Int J Pharm Bio Sci 2015 July; 6(3): (B) 724 - 731

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

EVIRDB – A COMPREHENSIVE VIRULENCE FACTORS DATABASE OF ENTEROBACTERIACEAE

THIRUNAVUKKARASU MURUGESAN[1,2], DINAKARAN KARUNAKARAN[2,3] AND *GNANENDRA SHANMUGAM[4]

[1] Department of Computer Science, Mahendra Arts & Science College, Kalippatti, Namakkal, TamilNadu, India.
[2] Research and Development Center, Bharathiar University, Coimbatore, TamilNadu, India.
[3] Department of Computer Science and Engineering, RMD Engineering College, Kavaraipettai, Tiruvallur, TamilNadu, India.
[4] Department of Biotechnology, Mahendra Arts & Science College, Kalippatti, Namakkal, TamilNadu, India.

ABSTRACT

The continuous emergence of multidrug resistant bacteria has become a greater concern for the public health sectors. The knowledge on the mechanism of bacterial pathogenesis, surveillance in the host system and its resistance to antibiotics has attained a greater attention of researchers to understand the critical virulence factors of bacterial pathogens. The identification of novel virulence factors that can prioritize its importance as a drug target is most promising challenge of many pharmaceuticals. Thus the development of the virulence factor database as a specialized repository that can facilitate the researchers to quickly mine the virulence factors of specific bacteria has become necessary. Hence, in the present research, the database EVirDB (http://www.evirdb.info), presented as a source for the rapid access of different virulence factors based on the gene ontology terms. The EvirDB is designed as a comprehensive and user-friendly database. The database provides the users to submit the query and to retrieve the virulence factors data of bacterial pathogens that belong to *Enterobacteriaceae*.

KEYWORDS: **Virulence factors, Gene ontology, Pathogens, *Enterobacteriaceae*, Database.**

*Corresponding author

GNANENDRA SHANMUGAM
Department of Biotechnology, Mahendra Arts & Science College, Kalippatti, Namakkal, TamilNadu, India.

INTRODUCTION

Even though the genomes of many bacterial pathogens have been completely sequenced, the limitation in the selection of suitable drug targets and the validation for determining the host-pathogen interactions is observed [1]. The drastic increase in the bacterial genome projects, have paved a path for the identification of potentially novel virulence-associated factors and their possibility as a novel drug target. Even though many significant sequence databases harboring huge amount of genome and protein sequence data, we are unfortunate to have the virulence factors that determine the pathogenesis of the bacteria [2-4]. In recent years, the continuous usage of antibiotics resulted a dramatic increase in the emergence of multidrug resistance in the bacteria [5-8]. This has led an alarming concern to all researchers to significantly identify the potential virulence factors as a most promising and alternative candidate drug targets that can overwhelm the worse situations of multidrug-resistant bacteria that has a concerning issues of public health sectors. The increased usages of extended spectrum antibiotics have led to the emergence of multidrug resistant bacteria[9]. Thus the need for specialized databases that can address the genes encoding for the virulent factors and their functions is in demand for the identification of novel drug targets. The *Enterobacteriaceae* is a bacterial family with the most important groups of Gram-negative bacteria. This family includes a number of important foodborne pathogens such as *Salmonella spp., Escherichia spp., Shigella spp.,* and other opportunistic pathogens such as *Klebsiella spp., Proteus spp., Enterobacter spp* especially in the clinical setups. Thus, in the present work, we emphasized a database of five major virulence factors from six most significant human pathogens of *Enterobacteriaceae*. This comprehensive database can serve as a source for the selection of significant virulence factor based on the intellectual Gene Ontology terms [11] that plays a critical role in the pathogenesis and its surveillance in the host.

METHODOLOGY

Data collection

The virulence factors from six pressing important human pathogens of *Enterobacteriaceae* were retrieved from the UniProt database [12]. The virulence factors information's such as Protein Name, Gene Name, Organism, Sequence Length, Function, Domain, InterproID, GoID, Gene Ontology, Keywords and Sequence were fetched by our IR (Information Retrieval) algorithm using python script (Fig 1).

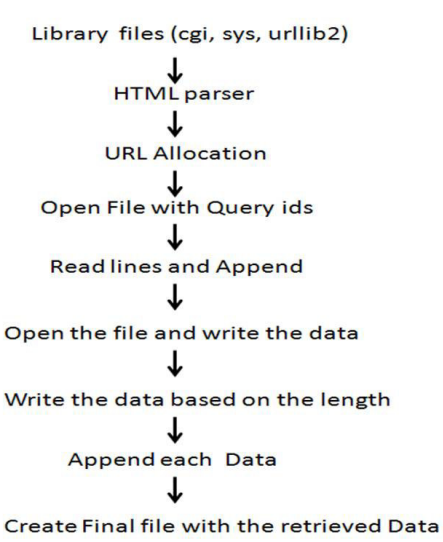
Data Classification

The Virulence factors information is classified as Capsule, Cell Wall, Flagella, Pili and Toxins. The Gene Ontology term is classified as Biological Process, Molecular Function and Cellular Component. The data are sorted according to the 6 pathogens such as *Enterobacter* spp., *Escherichia* spp., *Klebsiella* spp., *Proteus* spp., *Salmonella* spp., and *Shigella* spp., This classification of data helps the users to search and easily fetch the particular virulence factor from the desired pathogen based on the GO term.

Database Implementation

EVirDB was implemented using the MySQL relational database system (http://www.mysql.com) to construct the background relational database to store the information of virulence factors and the gene sequence data. The PHP programming and certain modules of DBI and CGI, were also used to communicate with the database and sends all kinds of pages to the Apache web server for users. Some webpage effects were also made by client-side JavaScript (Fig 2).

Algorithm and Python script used for data collection




```
import cgi
import HTMLParser
import os
url = 'http://www.uniprot.org/'
var1= open("proteinID.txt", "r")
for line in variable:
        strip()
        append(e)
var1.close()
var2=open (filename,'a')
for k in range(0, length, 1):
        write(seq)
Var2.close()
var3=open("finalfile.txt","a")
        write (seq)
        write("\n");
        if k % 2 == 0:
         lenh=len(lee[k])
        else:
         var3.write(lee[k])
         var3.write("\n");
var3.close()
```

Figure 1
Information Retrieval Algorithm implemented for data collection

Modules used in the development of EvirDB

Figure 2
The modules used in the implementation of EVirDB

RESULTS AND DISCUSSION

The EVirDB holds the virulence factors and its related information's collected from various sources. The warehouse is committed to envisage the functional and structural aspects which can ignite researchers to elucidate its role as pathogenesis and its mechanism to emerge as a most dreadful infectious disease. This can also serves as a specialized source to retrieve the novel virulence factors that are ignored in due course of time. These understandings can ecstasy the researchers to attain the immediate investigations for the development of novel approaches and drugs to treat and prevent the diseases caused by Multidrug resistant (MDR) bacteria. The current release of EVirDB harbors the 5 virulence factors, Capsule, Cell Wall, Flagella, Pili and Toxins from 6 noteworthy emerging Multidrug resistant bacteria of ***Enterobacteriaceae*** namely,

Enterobacter spp., *Escherichia* spp., *Klebsiella* spp., *Proteus* spp., *Salmonella* spp., and *Shigella* spp., The collective information's such as, Protein Name, Gene Name, Organism, Sequence Length, Function, Domain, InterproID, GoID, Gene Ontology, Keywords and Sequence were deposited. The warehouse is designed as a comprehensible resource for researchers to search for Gene Ontology terms and the intellectual virulence factors and their related information's. The EVirDB is accessible via the World Wide Web at http://www.evirdb.info. The data's were collected from UniProt and classified based on the Gene Ontology terms & keywords and deposited in the database. The scheme of database development is depicted in Fig 3. The HTML interface allows the user to query the database using Gene Ontology terms, Virulence Factors or by Organism names. The information of virulence factors and related gene and protein sequences can be retrieved, from EvirDB using appropriate Query. The Hypertext Pre-processor scripting language was used to retrieve the data and also as a server- side HTML embedded scripting language for building dynamic pages based on the user's query. The EvirDB Home page is shown in Fig 4.

EvirDB Development Process

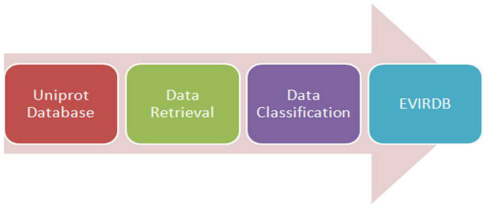


Figure 3
The overview of EVirDB Development

Home page of EVirDB

Figure 4
The EVirDB home page accessed at http://www.evirdb.info

The query based on Gene Ontology terms, virulence factors, organism to retrieve the information from EVirDB and the result page retrieving the information based on the search Gene Ontology were shown in Fig 5 and Fig 6.

EvirDB Query search Page

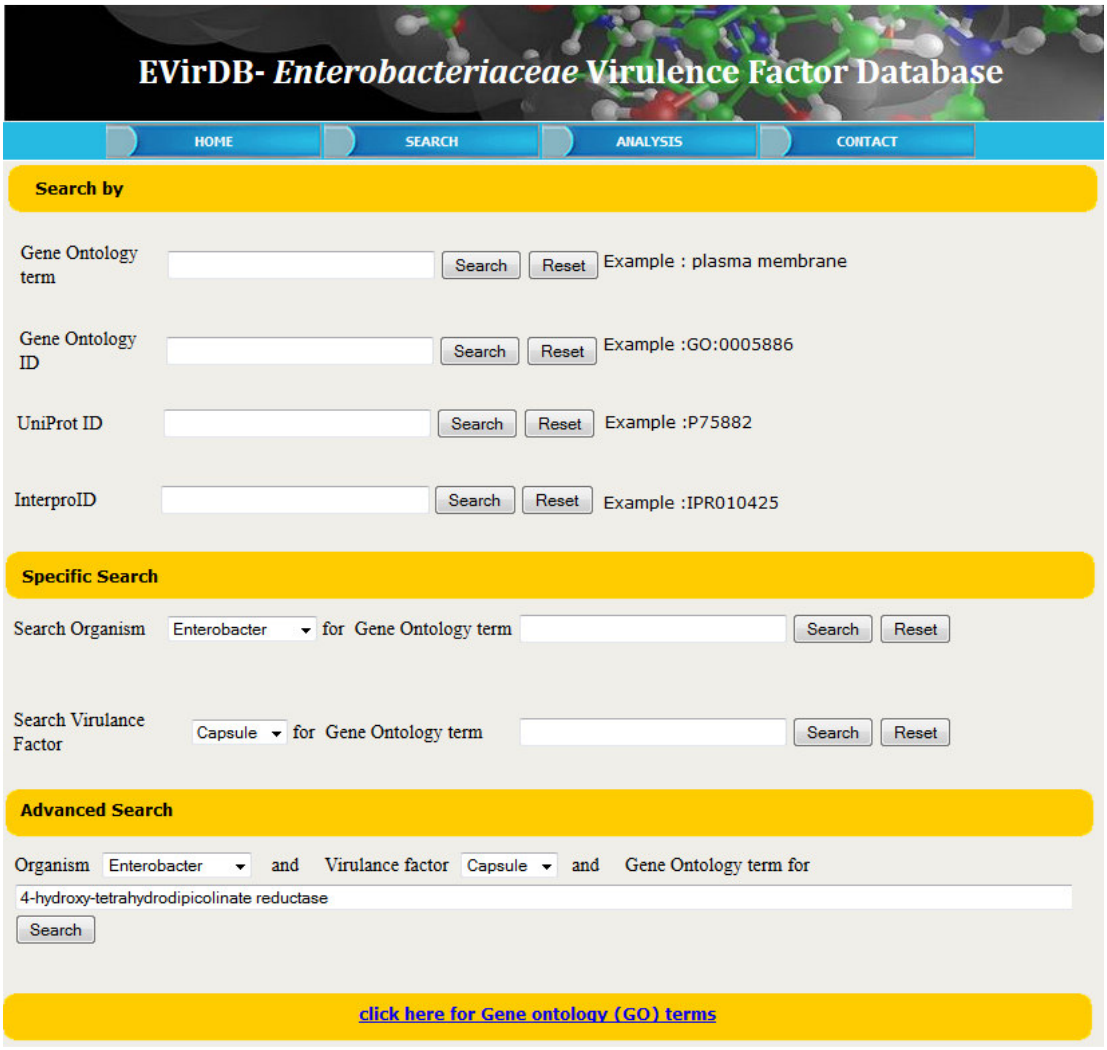


Figure 5
EVirDB query page based on Gene Ontology, Organism and Virulence Factors search

EvirDB result page based on user search query

Figure 6
The result page based on the query submission of Gene Ontology term.

In brief, the current release of EVirDB focuses six important pathogens namely *Enterobacter* spp., *Escherichia* spp., *Klebsiella* spp., *Proteus* spp., *Salmonella* spp., and *Shigella* spp., that belongs to *Enterobacteriaceae*. The 5 noteworthy virulence factors, Capsule, Cell Wall, Flagella, Pili and Toxins data's that mount to ~36,000 were included in the database. The virulence factors of each organism stored in the EVirDB were shown in Table 1. The already available VFDB[13] address only 459 virulence factors from the 26 pathogenic bacteria and 2505 VF related genes. That database focuses the virulence factors based on the functionality of the proteins. Whereas our EVirDB consists of more number of Virulence factors from the 6 pathogens. This difference in the data is due to the usage of Gene Ontology terms in our algorithm during the collection of data. The total of Gene Ontology terms used to retrieve the virulence factor information was alphabetically sorted to make the query search as user friendly. The number of Gene Ontology terms used to query the database is listed in Table 2. The links to various annotation tools such as ProtParam, Motif Scan, TargetP, TMPred, GOR, Swiss Model and WhatIF links were provided under Analysis menu.

Table 1
The virulence factors of Enterobacteriaceae deposited in EvirDB

Organisms	Virulence Factors					Total
	Capsule	Cell Wall	Flagella	Pili	Toxin	
Escherichia spp.	64	25	574	285	76	1024
Enterobacter spp.	67	38	1028	339	17	1489
Klebsiella spp.	343	189	533	1972	58	3095
Proteus spp.	5	4	255	153	0	417
Salmonella spp.	758	711	16483	8818	347	27117
Shigella spp.	153	87	1816	672	21	2749
Total	1390	1054	20689	12239	519	35891

Table 2
The number of Gene Ontology terms used to query the database EvirDB

Organisms	Virulence Factors					Total
	Capsule	Cell Wall	Flagella	Pili	Toxin	
Escherichia spp.	61	22	87	34	71	275
Enterobacter spp.	34	14	67	14	04	133
Klebsiella spp.	36	14	20	27	04	101
Proteus spp.	02	09	61	15	00	87
Salmonella spp.	47	22	83	69	09	230
Shigella spp.	33	11	71	17	13	145
Total	213	92	389	176	101	971

CONCLUSION

EVirDB is committed to provide the rational information on virulence factors of most important bacteria. The current release is focused on *Enterobacteriaceae* namely *Enterobacter spp., Escherichia spp., Klebsiella spp., Proteus spp., Salmonella spp., and Shigella spp*. thus database is aimed to serve as a specialized database that focuses on the 5 important virulence factors such as Capsule, Cell Wall, Flagella, Pili and Toxins of important human bacterial pathogens. As the expression of virulence genes and their pathways plays a significant role in the regulation of pathogenicity of the bacteria and also to colonize the host system, we have future plan to include much details on regulation and host interactions of the virulence factors.

REFERENCES

1. Mulder NJ., Mazandu GK., Rapano HA. Using Host-Pathogen Functional Interactions for Filtering Potential Drug Targets in Mycobacterium tuberculosis. J. Mycobac. Dis., 3:126 (2013).
2. Warner DF., Mizrahi V. Approaches to target identification and validation for tuberculosis drug discovery: a UCT perspective, South African Medical Journal, 102: 457-460 (2012).
3. Waldvogel FA., Infectious diseases in the 21st century: old challenges and new opportunities, Int. J. Infect. Dis., 8: 5–12 (2004).
4. Weiss RA., Virulence and pathogenesis, Trends in Microbiology, 10: 314–317 (2002).

5. Hogan D., Kolter R., Why are bacteria refractory to antimicrobials? Curr. Opin. Microbiol., 5:472–477 (2002).
6. Byarugaba DK., Antimicrobial resistance in developing countries and responsible risk factors, Int. J. Antimicrob. Agents., 24:105–110 (2004).
7. Docampo R., New and reemerging infectious diseases, Emerg. Infect. Dis., 2003; 9: 1030– 1033 (2003).
8. Morens DM., Folkers GK., Fauci AS. The challenge of emerging and re-emerging infectious diseases, Nature, 430: 242–249 (2004).
9. Dhandapany Senthil Pragash, Mangaiyarkarasi., Latha Ragunathan., Sageera Banoo., Vijaya Rayapu., Ivvala Anand Shaker. Occurrence of CTX-M and SHV genes in ESBL producing gram negative organisms causing pyogenic infections in a tertiary care hospital in puducherry, Int J Pharm Bio Sci, 3(4): (B) 782 – 78 (2012).
10. Toth IK., Pritchard L., Birch PR., Comparative genomics reveals what makes an enterobacterial plant pathogen, Annu. Rev. Phytopathol., 44: 305-36 (2006).
11. Ashburner M., Ball CA., Blake JA., Botstein D., Butler H., Cherry JM., Davis AP., Dolinski K., Dwight SS., Eppig JT., Harris MA., Hill DP., Issel-Tarver L., Kasarskis A., Lewis S., Matese JC., Richardson JE., Ringwald M., Rubin GM., Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25: 25-9 (2000).
12. Apweiler R., Bairoch A., Wu CH. Protein sequence Databases, Curr. Opin Chem. Biol., 8: 76-80 (2004).
13. Chen L., Yang J., Yu J., Yao Z., Sun L., Shen Y., Jin Q. VFDB: a reference database for bacterial virulence factors, Nucleic Acids Res., 33(Database issue):D325-8 (2005).