Identification of Protein Biomarkers in Cholera and Rational Drug Designing

Suma, N.*, Tanusree Chaudhuri and Kusum Paul

Department of Biotechnology, The Oxford college of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India.

ABSTRACT: Cholera is an acute diarrheal infection caused by ingestion of the bacterium Vibrio cholera of the O1 or O139 serogroup. Transmission occurs through direct fecal-oral contamination or through ingestion of contaminated water and food. The virulence factors include flagellin; cholera toxin (CT), TcpP/TcpH proteins, which, together with the ToxR/ToxS proteins, activate the expression of the ToxT regulatory protein. A bacterial flagellins and cholera toxins are known to be potent inducers of proinflammatory molecules via activation of Toll-like receptor. The cholera toxins encodes ctxA and ctxB genes carried on the transmissible prophage CTX-Φ which regulates the transcriptional activator ToxR and intern activates the ToxT, a secondary transcriptional regulator that activates the expression of CR and the toxin coregulated pilus (TCP). In the present method of characterizing Vibrio cholera strains involve serotyping of presence or absence of genes thought to be markers of an organism’s pathogenicity. It is unclear that the protein assays detect all pathogenic Vibrio cholera strains since a clear correlation between the presence of a particular gene and the organism’s pathogenicity has not yet been observed. In the present study includes protein biomarkers is helps to identify the signaling proteins that makes the cholera bacteria stop swimming and start to colonize in the human intestine.

KEYWORDS: Biomarker; Cholera; Rational Drug Design.

Introduction

During the 19th century, cholera spread repeatedly from its original reservoir or source in the Ganges delta in India to the rest of the world, before receding to South Asia. Seven pandemics were recorded that killed millions of people across Europe, Africa and the Americas[1]. The 17th pandemic, which is still ongoing, started in 1961 in South Asia. The disease is now considered to be endemic in many countries and the pathogen causing cholera cannot currently be eliminated from the environment. Two serogroups of V. cholerae - O1 and O139 - can cause out breaks. The important reservoirs are human beings and aquatic sources such as brackish water and estuaries, often associated with algal blooms (plankton). Recent studies indicate that global warming might create a favorable environment for cholera and increase the incidence of the disease in vulnerable areas[2].

The G protein-coupled receptors (GPCRs) are a family of proteins expressed at the cell membrane. The important functions of the organism and are privileged therapeutic targets and r expression is highly modulated depending on the metabolic state of the cells, in particular in[3] pathological situations.

Methodology

Insilco Discovery

The strains used in this study are O1 and O139 of V.Cholerae. The genotypic profiles for V. cholerae-associated determinants were selected according to pathogenesis on clinical tests. There are several cholera databases contains genome, proteome and pathogenic information. The cholera genome database (broad institute of cholera genome database) contains pathogenic sequences. There are six pathogenic cholera strains present in cholera bacterial membrane. There are several databases to search genes and proteins to identify protein markers with highly specific to or strongly expressed in different tissues.

Gene and Proteins in Pathogenic to Cholera Disease

Based on complete genomes of seven V.Cholerae representative stains. We search repeats of membrane protein sequences were selected from NCBI and Uniprot databases. All nucleotide sequence analyses and alignments were performed with MacVector, version 11.0.2. CTXφ, tcp, toxR, o-antigen, Accessory cholera enterotoxin (ACE), cholera toxin (CT), zonulaocculdentstoxin (ZOT) virulence factor protein sequences was performed using BLAST, HHpred, Phylogenetic analysis of all virulence factors were
analyzed using clustal W. One hundred bootstrap data sets were analyzed to evaluate the significance of the tree.

**Protein Peptides Prediction**

The Peptide Cutter tool is used to predict protein peptides based on the protein digestive enzymes. The protein sequences were sequentially cleaved from the N-to C-terminal. The digestive peptides provide the respective lengths in amino acids and the molecular weight in Daltons. The peptides displayed are calculated in a way of assuming that all chosen enzymes are present during digestion. We listed peptides resulting from the cleavage of only one enzyme or chemical, modification neither of the protein sequence nor of modifications evoked by the cleavage.

**2D GEL Predictions of Protein Peptides**

The selected protein peptides mass is routinely used for the post-electrophoretic identification of proteins extracted from 1D and 2D electrophoresis gels. The automated JvirGel software to develop 2D gel of selected protein sequences and processed for peptide mass mapping and uses the Genomic Solutions robots for spot cutting from gels, sample digestion, peptide extraction and sample spotting for mass spectrometry.

This software provides peptide mass information of gram positive and gram negative bacterial peptides.

**Mass Spectrometry Data Analysis**

To identify large amounts of high-quality data that in turn allows protein identification, annotation of secondary modifications, and determination of the absolute or relative abundance of individual proteins. We analyzed the spectral data based on the molecular weight of peptide. The main purpose is to quantify the sample peptide fragments and generate the mass spectrum which visualizes the peptide abundance and distribution in a 2-D electrophoretic gel image. There are two methods to analyze the mass data. The first process is ionization, which can convert atoms or molecules to gas phase ions by adding or removing charged particles. The peptides peaks were analyzed based on higher mass spectrometry. The mass is analyzed by capture of ions and separate ions according to their mass-charge ratio (m/z).

**Selection of Peptide Mass**

Using Protein Prospector to identify the best selected protein peptides. We predicting peptide library and data resource of >100,000 synthetic, unmodified peptides and their phosphorylated sites. Analysis of the library by mass spectrometry yielded a data set that we used to evaluate the merits of different search engines and fragmentation methods for peptide identification.

We also compared the sensitivities and accuracies of phosphorylation-site localization tools and we characterized the chromatographic behavior of peptides in the library. We found phosphorylates sites of unmodified peptides and those current computational tools for proteomics. These peptides and spectra will facilitate the development, evaluation and improvement of experimental and computational proteomic strategies, such as separation techniques and the prediction of retention times and fragmentation patterns.

**Biomarker Identification**

Selected protein peptides were predicted using bioinformatics databases. The gene and protein databases were mined to identify proteins highly specific to or strongly expressed in one or more tissues. We examined stomach, small intestine, large intestine, liver, kidney tissues were examined. In our study we used five gene databases and one protein database. The C-It, Tissue-specific and Gene Expression and Regulation (TiGER) and UniGene databases are based on expressed sequence tags (ESTs). The BioGPS and VeryGene databases are based on mass spectrometry data. The Human Protein Atlas (HPA) is based on immunohistochemistry (IHC) data.

- **C-IT Database**
  
  We used cholera tissue for protein enriched in that selected human tissue. Since the C-It database is used to analyze the tissue on cholera. The literature search provide less information of tissue specific option to add z-scores of the corresponding SymAtlas microarray probe sets to the protein list. The SymAtlas z-score of ≥ 1.96, corresponding to a 95% confidence level of enrichment.

- **Tiger Database**
  
  The TiGER database was searched for proteins preferentially expressed in each tissue based on ESTs by searching each tissue using ‘Tissue View’. The UniGene database was searched for tissue-restricted genes using the following search criteria: [tissue][restricted] + “Homo sapiens”, for the stomach, small intestine, large intestine, liver, kidney. Since the UniGene database did not have data for cholera toxic to small intestine tissue.

- **BIOGPS Database**
  
  The BioGPS database plug-in ‘Gene expression/activity chart’ using the default human data set ‘GeneAtlas U133A, germa’ was searched with a protein whose gene expression profile using the BioGPS plugin showed it to be specific to and strongly expressed in one tissue of interest.

- **HPA (Human Protein Atlas)**
  
  The HPA was searched for proteins strongly expressed in each normal tissue with annotated expression. Annotated protein expression is a manually curated score based on IHC staining patterns in normal tissues from two or more paired antibodies binding to different epitopes of the same protein, which describes the
distribution and strength of expression of each protein in cells.

- **Identification of Protein Overlaps in Databases**
  We evaluate the number of cholera proteins were analyzed using orientation of proteins in membranes (OPM) database. The proteins were identified in two or more databases could represent candidates that are more promising at this stage, since databases based on varying sources of data identified the protein as being highly specific to or strongly expressed in one tissue.

- **Secreted Proteins**
  The list of protein expressions were secreted using secretome algorithm to identify proteins that are either secreted or shed. The secretome algorithm designates a protein as secreted or shed if it is either predicted to be secreted based on the presence of a signal peptide or through non-classical secretion pathways, or predicted to be a membranous protein based on amino-acid sequences corresponding to transmembrane helices. Proteins that were not designated as secreted or shed were eliminated.

**Biomarkers Identification and Validation**
Biomarker data can be retrieved efficiently through establishment of entry portals for search functions. The infectious disease biomarker databases are providing functional enrichment information on cholera disease. The plasma proteins with differentially expressed gene sets with our human biofluid proteome database to yield potential protein biomarkers for cholera disease.

**Vaccine Designing and Validation**
Genome to vaccine design using T-cell epitopes, short peptide sequences displayed by antigen presenting cells to T cells, to create immunogenic and protective vaccines from whole genomes. Four major steps for vaccine design and vaccine redesign.
- Genomes are mined using computational tools to identify genes that encode proteins with promising vaccine antigen properties such as secretion, upregulated expression, immunogenicity and virulence.
- Immuno informatics tools are then used to map protein sequences for short, linear putative T-cell epitopes.
- Sequences are synthesized as peptides and evaluated for human leukocyte antigen (HLA) binding and antigenicity in survivors of infection or vaccines.
- Prototype epitope-based vaccines are evaluated for immunogenicity.

**Results and Discussion**
To identify the sequence similarity to predict genomic and proteomic islands using reference database NCBI. The results were refined the gene identification numbers covered. The presence of ctxA, tcpA, toxR, and toxT in environmental strains of V. cholerae. A total of 539 genome sequences of all V. cholerae strains and 787 protein sequences collected from uniprot database. We conventionally analyzed different virulence genes of V. cholerae, including tcpA-E, tcpA-C, ctxA, ctxB, and sto (gene encoding the heat-stable enterotoxin of V. cholerae). The list of virulence factors including genome and proteome data were presented in table 1.

<table>
<thead>
<tr>
<th>Virulence Factors</th>
<th>Serotype</th>
<th>No. of Genes</th>
<th>No. of Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flagellin</td>
<td>O1</td>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>Cholera Toxin</td>
<td>O1</td>
<td>137</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>TcpP</td>
<td>O1</td>
<td>48</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>32</td>
<td>15</td>
</tr>
<tr>
<td>TcpH</td>
<td>O1</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>ToxR</td>
<td>O1</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>ToxS</td>
<td>O1</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ToxT</td>
<td>O1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CTX-Φ</td>
<td>O1</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>ZotA</td>
<td>O1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>ACE</td>
<td>O1</td>
<td>48</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>OmpU</td>
<td>O1</td>
<td>32</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>hlyA</td>
<td>O1</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>O139</td>
<td>12</td>
<td>18</td>
</tr>
</tbody>
</table>

The proteins and genes are mostly involved in pathogenesis to cause cholera. We studied the virulence factors cholera toxin (CT) and toxin corregulated pili (TCP) shows main genetical alterations in genome structure that often integrates to cause pathogenesis. The ctx AB genes coding for CT are encoded on a filamentous bacteriophage CTXφ. TCP, is an essential colonization factor, was originally designated as part of a pathogenicity island named Vibrio pathogenicity island VPI, but this island has later on been proposed to be the genome of a filamentous phage, VPIφ.
There are two layers of search criteria were used to sort the search results. Briefly, proteins were ranked by a first criterion and a second criterion was activated if multiple proteins were matched by the first criterion. There are several options to choose from for each criterion, e.g., number of peptide matches, sequence coverage, or m/z deviation in ppm. In this study, the number of peptides matched and m/z deviation in ppm were used as the first and second criteria, respectively. If a protein ranked number 1 from a search, it was considered identified if it met the following criteria: (1) A minimum of 3 peptide m/z values was required to match values predicted from the theoretical digestion, and (2) the search result must be consistent with another database search with a second data set obtained from the same sample spot.

**Conclusion**

In this project, an Insilco proteomic approach was used for development of vaccine. We preferred several Insilco revolutionary tools for the identification of surface associated proteins and virulence factors. There are, however, several critical points that still limit the full exploitation of genomics technologies. We predicted the peptide sequences of all virulence factor proteins and have studied the mass of each peptide based on overlapping of peptide sequences using 2D gel electrophoresis for Protein identification by computer-assisted image analysis methods. Using MS analysis of designed 2D gel image was quantified by MALDI-TOF of MS/MS analysis. The highest peaks with red colored peptide sequence shows best vaccines and potentially used for treatment of cholera disease.

**Acknowledgement**

I extend my sincere thanks to the management of The Oxford College of Engineering for their support. I thank our principal Dr. Nagaraj R and our HOD Dr. Kusum Paul, Department of Biotechnology for providing necessary resources. I would like to express my sincere gratitude to my guide, Mrs. Tanusree Chaudhari, Assistant Professor, Dept. of BT, for her guidance, encouragement and constantly monitoring the progress of the project. She provided me the chance to do my project work under her guidance and gave advice at times. Her constructive and critical questions were an incomparable stimulus and always helped me to consider issues in greater depth. I would like to express my sincere gratitude to Mr. Shambu M.G, Assistant professor, Dept.of BT, for his continuous inspiration and encouragement throughout my year of work.
References


