

## Original article

# DFVF: database of fungal virulence factors

Tao Lu, Bo Yao and Chi Zhang\*

School of Biological Sciences, Center for Plant Science and Innovation, University of Nebraska, Lincoln, NE 68588, USA

\*Corresponding author: Tel: +1 402 472 1319; Fax: +1 402 472 8722; Email: czhang5@unl.edu

Submitted 15 May 2012; Revised 13 August 2012; Accepted 14 August 2012

Fungal pathogens cause various diseases for plant and animal hosts. Despite the extensive impact of fungi on human health and life, the threats posed by emerging fungal pathogens are poorly understood. Specifically, there exist few fungal virulence gene databases, which prevent effective bioinformatics studies on fungal pathogens. Therefore, we constructed a comprehensive online database of known fungal virulence factors, which collected 2058 pathogenic genes produced by 228 fungal strains from 85 genera. This database creates a pivotal platform capable of stimulating and facilitating further bench studies on fungal pathogens.

**Database URL:** <http://sysbio.unl.edu/DFVF/>

## Introduction

Fungal organisms comprise one of the most diverse kingdoms on Earth; the number of fungal species is estimated to be ~1.5 million (1). Many fungal taxa are pathogens under certain conditions. For instance, ~80 000 fungal taxa have pathogenicity on 56 000 vascular plant hosts, according to the Fungus–Host Distributions of US Department of Agriculture (<http://nt.ars-grin.gov>). Fungal pathogens have a broad spectrum of hosts, including plants and animals, and may cause death and disability in humans, yield loss in agricultural crops and even alteration of forest ecosystem dynamics. For example, fungi in the genus *Fusarium* cause a variety of blights, seedling disease, root rots or wilts on nearly all species of cultivated plants, and they strongly affect the agriculture economy. Virulence factors are the most important proteins in pathogens, such as toxin synthetic enzymes and secreted biodegradation enzymes (2), that permit them to evade the defense mechanisms of the host and, thus, cause diseases (3). Currently, the number of reported fungal virulence factors is limited, and many of the data are only available in the literature texts. Please refer to references (4–8) for reviews on fungal virulence factors. To maximize the value of this type of data, it is essential that a method of data storage and sharing should be implemented for efficient and effective data

mining. To this end, we believe that the utilization of a web-based comprehensive database will significantly facilitate such activities. Currently, there exists only one database, PHI-base (9), for fungal pathogens. Although PHI-base collected pathogenic genes for all types of fungal and bacterial pathogens, it contained limited number of fungal virulence factors, and most of its records are about bacterial pathogens. In comparison, advanced databases and prediction tools similar to what we are proposing have been developed for many other bacterial pathogens and were proven useful (10–13). Therefore, we constructed a comprehensive online database of fungal virulence factors for public use to fulfil this important need.

## Database generation and its content

The state-of-the-art text-mining technique is used by the PubMed database and the Internet by searching keywords, such as fungal virulence factors, pathogenic genes and so forth. An in-house tool, programed in Python, is used to fetch article titles and abstracts from the PubMed database, and algorithms used by MedTAKMI (14) are implemented locally for entity extraction. Human intelligence is also involved to screen the output of automatic literature

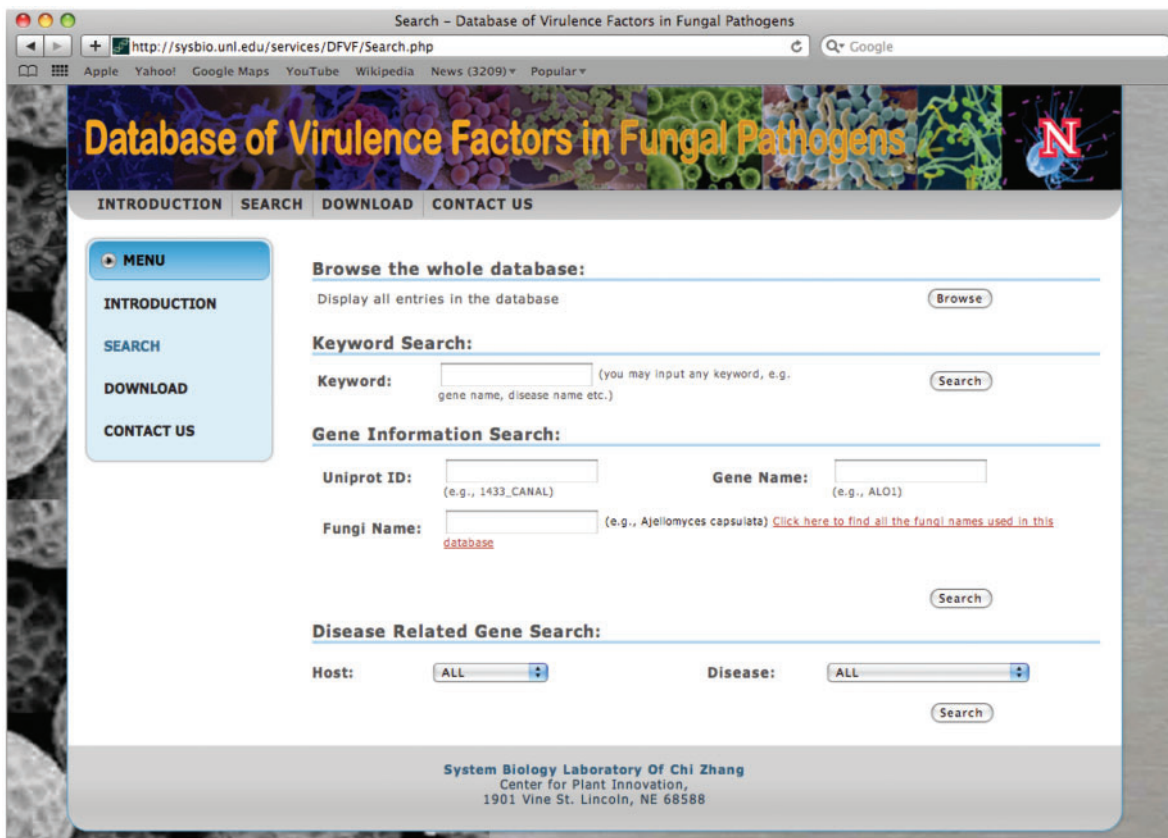
mining methods. Virulence factors are protein products of virulence genes, which are helpful for induction and development of disease (7, 15), but this database focuses on genes and their protein products. Some fungi exist in normal human body flora, such as *Saccharomyces cerevisiae*, and are normally non-pathogenic. However, they could also cause life-threatening infections, which often occur in immunocompromised patients or vulnerable population with weakened immune systems (16, 17). Therefore, virulence factors from this kind of fungi are also included in the database. As a result, 2058 fungal virulence factors are collected, which belong to 85 fungal genera, 228 fungal strains (by NCBI taxonomy ID). These virulence factors come from 593 peer-reviewed journal articles and sequences submitted to GenBank or UniProt databases. Although they are taxonomically different from fungi, oomycetes, as originally being classified among the fungi (18–20), are included this database, and only 79 records come from oomycetes. Of the 2058 proteins, 320 virulence factors are predicted to be secreted by fungi using WoLFPSORT (21) and signalp (22), whereas ~30 proteins are related to biosynthetic fungal toxin, such as mycotoxin. As a comparison, there are 600 fungal pathogenic genes in the PHI-database (9).

The related information of all virulence factors, such as gene symbol, NCBI database IDs, taxonomy and the protein sequence, is collected. In database of fungal virulence factors (DFVF), the fungus strain taxonomy is recorded with its original reference, as this method may help investigators find the references. In addition, the Pfam (23) domain annotation of each virulence factor and their Gene Ontology annotation are also recorded. The phenotypic information,

**Table 1.** Statistics of virulence factors as their hosts

Host	Genus	Species	Factors
Animal			
All <sup>a</sup>	2	25	451
Vertebrata	36	71	1308
Invertebrata	5	8	45
Plant			
All <sup>a</sup>	23	90	346
Herb	20	65	539
Xyloid	17	67	261

<sup>a</sup>All means that the hosts of the virulence factors have a broad spectrum and cover many different kinds of hosts.



**Figure 1.** The searching page of the database.

such as disease information and host, is also collected. In DFVF, all plant diseases and host descriptions are from the US Department of Agriculture database records. Table 1 shows the statistics of hosts and their pathogenic fungi as well as virulence factors in DFVF. For example, there are 1308 virulence factors from 71 pathogens, whose hosts are all vertebrata animals, whereas 539 factors from 65 fungal pathogens have hosts of herbal plants.

conveniently locate the genes of interest. If a user knows the UniProt ID, gene name or a fungus name, the gene information search can be directly applied. If a user wants to get all virulence factors related to a certain disease or host, a disease related gene search can also be conducted. The most convenient searching approach is the keyword search in which the database will return all virulence factors for which the keyword is contained in any piece of the information under the factors.

## USER interface

### Search

The database system provides interactive access to all of the collected data, and users may connect to the database using a web browser. Figure 1 shows a snapshot of the user interface for users to browse or search the database. The 'Browse' button allows users to get a list of all records in one table. Variable search options are provided to

## Results

Once a user browses the whole database or searches with a specific option, the database first returns a table of related records, and then displays the UniProt ID, gene information and disease information in three columns, shown in Figure 2. The details of each factor can be displayed by clicking on the link of UniProt ID. The information for

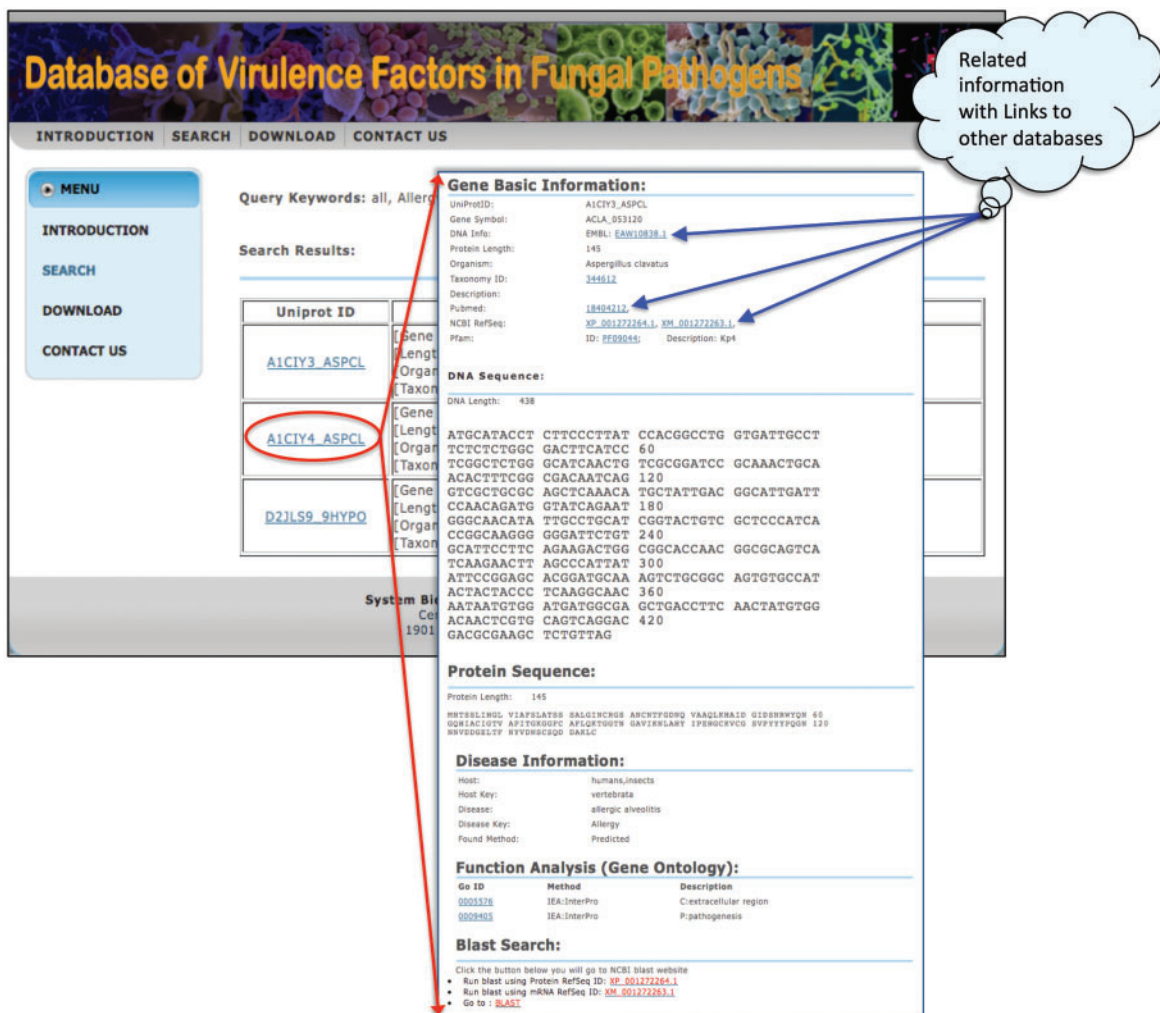


Figure 2. The display page of searching result and information of each gene.

each virulence factor includes the basic information, DNA/protein sequences, disease information and **Gene Ontology** annotation. The basic information consists of UniProt ID, gene symbol, taxonomy, ID and links to other databases and description. The PubMed links to all related publications are also presented. If the protein of a virulence factor has one or more Pfam domains, the links to those domains are provided. Descriptions on phenotypic data, including diseases and hosts are shown in the section of 'Disease information'.

## Implementation

We adopted the LAMP (Linux, Apache, MySQL, PHP) platform to construct the online database system. The user interface has been designed using the JavaScript application framework. The user interface additionally accepts parameters through a URL for direct searching. This feature facilitates a link to the database from external sites, and it also allows users to bookmark and to cite specific results.

## Accessibility

The database is freely available to all users without restriction at <http://sysbio.unl.edu/DFVF>. All data are downloadable from the same website. In addition to the link to download the whole database, we provide various links to download data in different categories for convenience. The source codes and other detailed information are available on request.

## Acknowledgements

T.L. initialized this project and curated all data. B.Y. constructed the database and web pages. C.Z. supervised this project and drafted the manuscript.

## Funding

This work is supported by the University of Nebraska—Lincoln start-up funds (to C.Z.) and the Nebraska Soybean Board Funds (to C.Z.). Funding for open access charge: University of Nebraska - Lincoln start-up funds.

*Conflict of interest.* None declared.

## References

- David, L.H. (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycol. Res.*, **105**, 1422–1432.
- Yike, I. (2011) Fungal Proteases and their pathophysiological effects. *Mycopathologia*, **171**, 299–323.
- Bhatnagar, D., Yu, J. and Ehrlich, K.C. (2002) Toxins of filamentous fungi. *Chem. Immunol.*, **81**, 167–206.

- Stergiopoulos, I. and de Wit, P.J. (2009) Fungal effector proteins. *Annu. Rev. Phytopathol.*, **47**, 233–263.
- De Wit, P.J., Mehrabi, R., Van den Burg, H.A. et al. (2009) Fungal effector proteins: past, present and future. *Mol. Plant Pathol.*, **10**, 735–747.
- van de Wouw, A.P. and Howlett, B.J. (2011) Fungal pathogenicity genes in the age of 'omics'. *Mol. Plant Pathol.*, **12**, 507–514.
- Idnurm, A. and Howlett, B.J. (2001) Pathogenicity genes of phytopathogenic fungi. *Mol. Plant Pathol.*, **2**, 241–255.
- Gonzalez-Fernandez, R. and Jorin-Novo, J.V. (2012) Contribution of proteomics to the study of plant pathogenic fungi. *J. Proteome Res.*, **11**, 3–16.
- Winnenburg, R., Baldwin, T.K., Urban, M. et al. (2006) PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res.*, **34**, D459–D464.
- Winsor, G.L., Lam, D.K., Fleming, L. et al. (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.*, **39**, D596–D600.
- Zhu, X., Chang, S., Fang, K. et al. (2009) MyBASE: a database for genome polymorphism and gene function studies of Mycobacterium. *BMC Microbiol.*, **9**, 40.
- Yang, J., Chen, L., Sun, L. et al. (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.*, **36**, D539–D542.
- Gillespie, J.J., Wattam, A.R., Cammer, S.A. et al. (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.*, **79**, 4286–4298.
- Uramoto, N., Matsuzawa, H., Nagano, T. et al. (2004) A text-mining system for knowledge discovery from Biomedical Documents. *IBM Syst. J.*, **43**, 516–533.
- Gijzen, M. and Nurnberger, T. (2006) Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. *Phytochemistry*, **67**, 1800–1807.
- Munoz, P., Bouza, E., Cuenca-Estrella, M. et al. (2005) Saccharomyces cerevisiae fungemia: an emerging infectious disease. *Clin. Infect. Dis.*, **40**, 1625–1634.
- Barta, Z., Zold, E., Nagy, A. et al. (2011) Celiac disease and microscopic colitis: a report of 4 cases. *World J. Gastroenterol.*, **17**, 2150–2154.
- Sleigh, M.A. (1989) *Protozoa and Other Protists*, 2nd edn. E. Arnold, London, New York.
- Margulis, L. (1990) *Handbook of Protozoa: The Structure, Cultivation, Habitats, and Life Histories of the Eukaryotic Microorganisms and Their Descendants Exclusive of Animals, Plants, and Fungi: A Guide to the Algae, Ciliates, Foraminifera, Sporozoa, Water Molds, Slime Molds, and the Other Protozoists*. Jones and Bartlett Publishers, Boston.
- Willey, J.M., Sherwood, L., Woolverton, C.J. et al. (2008) *Prescott, Harley, and Klein's Microbiology*, 7th edn. McGraw-Hill Higher Education, New York.
- Horton, P., Park, K.J., Obayashi, T. et al. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
- Petersen, T.N., Brunak, S., von Heijne, G. et al. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
- Finn, R.D., Mistry, J., Tate, J. et al. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.