Chapter 3 - ProtVirDB: a database of protozoan virulent proteins
3.1 Background

In the current scenario there is a deluge of information emanating from the successful completion of genomic projects of pathogens. However at the same time it becomes important to characterize proteins encoded in them and organize them in a way so as to enable comparison with the already discovered proteins and the imminent ones, to further the discovery of novel drug or vaccine targets. Virulence is the ability of any pathogen to cause disease. Virulent proteins are an important class of proteins enabling pathogens to evade host immune mechanisms to cause diseases in the host. There is an ever-growing interest to identify novel virulent proteins in variety of pathogens in order to counter growing drug resistance and to develop novel vaccines. Well-defined classes for bacterial virulent proteins (Prescott, et al., 1999) have been described, however, no such classification is available for protozoans virulent proteins, and reasonably so because of intricate and multifactor virulence mechanisms, involving adhesive factors, pore-forming proteins, apoptosis and survival molecules, relying on different and often multiple, mechanisms for pathogenesis like antigenic variation, etc. Databases of virulent proteins enable easy comparison of virulent proteins in organisms to facilitate identification of novel virulence factors, comparison of virulent factors in different organisms related to each other, or on the evolutionary scale and facilitate studies related to mechanism of adaptive evolution in parasites.

There are databases of bacterial virulent proteins like Virulence Factors Database or VFDB (Yang, et al., 2008) and MvirDB (Zhou, et al., 2007). For fungal pathogens also, there is an analogous database called the pathogen–host interaction database (PHI-base) (Winnenburg, et al., 2008). However, there is no report of any such database for protozoans. Parasitic protozoans have assumed great medical and veterinary importance, attributable to the cosmopolitan life-styles of the parasites (e.g. Toxoplasma), emerging drug resistance and significant levels of morbidity and mortality that they exact on their hosts. Comprehensive studies on protozoan pathogenesis have widened our viewing aperture for virulence mechanism and immune evasion factors, which is important to search relevant targets for chemotherapeutic intervention and prevention of the disease in the light of emerging drug resistance in pathogens. There is an umpteen number of proteins known to play a key role
in the virulence of protozoan parasites, however the information is rather scattered in literature. Therefore, we have assembled all this information in the form of a database called ProtVirDB and provided useful analysis tools with it.

3.2 Database content

The current release (V 1.0) holds a cumulative collection of 345 unique virulent proteins (and 1775 total entries) from twelve important parasitic protozoans (Table I and Figure 3.1). The database entries were manually curated from bibliographic (PubMed) and sequence (GenBank, RefSeq, SwissProt) databases (Figure 3.2). The database includes many polymorphic proteins like trans-sialidase in *T. cruzi*, VSG in *T. brucei* etc. These make up a substantial bulk of the number of sequences. But counting the polymorphic forms (slightly different protein sequences) of the same protein as one, these make one ‘unique’ entry. So the number of total entries is less than the number of total entries. For 30 proteins there is no sequence available and these have been identified by functional assays only.

We have attempted a function-based classification for these proteins. However the delineation amongst different categories is rather vague, for example adhesion and invasion (e.g. CSP protein from *Plasmodium falciparum*) and is for the purpose of a broad outline.

Based on the currently available literature, each protein was allotted to one of the following functional categories: Adhesin, Invasion, Establishment (within the host, i.e. proteins involved in nutrient acquisition or evasion of host immunity), Proteases, Cysteine proteases, Heat shock proteins and Others (which cannot be defined within any of these categories). Cysteine proteases serve a multitude of roles like cytoadherence, invasion etc., so an exclusive category has been devoted to these.
Table I. Statistics of the ProtVirDB database

<table>
<thead>
<tr>
<th>S. no.</th>
<th>Organism</th>
<th>Total number of repeat containing proteins/unique proteins</th>
<th>Total number of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cryptosporidium parvum</td>
<td>6/20</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>Toxoplasma gondii</td>
<td>13/74</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>Eimeria tenella</td>
<td>5/16</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>Giardia intestinalis</td>
<td>1/66</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>Entamoeba histolytica</td>
<td>11/29</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>Babesia bovis</td>
<td>10/15</td>
<td>132</td>
</tr>
<tr>
<td>7</td>
<td>Theileria anulata</td>
<td>6/9</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>Plasmodium falciparum</td>
<td>33/45</td>
<td>380</td>
</tr>
<tr>
<td>9</td>
<td>Trichomonas vaginalis</td>
<td>0/24</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>Leishmania donovani</td>
<td>2/13</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>Trypanosoma cruzi</td>
<td>4/22</td>
<td>830</td>
</tr>
<tr>
<td>12</td>
<td>Trypanosoma brucei</td>
<td>5/12</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td>101/345</td>
<td>1775</td>
</tr>
</tbody>
</table>

The total number of proteins is greater than the number of unique sequences because many of the virulent proteins are polymorphic and have many different known sequences. For 30 proteins there is no sequence available and these have been identified by functional assays only.

Figure 3.1 Statistical distribution of the species in the database

28
Chapter 3- ProtVirDB: a database of protozoan virulent proteins

Figure 3.2 ProtVirDB database development. The protein sequences were collected by keyword search from different databases and then filtered by retrieving related articles from PubMed. Value addition included classification of filtered pool sequences into different categories and additional analysis (like Pfam domains, pI/Mw, PDB code, vaccine potential, and presence of amino acid repeats). The web based user interface integrated with powerful bioinformatics tools facilitates database query and analysis.

3.3 Database architecture and data retrieval

ProtVirDB is implemented as a MySQL database (Figure 3.3), with a single table containing 13 fields with different types of information, i.e. basic information, hyperlinks to other servers wherever relevant and pre-calculated protein properties. The basic information includes organism name, accession number, sequence, functional category and a brief description about the protein. The hyperlinks are provided to PubMed for the research articles which describe that protein and TDR targets database (Aguero, et al., 2008) if the protein is listed as a target in it. Additional PubMed links for verified or predicted vaccine or immunotherapeutic targets are included. Each protein is linked to additional information comprising its molecular weight, pI, PDB code, Pfam domains (Finn, et al., 2008), amino acid repeats.
Figure 3.3 ProtVirDB database schema. The schema depicts the different types of information stored in the MySQL table

The web interface provides links to several web-based utilities for calculation of antigenic index and epitope prediction.

PHP is used to connect the database and dynamically generate user-friendly HTML front-end queries, using Apache web server. The web interface query form (Figure 3.4) allows users to selectively retrieve a table enlisting details (functional category and a brief description) of virulent proteins from one or more organisms within a single or multiple functional categories (Figure 3.4). Users can selectively download the proteins of interest as an excel table or FASTA file. Alternatively, the database can be queried with user-defined keywords, with accession number or molecular weight combined with the filter based on organism name (Figure 3.4 and 3.5).
Chapter 3: ProtVirDB: a database of protozoan virulent proteins

Figure 3.4 The query form of ProtVirDB

Figure 3.5 The results of a query with keyword ‘mitochondria’
3.4 Integrated web based tools

ProtVirDB is integrated with several useful tools to facilitate sequence retrieval and analysis. The ViroBLAST (Deng, et al., 2007) tool allows users to search for entries in ProtVirDB that have sequence similarity to query protein sequences. This provides the advantage of parsing the results according to an E-value or score cut-off chosen by the user. The integrated ClustalW (Thomson et. al., 1994) and Muscle (Edgar, 2004) tools, supplemented with the colorful display generated by Jalview (Clamp et al., 2004) perform multiple sequence alignment of selected sequences. The Java-based ATV program (Clamp et al., 2004) allows the viewing of phylogenetic trees obtained from the QuickTree program (Howe et al., 2002). The antigenic program from the EMBOSS package (Rice et al., 2000) predicts potentially antigenic regions of a protein sequence.

The detection of conserved motifs in protein sequences is critical for annotation of proteins. The available tools for this purpose like PPsearch (http://www.ebi.ac.uk/Tools/ppsearch/index.html), allow the user to scrutinize only the already recognized and conserved motifs in databases like Prosite.

Herein, we developed a simple and versatile tool called ProbeMotif (Figure 3.6). This facilitates search of user-defined motifs within the ProtVirDB database or any other user defined set of sequences (Figures 3.7 and 3.8). This is a PERL based tool that allows the users to search for motifs using regular expressions including wildcards. This serves as a supplementary tool, especially in cases where a newly discovered motif can be quickly searched in the database.
Chapter 3- ProtVirDB: a database of protozoan virulent proteins

Figure 3.6 The tool ProbeMotif allows the user to search for pre-defined motifs in any set of sequences.

<table>
<thead>
<tr>
<th>Identifier (aa len)</th>
<th>Motif</th>
<th>Size</th>
<th>Frequency</th>
<th>Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;AAA83031 Hemolysin (Invasion) [Cryptosporidium parvum]</td>
<td>CPETLNEFNSMYK</td>
<td>13</td>
<td>1</td>
<td>17..29</td>
</tr>
</tbody>
</table>

Figure 3.7 Results of the above query given by ProbeMotif

Figure 3.8 Clicking on the Graphical Display displays the sequence with the motif highlighted in red color
3.5 Amino acid repeats analysis

Amino acid repeats have been correlated with virulence in bacteria as well as protozoans (Fankhauser et al., 2007). We scanned the entire ProtVirDB database for the presence of repeats using the DIREP program (Kalita et al., 2006). Interestingly, both homo and hetero-repeats were detected in 32% of the total sequences (101 unique proteins out of 315). The repeat-containing protein sequences are available on the webserver at http://bioinfo.icgeb.res.in/protvirdb/repeats. In the database, the Plasmodium falciparum sequences alone accounted for 33 proteins with repeats (out of 45), followed by Entamoeba histolytica 11 (out of 29), see Table I for details. These figures are strikingly high when compared with the percentages of repeat containing proteins in the entire proteomes (33.49 and 2.79%, respectively, Depledge et al., 2007). This may well be an underestimation of the proportions since the set of ProtVirDB virulent proteins represents only the currently annotated virulent proteins in the parasite genomes. Yet the presence of repeats in almost one-third of the proteins within a small collection is certainly intriguing and reinforces that repeat-containing proteins play an indispensable role in the parasite’s virulence. It is noteworthy that we did not observe any bias of repeat-containing proteins within any specific functional category. Cysteine proteases, proteases and heat shock protein categories were scantily represented in this set but this bias could be due to their under-representation in the database.

Such a high preponderance of repeats in virulent proteins demands for a further study to verify in vivo how exactly the repeats modulate protein’s function within the parasite and whether they actually play some role in enhancing its virulence. If this is definitively proved, we would have already identified plenty of vaccine or drug targets to aim at by using repeat finding tools.

3.6 Perspectives

Diseases caused by parasitic protozoans are often studied in isolation; however comparative studies may provide a key to hitherto undiscovered but common mechanisms of virulence. The ProtVirDB database can assist in research efforts aimed at such comparative studies. It also provides ground for further studies related to the significance of repeat-containing proteins in the virulence of the protozoan parasites. The
database will be updated regularly and additional tools incorporated. Given the mounting interest in protozoan parasitic diseases, we expect ProtVirDB to serve as a valuable resource to the scientific community.
Chapter 3- ProtVirDB: a database of protozoan virulent proteins

References


36
Chapter 3 - ProtVirDB: a database of protozoan virulent proteins


