Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Haemophilus influenzae Genome Database (HIGDB): A single point web resource for *Haemophilus influenzae*

Rayapadi G Swetha [a], Dinesh Kumar Kala Sekar [b], Sudha Ramaiah [a], Anand Anbarasu [a,*], Kanagaraj Sekar [b]

[a] *Medical & Biological Computing Laboratory, School of Biosciences and Technology, VIT University, Vellore 632 014, India*
[b] *Laboratory for Structural Biology and Bio-computing, Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India*

ARTICLE INFO

Article history:
Received 29 July 2014
Accepted 1 October 2014

Keywords:
Haemophilus influenzae
Annotation
Genome
Database
Resistance

ABSTRACT

Background: *Haemophilus influenzae* (*H. Influenzae*) is the causative agent of pneumonia, bacteraemia and meningitis. The organism is responsible for large number of deaths in both developed and developing countries. Even-though the first bacterial genome to be sequenced was that of *H. Influenzae*, there is no exclusive database dedicated for *H. Influenzae*. This prompted us to develop the *Haemophilus influenzae* Genome Database (HIGDB).
Methods: All data of HIGDB are stored and managed in MySQL database. The HIGDB is hosted on Solaris server and developed using PERL modules. Ajax and JavaScript are used for the interface development.
Results: The HIGDB contains detailed information on 42,741 proteins, 18,077 genes including 10 whole genome sequences and also 284 three dimensional structures of proteins of *H. influenzae*. In addition, the database provides “Motif search” and “GBrowse”. The HIGDB is freely accessible through the URL: http://bioserver1.physics.iisc.ernet.in/HIGDB/.
Discussion: The HIGDB will be a single point access for bacteriological, clinical, genomic and proteomic information of *H. influenzae*. The database can also be used to identify DNA motifs within *H. influenzae* genomes and to compare gene or protein sequences of a particular strain with other strains of *H. influenzae*.

1. Introduction

Haemophilus influenzae is an important community-acquired bacterial pathogen, causing respiratory tract infections in both children and adults [1,2]. Even though, *H. influenzae* is a member of normal respiratory bacterial flora in the human upper respiratory tract, particularly nasopharnyx; it causes invasive infections by extending the organism from nasopharnyx to the lower respiratory tract [3,4]. Among the different serotypes, *H. influenzae* type b capsular strains are predominantly associated with severe systemic infections such as pneumonia, septicaemia, meningitis, empyema and septic arthritis [5–9]. The non-type-able *H. influenzae* strains commonly cause sinusitis, otitis media, acute lower respiratory tract infections and conjunctivitis [10]. The prevalence of antibiotic resistance in *H. influenzae* is rising and the optimum treatment for these severe infections has became more complicated [2,11]. The World Health Organization (WHO) estimates that around 386,000 child deaths occur annually due to *H. influenzae* meningitis and pneumonia in developed countries [11,12]. Thus, the infections by this bacterium are a major global public health problem, and therefore *H. influenzae* is being broadly investigated at the genome level [11,13]. Presently, 10 strains of *H. influenzae* have complete genomes and 15 strains have high throughput genome data. Thus, the number of *H. influenzae* genomes being sequenced is increasing subsequently leading to significant interest in comparing the genome of each strain with other strains. Strain level comparisons leads to better understanding of strain specific characteristics that may play an important role in virulence and antimicrobial resistance. The first bacterial genome to be sequenced was that of *H. influenzae*; however as per our understanding there is no database available exclusively for *H. influenzae*. Hence, in view of the above, we have attempted to develop a database, *Haemophilus influenzae* Genome Database (HIGDB). The HIGDB database provides a dynamic, user-friendly interface to execute varied Boolean searches or sequence based searches. The database provides links to tools like BLAST and DNA motif search that facilitate the comparison between multiple genomes of *H. influenzae* and identification of DNA motifs in *H. influenzae* genomes. The database is also interfaced with the genome map of *H. influenzae* strains and the three-dimensional structures of proteins of *H. influenzae* available in Protein Data Bank (PDB)

[*] Corresponding author at: VIT University, Tamil Nadu, India. Tel.: +91 416 2202547; fax: +91 416 2243092.
E-mail address: aanand@vit.ac.in (A. Anbarasu).

http://dx.doi.org/10.1016/j.compbiomed.2014.10.001
0010-4825/© 2014 Elsevier Ltd. All rights reserved.

[14]. These structures can also be exploited further for structural analysis based on the user requirements. Also, the database provides detailed information on the bacteriological characteristics, laboratory diagnosis, virulence factors and pathogenesis of *H. influenzae*. The purpose of HIGDB is to act as the universal single point access (one stop shopping) for researchers studying complete genomic and proteomic information of *H. influenzae* and to perform comparative studies on *H. influenzae* genomes. This analysis may give vital clues to understand the functions of most putative genes and to recognize the genome components of medical importance.

2. System design and implementation

The HIGDB database was developed with PERL/DBI and PERL/CGI modules and it has been hosted on Solaris server. This server has been particularly chosen for its adaptability, security and performance and it has been power-driven by 2.66 GHz Xeon (R) processor with 4 GB FDIMM main memory. The complete data of HIGDB were implemented in MySQL relational database. The front-end input data part was coded in HTML, JavaScript and Ajax which allows user-friendly web forms. The complete genomes of *H. influenzae* strains available in NCBI genome FTP site [15] were downloaded in Genome Feature Format (GFF3) and FASTA format. Then, they were loaded into GBrowse. The database has been completely validated and displays the results rapidly; however, it may differ based on the user network speed and traffic. The database has been tested on multiple platforms (iOS, Linux, Windows and Solaris) with different web browsers (Firefox, Chrome, IE and Opera).

3. Complex, user-friendly search options

The HIGDB database affords a powerful and user-friendly search engine. The complete annotations of genes/proteins of different strains of *H. influenzae* may be scrutinized by using either simple or advanced Boolean-based search tools. In simple search, the user can browse for various strains, genes and proteins of *H. influenzae* by entering strain/gene/protein name in text box, respectively. In addition, the hypothetical genes can be identified by entering the gene number in “gene search”. The advanced search has the options to return the list of proteins, localizing to a particular cellular localization. To serve downstream system level analysis, the database enables searching of proteins based on the Cluster of Orthologous Groups (COGs) category and on a specific pattern/profile. Further, it facilitates the user to fetch proteins, based on status (review/unreviewed) and virulence.

4. Facilitating sequence based DNA motif and BLAST searches

The DNA sequence motifs with major biological function have been becoming an important factor in the analysis of gene regulation [16] and they are located non-randomly in the genome [17]. We provided a search tool, “DNA Motif search” in HIGDB. This search tool is used to identify user-specified DNA motifs within the coding sequences of genes of *H. influenzae* strains. The tool accepts a stretch of DNA sequence with varying lengths in IUPAC format as an input sequence which is then converted into a regular expression. Additionally, BLAST tool [18,19] is also interfaced in HIGDB with which the user can perform the sequence similarity searches for both nucleotide and protein sequences against a particular or complete *H. influenzae* strains. When working with protein sequences, the BLAST tool locates the known domain within the sequence of interest. The tool allows users to set parameters like word size, gap open and extension penalty and substitution matrix. The links out from the BLAST results allow the researchers to look in further detail at a gene/protein of interest and a link to NCBI BLAST is provided; in case if the user wishes to perform the search against other genomes. The results generated by both DNA motif search and BLAST search can be stored in the hard disk of a local computer as a text document or in a Portable Document Format (PDF) file.

4.1. Case study

One of the important characteristics of *H. influenzae* is the preferential binding of its own DNA over foreign DNA [20–22]. *Smith* et al. deduced that the consensus uptake signal sequence in *H. influenzae* is “AAGTGCGGT” which is supported by Mell et al. [21,23]. The DNA motif “A{2}GTGCGGT” has been searched through the HIGDB DNA motif search tool against all *H. influenzae* strains and the results are shown in Table 1. The *H. influenzae* 10,810 genome has the highest number of occurrences (382) of this binding signal compared to genome of other *H. influenzae* strains (Fig. 1). This is just one example of how integration of this tool can led to new insights through the *H. influenzae* genome analysis.

5. Genome sequences utilizing GBrowse

In recent decades, the amount of genetic material available for *H. influenzae* related study is increasing due to the increasing number of genomes being sequenced. To ease this, a platform-independent web based application, Generic Genome Browser (GBrowse) has been incorporated in HIGDB. The GBrowse is a feasible and interactive viewer and it was developed by Stein et al. [24] of the Generic Model Organism System Database Project (GMOD). The browser has features like scroll, navigate and zoom in and out over the random regions of the genome. The user can fetch the region of genome or a landmark by specifying them in a search text box provided at the top left corner of the page. The search results show five tracks (i) genes (ii) proteins (iii) GC content (iv) 3-frame translation and (v) 6-frame translation. The landmark on each track carries a link to the corresponding information in HIGDB database or NCBI [15]. Thus, the HIGDB GBrowse makes the user to efficiently view the genomic content of different strains of *H. influenzae*.

5.1. Case study

The *H. influenzae* Rd KW20 is the first free-living organism to have its complete genome sequenced by the Institute for Genomic Research [25]. The strain is a derivative of a serotype d strain and it is considered to be avirulent as it lost the genes encoding its

Table 1
The number of occurrences of the motif “A{2}GTGCGGT” in each strain of *Haemophilus influenzae* identified by ‘DNA Motif search tool’.

Strain name	Number of occurrences
Haemophilus influenzae Rd KW20	329
Haemophilus influenzae 10810	382
Haemophilus influenzae 86-028NP	348
Haemophilus influenzae F3031	356
Haemophilus influenzae F3047	349
Haemophilus influenzae KR494	380
Haemophilus influenzae PittEE	301
Haemophilus influenzae PittGG	300
Haemophilus influenzae R2846	332
Haemophilus influenzae R2866	363