Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Original Article

Computational drug target identification in *Streptococcus pneumoniae* using available microarray data

*Limpon Bora**

Assistant Professor, Dept of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam 784028, India

ARTICLE INFO

Article history:
Received 2 March 2013
Accepted 3 April 2013
Available online 1 May 2013

Keywords:
Clusters
K-mean clustering
Drug targets

ABSTRACT

Background: In spite the availability of several drugs and vaccines, bacterial pathogenic diseases remain a major health problem and concern worldwide, warranting identification of new drug targets for the design of more efficacious drugs. This is due to the fact bacteria's become resistant to a particular antibiotic over the course of usage.

Methods: In this work the microarray expression data is selected to identify new genes that co-express along with the already known virulent genes. The study was undertaken using microarray data available on SMD.

Results: The microarray data was analyzed and new genes were selected that co-expressed along with known virulent genes available on VFDB. Later genes essential and also dissimilar to the host genome were selected as putative drug targets. The study started with 4508 expressed genes with a time gap of approximately 8–10 min of which 21 were known to be virulent. Of these genes 450 clusters were made out of which 21 were selected. In these 21 clusters there were 50 essential genes from which 18 were similar to humans and 19 similar to gut flora which were discarded. 13 New genes identified may be used as putative drug targets for further studies.

Conclusion: The study will help in the identification of new drug molecules and their validation.

1. Introduction

Streptococcus pneumoniae, or pneumococcus, is Gram-positive, alpha-hemolytic, bile-soluble aerotolerant, anaerobic member of the genus *Streptococcus*[1] a significant human pathogenic bacterium, recognized as a major cause for pneumonia in the late 19th century. Pneumonia is an inflammatory condition of the lung and often characterized as inflammation of the alveoli and abnormal alveolar filling with fluid.[2–4] There is growing momentum to sequence bacterial genomes with a focus primarily on pathogens which encompass the majority of all genome projects, and has generated a large amount of raw material for computational analysis.[5–7] These data pose a major challenge in the post-genomic era, i.e., to fully exploit this treasure trove for the identification and characterization of virulent factors in these pathogens, and to identify novel putative targets for therapeutic intervention.[8–10] The target must be essential for the growth, replication, viability or survival of the microorganism, i.e., encoded by genes critical for pathogenic life-stages. The microbial target for treatment should not have any well-conserved homolog in the host, in order to address

* Tel./fax: 91 03712 275415.
E-mail address: limponbioera@gmail.com.

0974-6943/$ – see front matter
http://dx.doi.org/10.1016/j.jopr.2013.04.003

cytotoxicity issues. Genes that are conserved in different genomes often turn out to be essential.[11,12]

The possibilities of selecting targets through genomics-related methodologies are increasing. An interesting approach designated "differential genome display" relies on fact that genomes of parasitic microorganisms are generally smaller than the genomes of free-living organisms.[13,14] The genes present in the genome of parasitic bacterium and absent in the genome of a closely related free living bacterium, and are likely to be important for pathogenicity and may be considered candidate drug targets.[15,16] Another method that is drug target identification using side-effect similarity[6] uses targets for drugs which have so far been predicted on the basis of molecular or cellular features, for example by exploiting similarity in chemical structure or in activity across cell lines. The study of gene expression has been greatly facilitated by DNA microarray technology.[17] The anticipated floods of biological information produced by these experiments will open new doors into genetic analysis.[18] Expression patters have already been used in a variety of tasks.

Most bioresearch involves through the development of technology used for carrying them out. It is not possible to research on a large number of genes using traditional methods. Microarray is one such technology which enables researchers to investigate an issue which were once thought to be non-traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. Microarray technology has empowered the scientific community to understand the fundamental aspects the underlying the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of human body. Researchers hope to find molecules that can be targeted for treatment with drugs amount the various protein encoded by disease- associated genes. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1996.[6]

Clustering is the assignment of a set of observations into subsets called *clusters* so that observations in the same cluster are similar in some sense. It is also a common technique used for statistical data analysis in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics. Despite the availability of several drugs and vaccines, bacterial pathogenic diseases remain a major health problem and concern worldwide. This is due to the fact that bacteria become resistant to a particular antibiotic over the course of usage.

The objectives of the present study are prediction of probable virulent gene, identification of paralogous genes and co-expressed genes, prediction of essentiality of corresponding proteins and prediction of Putative Drug targets.

2. Materials and methods

2.1. Databases

2.1.1. Virulence factor database (VFDB)

VFDB is an integrated and comprehensive database of virulence factors of 24 bacterial pathogens.[11,18] VFDB is comprehensive and user-friendly and one can search VFDB by browsing each genus or by typing keywords (www.mgc.ac.cn/vfs). Furthermore, a BLAST search tool against all known VF-related genes is also available. VFDB also provides a unified gateway to store, search, retrieve and update information about VFs from various bacterial pathogens.

2.1.2. Stanford microarray data (SMD)

The SMD contains the largest amount of gene expression data from about 67 organisms.[14] It stores raw, normalized data and their corresponding image files. It also provides interfaces for data retrieval, analysis and visualization. SMD has its source code fully and freely available to others under an Open Source Licence, enabling other groups to create a local installation of SMD (www.ncbi.nih.gov/pubmed).

2.1.3. Database of essential genes (DEG)

The DEG holds information on essential genes from a number of organisms.[16,17] The current release 6.3 contains information on 11,392 essential genes from various organisms both prokaryotes and eukaryotes. A typical entry includes a database specific accession number, the common gene name. GI reference, function, organism, reference and nucleotide sequence (www.essentialgene.org).

2.2. Tools

2.2.1. Cluster tool

With the explosion of microarray data there is an emerging need to develop tools that can statistically analyze the gene expression data. There are many tools available on net for the same. Cluster is a tool for data clustering of genes on the basis of gene expression data. It is available at Eisen Lab and can run on Windows. It uses many clustering algorithms which include K-means, hierarchical, self-organizing map. The genes were clustered assuming the fact that genes that co-express along with the known virulent genes may also be responsible for the virulence.[15]

2.2.2. BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool, or **BLAST**, was used for primary biological sequence information comparison. BLAST2 was used for the identification of paralogs for virulent genes. BLASTP was used for protein sequence comparison available on the home page of DEG and also was done for human genome and microbial genome BLAST (www.blastncbi.nlm.nih.gov.in).

3. Methodology

3.1. Prediction of probable virulent gene products

In the study well-reported virulent genes for *S. pneumoniae* were taken from VFDB. Next the gene expression data was downloaded with a time gap of 8–12 h. Data was normalized for further study. To predict probable virulent genes normalized gene expression data was analyzed by the help of cluster software using K-mean clustering algorithm and found 450 clusters.

In *K-means clustering*, the numbers of clusters are designated (450), and then each gene is assigned to one of the K clusters by this algorithm before calculating distances. When