Informatics in Medicine Unlocked 7 (2017) 34–38

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Internal transcribed spacer sequence database of plant fungal pathogens: PFP-ITSS database

Aman Chandra Kaushik[a,g], Anirudh Pal[a], Akash Kumar[c], Vivek Dhar Dwivedi[c,e,f], Shiv Bharadwaj[b], Amit Pandey[d], Sarad Kumar Mishra[e], Shakti Sahi[a,*]

[a] School of Biotechnology, Gautam Buddha University, Greater Noida 201308, India
[b] Nanotechnology Research and Application Center, Sabanci University, Istanbul, Turkey
[c] Department of Biotechnology and Bioinformatics, Uttaranchal College of Science & Technology, Dehradun, India
[d] Forest Pathology Division, Forest Research Institute-Dehradun, India
[e] Department of Biotechnology, Deen Dayal Upadhyay Gorakhpur University, Gorakhpur, India
[f] Mahatma Ghandhi Chitrakoot Rural University Chitrakoot Satna, India
[g] The Shraga Segal Dept. of Microbiology, Immunology and Genetics, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

ARTICLE INFO

Keywords:
ITS
Wrapper
Integrity
Mining
BLAST
Primer

ABSTRACT

The nurseries and plantations of medicinally and economically important plants are facing challenge on health front as they are being rapidly exposed to fungal pathogens thereby affecting their health and productivity adversely. The plant pathogenic fungi significantly damage and reduce the flora of subcontinent. Lack of knowledge and trouble in identification of fungal pathogens, are imposing a peril to both the flora and human health all over the world. Albeit routine pathological techniques can be used for the identification of plant fungal pathogens but these methods are time consuming and often need sound knowledge in mycotaxonomy. Recently, molecular (DNA sequence) data has emerged with the nuclear ribosomal internal transcribed spacer (ITS) region (DNA barcoding) as crucial biomarker to disclose the necessary information for the taxonomic identification of plant pathogenic fungi. However, development of nucleotide sequence database of plant pathogenic fungi will enable authentic pathogen identification easy and quickly even by a person not trained in fungal taxonomy. This study presents the development of a new plant fungal pathogen -Internal Transcribed Spacer Sequence database (PFP- ITSS Database) holding 1215 ITS sequences collected from various sources. It represents 1215 plant fungal pathogens (PFP) in relation to their respective economical and medicinal plants; and is available at http://shaktisahislab.com/include/ITSdb. PFP-ITSS Database will provide useful information for better understanding of plant pathogenic fungi, which cause disease in both economical and medicinal plants.

1. Introduction

Pathogens are a group of species that infect and disrupt the normal physiology of the hosts to complete their life cycle. Plant pathogens include fungi, nematodes, bacteria, and viruses which can cause diseases or damages in the plants [1]. Amongst these pathogens, fungi are known to cause maximum yield loss in numerous economically important crops [2]. In developing country like India, the yield of crop production is very low due to infection by pathogenic fungi [3]. Additionally, recent reports on emerging plant fungal pathogens and their cross-kingdom infections to animals and immune compromised persons emphasize the need for correct and quick identification [4,5]. However, the identification of the PFP using the traditional taxonomy methods at species level are complicated due to lack of an adequate information on morphological characteristics or different phases of the life cycle [6,7]. Also, it has been reported that different PFP species have evolved a particular mode to infect and cause diseases in the plants [8]. There is a need for a novel, fast and accurate technique for identification of the PFP at species or strain level in the environment to carry out disease surveillance and implementation of a disease management strategy. Recently, the genomic data of the organisms have been widely explored to collect and scrutinize the relevant detailed information such as identification of PFP irrespective of their morphological characteristics, degree of cultivability and different phases of life cycle [9]. These studies have contributed towards a huge collection of taxonomically and technically conciliated DNA sequences, further

* Corresponding author.
E-mail address: shaktis@gbu.ac.in (S. Sahi).

http://dx.doi.org/10.1016/j.imu.2017.02.006
Received 24 December 2016; Received in revised form 17 February 2017; Accepted 22 February 2017
Available online 24 February 2017

pooled and shared by the international nucleotide sequence databases. Albeit the availability of these molecular databases can be used as reference data for the identification of unknown species; but this type of identification raises the problems, because novel sequence data produced may not be in a locus to evaluate whether a suggested taxonomic affiliation is reliable. For instance, recent studies reveal that the reliability of the produced molecular data in certain groups of organisms have been severely compromised without the adequate means to differentiate a significant data from the insignificant. As a consequence, these errors and inconsistencies are regularly included into the data and used by the research communities over the time. It eventually results in the misidentification of the species names and ecological properties from sequence resemblance searches.

Molecular identification through DNA sequences have promptly advanced the understanding of species frontiers and interactions in several important plant pathogenic genera, revealing numerous enigmatic species. In this concern, fungus identification usually relies on the sequencing of the nuclear ribosomal internal transcribed spacer (ITS) region and has been proposed as the formal universal fungal barcode. The largest database designed for fungal ITS sequences is UNITE. UNITE mirrors and curates the International Nucleotide Sequence Database Collaboration INSDC: GenBank, ENA, and DDBJ for fungal ITS sequences and deals with broad propensities for the sequence analysis and third-party annotation of sequences to its users. Though the molecular databases have been gradually cited and recorded with fungal taxon in dated publications, but as our familiarity and survey of disease associated with fungi increases, the checklists and databases have become more and more inaccurate. For example, [10] studies concluded that the frequent occurrence of the fruit rot disease in tropics is not caused by *Colletotrichum gloeosporioides*. It is, therefore, required to group these pathogens in Plant-pathogen species complexes based on the modern molecular data protocols to present them as reference databases for the identification of unknown PFP. We report here the collection and identification of plant pathogens associated with medicinal and economical plants. The present database readily contains the large sets of plant pathogenic fungi information and internal transcribed spacer sequences. As fungal internal transcribed spacer region of ribosomal RNA gene sequence are available in plant, it will be updated into the database regularly. The output results are highly valued and easy to process, although it still needs to be updated regularly. The database is the first platform concerning fungal pathogen of economic and medicinal plant. It will assist users in related fields by providing comprehensive information (fungus name to reference column) on database. The database resource will be freely available for use in public.

2. Methodology

2.1. ITS sequence data collection

Total 1215 barcoding sequences (18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28 S ribosomal RNA gene, partial sequence) of 1215 species were retrieved as shown in Supplementary A. The PFP-ITSS data were retrieved from various sources.

3. Development of PFP- ITSS database

3.1. ITS database architecture and web user interface

This Database was assembled and configured upon platform JavaScript, AJAX, MySQL, PHP and HTML. Java script language was used for flash designing as well as control development. The ITS Database architecture and web user interface cycle is shown in Fig. 1.

Java script was mainly implemented as part of a web browser in order to create enhanced user interface and dynamic website. AJAX is a technique for creating fast and dynamic web pages. Since, AJAX allows web pages to be updated asynchronously by exchanging small amounts of data with the server at the back end, therefore, we employed AJAX for building the database web pages. This means that it is possible to update parts of a web page, without reloading the whole page. MySQL was used for back end designing as well as ITS Database deposition. For designing of the database, PHP- Hypertext Preprocessor v5.5.0 Alpha1 version was used to perform the control designing and back end connectivity. Additionally, Cascading Style Sheets (CSS) were used for front end designing of ITS Database. CSS is a simple mechanism for adding style like fonts, colors, spacing to Web documents.

3.2. Approaches

3.2.1. Wrapper

The interesting feature of ITS Database is usage of wrappers; which explicitly optimizes the performance criterion and verifies database by comparing the results with FILTER approach FCBF method, where if the data about the Fungus name, Region, Fungus Features, Affected part of the Host, Name of the Host Plant, Disease, Impact on Plant, ITS Sequence are not available in ITS Database then it will search the data in other similar databases using wrapper method and it will search out complete query unless and until the user is not satisfied.

3.3. Data integrity

Data integrity is the best features of ITS Database. It has been incorporated to make the ITS database more accurate and consistent, though data warehousing require this accuracy to stand against the errors occurring either due to human error, software or hardware. Federator Data warehouse approach has been highly envisaged in inter operating ability of the medical care and its research.

3.4. Data mining and data warehousing

Enterprise data warehouse was used for extracting the ITS data into data mart as it provided the clean data. The issues pertaining to data consolidation were addressed during maintenance procedure. ITS data was retrieved from the various databases.

3.5. Normalization

Normalization was done to validate the data. The data was arranged data in the database by creating tables and establishing relationships between tables thereby eliminating redundancy and inconsistent dependency. The data deposition was done using natural language processing (NLP) methods. ITS generated output can identify and classify different values like fungus name, region, fungus features, affected parts of the host plant, name of host plant, disease, impact on plant, ITS sequence and references of the sequence available in the literature.

3.6. Development of homology searching tool

Homology searching is a method used to find out the identical or similar homologs of an organism which is useful in the identification of particular species at molecular level. The homology search algorithms [11,12] were used for designing homology searching tool for PFP- ITSS Database.

3.7. Development of primer designing tool

ITS PRIMER Tool, a new primer designing program for finding primers from ITS sequences was developed. This tool is a very powerful and efficient PCR primer design program, which allows the user to have