

SolDB: A Database of *Solanum lycopersicum* and *Solanum tuberosum* Primers

Hassan Tariq^{1*}, Shahid Nadeem¹, Muhammad Sohail Raza¹,
Sobia Idrees¹, Aisha Yousaf¹, Muhammad Ibrahim Rajoka¹,
Bilal Hussain¹, Muhammad Nawaz¹, Tariq Niaz²

¹Department of Bioinformatics and Biotechnology
Government College University, Faisalabad, Pakistan
E-mail: hassantariq9@hotmail.com

²Directorate of Entomology
Ayub Agriculture Research Institute, Faisalabad, Pakistan

*Corresponding author

Received: October 13, 2010

Accepted: January 5, 2011

Published: January 31, 2011

Abstract: SolDB is the Database of Solanaceae Family. It is an interactive, free online specialized database for Solanaceae family. Currently, it spans complete nucleotide sequences of expressed genes of *Solanum lycopersicum* and *Solanum tuberosum* along with their annotation. We have designed PCR oligonucleotide primer sequences for each gene, with their features and conditions given. This feature alone greatly facilitates researchers in PCR amplification of genes sequences, especially in cloning experiments. We also provided chloroplast genome section which gives access to fully sequenced plastid genomes and their annotations. Flexible database design, easy expandability, and easy retrieval of information are the main features of SolDB. The Database is publicly available at www.solddb.pakbiz.org.

Keywords: *Solanum lycopersicum*, *Solanum tuberosum*, Database.

Introduction

Historically, databases have been arisen, to satisfy diverse needs, whether it address a biological question of an interest to an individual scientist, to better serve a particular section of biological community, to co-ordinate data from sequencing projects, or to facilitate drug discovery in pharmaceutical companies. According to Nucleic Acids Research annual database issues, in 2010 update, the online Database Collection that accompanies the issue holds 1230 data resources, a growth of 5% over last year [5].

Numerous bio-molecular data are available, but they are scattered in many databases and only some of them are curated by experts. Most available data are computationally derived and include errors and inconsistencies. Effective use of available data in order to derive new knowledge hence requires data integration and quality improvement [8].

The exponential growth of biological data over the past decade has created an enormous challenge to make effective use of the accumulated information. Correctly cataloging, labeling and connecting sequence, structural and functional information of genes and proteins of various trends and laws crucial to our understanding of life on earth as complex systems. Information stored must be correct, complete and internal relationships among elements easy to navigate. Computational tools and databases are essential to the management and identification of patterns among database elements that reflect biological systems [4].

The primary goal of the current project is to develop a specialized, minimally redundant, and curated nucleotide sequence database of *Solanaceae* family that strives to provide high-level annotations, including species based categorization of expressed genes, primers designed for the amplification of expressed genes and chloroplast genome section for species in *Solanaceae* plant family. Currently our focus was only on two important members of the *Solanaceae* family i.e. *Solanum lycopersicum* and *Solanum tuberosum*.

Materials and methods

Data collection

In the present study to develop the desired database, genes sequences that are expressed in the species of '*Solanaceae*' family (especially for tomato & potato) and their relevant annotations were required. To collect the data we searched for protein coding genes in NCBI's 'GenBank' and 'Gene' databases. We used nucleotide sequences in FASTA format and designed primers using Primer3. Then we analyzed the designed primers for primer-dimer formation and secondary structures using NetPrimer.

Database design

Family Section: In this section, we can add new families; edit the information about currently existing families and delete the currently existing family.

Species Section: In this section we can also add a new species, secondly we can edit the current species or other information and thirdly delete the existing species. Currently there are two species in focus: *Solanum lycopersicum* and *Solanum tuberosum*.

Genes Section: In this section we can add new genes, second, edit the current genes or other information related to it and third, delete existing genes.

Chloroplast genome Section: As the site in the eukaryotic cell where photosynthesis takes place, chloroplasts are responsible for much of the world's primary productivity, making chloroplasts essential to the lives of plants and animals alike. The oxygen in our atmosphere, all agricultural commodities and fossil fuels such as coal and oil are 'products' of photosynthesis [9].

Because of chloroplasts' huge importance, Chloroplast Genome section is designed by providing direct links to chloroplast genomes entries in GenBank database of species related to *solanaceae* family. All the collected represented links associated with annotations were aimed to develop a single site that represents all the *solanaceae* species' known chloroplast genomes.

Results

In the present study, we included complementary DNA nucleotide sequences for each gene. The primers designed for these complementary DNA sequences are really useful in their PCR amplification when they are cloned into some sort of vector. The current database also includes protein information of the relevant genes and their function. These features are the result of our flexible database design. Followings are the salient features of the SolDB:

1. Data searching

SolDB provides a very stylish way of searching the data. We can search our required information in two ways.

➤ Data searching by Search field

SolDB facilitates the users to search data by giving keyword related to function, protein, gene. If the record is found in the database then it will show all the results in all possible species.

➤ Data searching through Navigation

SolDB provides the facility for the users to search their relevant data by navigating the database. Whenever we click a family, a list of species will appear. From this list of species we can choose one specie and a list of genes present in it will appear, we can view its details by further navigating into it.

2. Easy and fast access to the information

We can get access to data in no time. Data searching is so easy in SolDB that even a new user can search through it with almost no difficulty.

3. Built in Primers

Primer designing has been the most distinguishing feature of this database. It is a new concept in database designing. It will help the scientist in PCR amplification of specific gene. Additionally, the conditions and features given pertaining to a particular primer also facilitate scientists to work effectively.

Discussion

As part of the present study, we have also developed a specialized database (SolDB) to store species based categorized expressed genes nucleotide sequences and their annotations related to species in Solanaceae plant family. Currently, it is focused on potato (*Solanum tuberosum*) and tomato (*Solanum lycopersicum*) of Solanaceae family.

The goal of biological databases is two fold: information retrieval and knowledge discovery [10]. Similarly, the chief objective in database development was to organize data in such a way in a set of structured records that user community can easily retrieve the desired information.

Keeping a good eye on the usage details of the database and the needs of the people using it is the only way to stay grounded [3]. In this project, special efforts are employed to get right details for effective database development, because designing, implementing and running databases are predominantly a series of decisions about intricate details [3].

The sequences in this database may overlap with the primary databases like GenBank [2], but it also has newly submitted data, which was obtained by submitting genes nucleotide sequences in online analysis programs, and then from the outputs of programs different kinds of new data was obtained. Thus, SolDB has its own unique organization and unique related annotations associated with the genes nucleotide sequences.

Although many issues in creating a good database may transcend biology and be valid for all domains, there are special circumstances around biological databases that make them worth treating as a special group (i.e. the free availability that they are on internet, that they keep up with rapidly growing field, and that they maintain high biological relevance) [1]. So like other

plant databases SolDB is also free online database. It can be accessed through easy-to-use web interface. All the data in the database is freely available with no restrictions. Its data can be used in wide range of applications and scenarios by users ranging from laboratory scientists to experienced bioinformaticians.

Keeping in view the fact that the manual selection of optimal PCR oligonucleotide sets can be quite tedious and thus lends itself very naturally to computer analysis [7], SolDB also contained PCR oligonucleotide primer sequences for nucleotide sequences of genes. These primers' design is aimed at obtaining a balance between two goals: specificity and efficiency of amplification. In primer designing, this balance is obtained by analyzing the quality of primers with various programs, considering specially avoiding primer-dimer formation and secondary structure in primers.

Organelle genomes may be ideal proving ground for methods of analysis being developed to understand genome and gene order evolution [6]. With the mission to further progress comparative analysis of plastid/chloroplast genomes by explaining the community need for better uniform annotations, Chloroplast Genome Section is incorporated in SolDB website for species in Solanaceae family.

In conclusion, SolDB is currently covering genes of two major species of plant family Solanaceae i.e. potato and tomato. It provides easy methods to update and expand the database by submitting more plant genes nucleotide sequences together with their associated data. In future, as the SOL project (International Solanaceae Genome Project) will provide more euchromatic sequencing data and as the gene characterization in plants of Solanaceae increased, SolDB will expand and become more useful in specialized information retrieval.

Acknowledgement

The authors are thankful to all of the researchers of Database and Software Engineering Research Group of the Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan.

References

1. Altman R. B. (2004). Building Successful Biological Databases, Brief Bioinformatics, 5(1), 4-5.
2. Benson D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers (2009). GenBank, Nucleic Acids Res, 38, D14-D15.
3. Birney E., M. Clamp (2004). Biological Database Design and Implementation, Brief Bioinformatics, 5(1), 31-38.
4. Buehler L. K., H. H. Rashidi (2005). Bioinformatics Basics: Applications in Biological Science and Medicine, CRC Press, USA, 166.
5. Cochrane G. R., M. Y. Galperin (2010). The 2010 Nucleic Acids Research Database Issue and Online Database Collection: A Community of Data Resources, Nucleic Acids Res, 38, D1-D4.
6. Cui L., N. Veeraraghavan, A. Richter, K. Wall, R. K. Jansen, J. Leebens-Mack, I. Makalowska, C. W. Depamphilis (2006). ChloroplastDB: The Chloroplast Genome Database, Nucleic Acids Res, 34, D692-696.
7. Dieffenbach C. W., T. M. Lowe, G. S. Dveksler (1993). General Concepts for PCR Primer Design, PCR Method Appl, 3(3), S30-S37.

8. Ghisalberti G., M. Masseroli, L. Tettamanti (2010). Quality Controls in Integrative Approaches to Detect Errors and Inconsistencies in Biological Databases, *J Integr Bioinform*, 7(3), 2010-2119.
9. Halliwell B. (1978). The Chloroplast at Work. The Review for Modern Developments in our Understanding of Chloroplast Metabolism, *Prog Biophys Mol Biol*, 33, 1-54.
10. Xiong J. (2006). *Essential Bioinformatics*, Cambridge University Press, New York, USA, 10-27.

Hassan Tariq, M.Sc. in Bioinformatics

E-mail: hassantariq9@yahoo.com



Hassan Tariq graduated from the Government College University – Faisalabad in 2008 with the distinction of securing second highest percentage in the session. Since 2009 he is working as a Research Officer in the Department of Bioinformatics and Biotechnology. He is the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Bioinformatics Software Development, also publishing textbooks. Scientific interests: Biological Databases, Data Mining, Software Engineering in Bioinformatics, Web Engineering.

Shahid Nadeem, Ph.D.

E-mail: snadeem63@yahoo.com



Dr. Shahid Nadeem got his Ph.D. from the Punjab University, Lahore in 2002. He is working as a senior scientist in NIAB, Faisalabad. In 2009 he joined the Department of Bioinformatics and Biotechnology as a Chairman. He is among the founders of Department of Bioinformatics and Biotechnology in GC University Faisalabad. He is also the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Biotechnology and its applications, Social, Ethical aspects of Biotechnology also publishing textbooks. Scientific interests: Biotechnology, Plant breeding and Genetics, Microbiology.

Muhammad Sohail RazaE-mail: muhammadsohailraza@live.com

Muhammad Sohail Raza is a dedicated member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. He graduated from Government College University in 2010 and has the distinction of being the student with second highest percentage in the session. In Research group, he works as a database Developer and is working as an active participant in ongoing research projects. He has command over many Bioinformatics tools and software. He also works on various programming languages like java, Perl, BioPerl and Python. His core interests are Biological Database development, Bioinformatics software Engineering, Data mining, Drug discovery and Computer Aided Drug Designing.

Sobia IdreesE-mail: sobia_binm@live.com

Sobia Idrees graduated from the Government College University Faisalabad in 2010. She is working as a dedicated member of Database and Software Engineering Research Group at Government College University Faisalabad. She holds the position of an enthusiastic software engineer in research group and working on many ongoing research projects held by the Group. Based on her research project she obtained 1st position in the software competition at national level. She has worked as active student member of IEEE-GCUF for the year 2008 and organized and participated in many events. She has interests in Software Engineering, Web and Database Development, Bioinformatics Sequence Analysis and in various programming languages like C++, VB.Net, Java, BioJava, Perl, Bioperl and PHP.

Aisha YousafE-mail: aishayousaf@live.com

Aisha Yousaf is an active member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. She graduated from the Government College University Faisalabad in 2010. In research group she is working on ongoing research projects. She has command over many Bioinformatics data analysis and structure prediction tools. Her interests are software development, drug designing, phylogenetic analysis and protein structure prediction.

Muhammad Ibrahim Rajoka, Ph.D. in BiochemistryE-mail: mibrahimrajoka47@gmail.com

M. I. Rajoka graduated from the University of New South Wales – Australia in 1981 with the distinction of securing second highest percentage in the session. He did Ph.D. in Biochemistry from University of the Punjab in 1990. After serving 41 years in Nuclear Institute for Agriculture and Biology – Faisalabad (19 years) and National Institute for Biotechnology and Genetic Engineering – Faisalabad (22 years), he joined as Professor in the Department of Bioinformatics and Biotechnology. He has supervised 25, 21 and 7 M.Sc., M.Phil. and Ph.D. students respectively and has published 140 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. He is also an active member of teaching Faculty of the department, teaching bioprocess technology, research methods in biological sciences, technical writing, communication skills and guiding faculty in writing research proposals. Scientific interests: Industrial strain development through conventional breeding and rDNA technology, bioprocess engineering.

Bilal Hussain, M.Phil., Ph.D. StudentE-mail: profbilal@yahoo.com

Bilal Hussain is a Ph.D. Scholar from the Government College University – Faisalabad with the distinction Bronze Medal in the session. Since 2003 he is working as a Lecturer in the Department of Bioinformatics and Biotechnology. He is among the founder of Department of Bioinformatics and Biotechnology in GC University Faisalabad. He has specialization in the field of Molecular Biology and fisheries. He is performing many administrative duties in the department of Bioinformatics and Biotechnology and is an active member of research group of the department with many publications. He is working on molecular study of genetic disorders.

Muhammad Nawaz M.Phil., Ph.D. StudentE-mail: m.nawaz9@gmail.com

Muhammad Nawaz did his master in Biology from university of Arid Agriculture Rawalpindi, Pakistan and M.Phil. in plant biotechnology from National Institute for Biotechnology and Genetic Engineering Faisalabad, Pakistan affiliated with Quaid-i-Azam university Islamabad, Pakistan. Since 2007 he is a working as lecturer in the Department of Bioinformatics and Biotechnology GC University Faisalabad. He is the founder of Plant Molecular Biology and Biotechnology in the Department of Bioinformatics and Biotechnology. He is also an active member of Pakistan society for Biochemistry and Molecular Biology. Currently he is working in the area of gene characterization and transformation in sugarcane for better attributes. Scientific interests: Plant Molecular Biology and Biotechnology. Plant Genomics and Bioinformatics.

Tariq Niaz, M.Sc. in EntomologyE-mail: tniaz08@hotmail.com

Tariq Niaz graduated from the University of Agriculture – Faisalabad in 1979 with the distinction of securing third highest percentage in the session. He is serving Ayub Agriculture Research Institute, Faisalabad since 1980 as an entomologist. He has also supervised several thesis of MSc and M.Phil student. He has published 40 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. Scientific interests: Plant protection, Bioassay of pesticides, Biological databases, Bioinformatics, Software development, Web engineering.