

REFERENCES

[1] Henkel J: Soy: health claims for soy protein, questions about other components. FDA consumer 2000, 34.
[2] Han B-Z, Rombouts FM, Nout MJR: A Chinese fermented soybean food. International Journal of Food Microbiology 2001, 65(1-2):1-10.
[3] Chang, R.Z. (1989) Studies on the origin of cultivated soybean. Oil Crop of China 1, 1–6.
[4] Ding, Y.L., Zhao, T.J. and Gai, J.Y. (2008). Genetic diversity and ecological differentiation of Chinese annual wild soybean (Glycine soja). Biodiversity Science 16, 133–142.
[5] FAO (2012) FAOSTAT. Food and Agriculture Organization of the United Nations, Rome, Italy. Available at: http://faostat.fao.org.
[6] Shultz J, Kurunam D, Shopinski K, Iqbal M, Kazi S, Zobrist K, Bashir R, Yaegashi S, Lavu N, Afzai A, et al: The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. Nucleic Acids Research 2006, 34 (D): D758-D765.
[7] SoyBase and the soybean breeder's toolbox. http://soybase.org/.
[8] Duvick J, Fu A, Muppirala U, et al. PlantGDB: a resource for comparative plant genomics. Nucleic Acids Research 2008, 36 (D): D959-D965.
[9] Joshi T, Fitzpatrick M.R, Chen S, Liu Y, et al. SoyKB: a web resource for integration of soybean translational genomics and molecular breeding. Nucleic Acids Research 2014, 42(D): D1245-D1252.
[10] Singh T.R, Gupta A, Seal A, Mahalaxmi M, Riju A. and Arunachalam V (2011). Computational identification and analysis of single-nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of Eucalyptus, Journal of Genetics 90, e34-38.
[11] Dash S, Hemert J.V, Hong L, Wise R.P and Dickerson JA. PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Research 2012, 40(D): D1194-D1201.
[12] Yu J, Zhang Z, Wei J,et a, SFGD: a comprehensive platform for mining functional information from soybean transcriptiome data and its use in identifying acyl-lipid metabolism pathways. BMC genomics 2014, 15:271.
[13] Grant D, Nelson R.T, Cannon S.B, Shoemaker R.C. SoyBase: the USDA-ARS Soybean genetics and genomics database. Nucleic Acids Research 2010, 38(D): D843-D846.
[14] Cheng K, Stromvik M: SoyXpress: a database for exploring the soybean transcriptome. BMC genomics 2008, 9(1):368.
[15] Zdobnov E, Apweiler R: InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001, 17(9):847-848.
[16] Mulder N, Apweiler R, Attwood T, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P: InterPro: An integrated documentation resource for protein families, domains and functional sites. Briefings in Bioinformatics 2002, 3(3):225-235.
[17] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux P, Pagni M, Sigrist C: The PROSITE database. Nucleic Acids Research 2006, , 34 Database: D227-D230.
[18] Attwood T, Croning M, Flower D, Lewis A, Mabey J, Scordis P, Selley J, Wright W: PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Research 2000, 28(1):225-227.
[19] Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E: The Pfam protein families database. Nucleic Acids Research 2004, 32(1):276-280.