Genomics 104 (2014) 582–586

Contents lists available at ScienceDirect

Genomics

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

Streptococcus pneumoniae Genome Database (SPGDB): A database for strain specific comparative analysis of *Streptococcus pneumoniae* genes and proteins

Rayapadi G Swetha [a], Dinesh Kumar Kala Sekar [b], Ekambaram Durga Devi [b], Zaheer Zameer Ahmed [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 Biocomputing, Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India*

ARTICLE INFO

Article history:
Received 19 May 2014
Accepted 22 September 2014
Available online 28 September 2014

Keywords:
Pneumococci
Genome
Annotation
Resistance
Database

ABSTRACT

Streptococcus pneumoniae causes pneumonia, septicemia and meningitis. *S. pneumoniae* is responsible for significant mortality both in children and in the elderly. In recent years, the whole genome sequencing of various *S. pneumoniae* strains have increased manifold and there is an urgent need to provide organism specific annotations to the scientific community. This prompted us to develop the *Streptococcus pneumoniae* Genome Database (SPGDB) to integrate and analyze the completely sequenced and available *S. pneumoniae* genome sequences. Further, links to several tools are provided to compare the pool of gene and protein sequences, and proteins structure across different strains of *S. pneumoniae*. SPGDB aids in the analysis of phenotypic variations as well as to perform extensive genomics and evolutionary studies with reference to *S. pneumoniae*. The database will be updated at regular intervals and is freely accessible through the URL: http://pranag.physics.iisc.ernet.in/SPGDB/.

1. Introduction

Streptococcus pneumoniae or "*Pneumococcus*" is the most common cause of pneumonia and related invasive diseases such as bacterial meningitis, sepsis, otitis media and sinusitis. The bacterium causes extreme morbidity and mortality worldwide, particularly in infants and in the elderly. It spreads through immediate contact with respiratory discharges from patients and healthy carriers [1–3]. In several developed countries, the burden of the pneumococcal disease is intensified with chronic diseases (sickle-cell disease, chronic renal failure, chronic liver disease and asplenia), HIV and *Mycobacterial* infection, as well as an aging population [4]. *Pneumococci* are also well-known for their inherent antibiotic resistance, principally affecting beta-lactam, macrolide and sulfonamide sensitivity, thereby, it results in treatment failures. In 2000, the World Health Organization (WHO) reported that 14.5 million cases of pneumococcal disease occurred, resulting in 826,000 deaths in toddler aged between 1 and 59 months [5]. Thus, the diseases caused by *Pneumococci* constitute a major global public health problem and it is being extensively studied at the genome level [6,7]. In recent decades, the increasing numbers of *Pneumococci* genomes being sequenced entails significant interest in comparing the genome of each strain with other strains. This provides insight into strain specific characteristics that may act as a significant role in virulence and antimicrobial resistance. Hence, there is an urgent need to develop a database exclusively for *S. pneumoniae*, which we have attempted in the form of *Streptococcus pneumoniae* Genome Database (SPGDB). SPGDB provides links to tools that facilitate the comparison between multiple genomes of *Pneumococci* and to recognize the genome components of medical importance. The database, SPGDB, provides a powerful, user-friendly interface to perform various Boolean searches or sequence based searches. The database is also interfaced with the genome map of pneumococcal strains and the available three-dimensional structures of pneumococcal proteins in Protein Data Bank (PDB) [8]. This facilitates the users to explore functionally active proteins involved in pneumococcal disease pathogenicity and these structures can also be exploited further for structural analysis based on the user requirements. Additionally, it provides detailed information on the characteristics, virulence factors, pathogenesis and laboratory diagnosis of *Pneumococci* for researchers.

2. System design and implementation

The SPGDB database has been developed and hosted on Solaris server and is powered by 2.66 GHz Xeon (R) processor with 4 GB FDIMM main memory. The Solaris server was especially selected for its adaptability, scalability and security. The entire data of SPGDB were stored and managed in MySQL relational database. The search engine was written using PERL/CGI and PERL/DBI modules. The front-end input data part was coded in HTML, JavaScript and Ajax allows user-friendly web forms. The complete genome of pneumococcal strains available in NCBI genome

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

http://dx.doi.org/10.1016/j.ygeno.2014.09.012
0888-7543/

Table 1
The number of occurrences of com-box in each strain of *Pneumococci* identified by 'DNA Motif search tool'.

Strain name	Number of occurrences
Streptococcus pneumoniae 670-6B	12
Streptococcus pneumoniae 70585	10
Streptococcus pneumoniae A026	9
Streptococcus pneumoniae AP200	9
Streptococcus pneumoniae ATCC 700669	8
Streptococcus pneumoniae CGSP14	13
Streptococcus pneumoniae D39	7
Streptococcus pneumoniae G54	8
Streptococcus pneumoniae gamPNI0373	9
Streptococcus pneumoniae Hungary19A-6	9
Streptococcus pneumoniae INV104	4
Streptococcus pneumoniae INV200	7
Streptococcus pneumoniae JJA	7
Streptococcus pneumoniae OXC141	6
Streptococcus pneumoniae P1031	8
Streptococcus pneumoniae R6	7
Streptococcus pneumoniae SPN034156	4
Streptococcus pneumoniae SPN034183	5
Streptococcus pneumoniae SPN994038	5
Streptococcus pneumoniae SPN994039	5
Streptococcus pneumoniae SPNA45	5
Streptococcus pneumoniae ST556	9
Streptococcus pneumoniae Taiwan19F-14	8
Streptococcus pneumoniae TCH8431/19A	9
Streptococcus pneumoniae TIGR4	10

database [9] FTP site were obtained in Genome Feature Format (GFF3) and FASTA format and loaded into GBrowse. The database has been tested on multiple platforms (Windows, Linux and Solaris) with different web browsers. For better view of the database, the user can use recent versions of web-browsers. The database has been thoroughly validated and produces the results quickly; however, it may vary depending on the user network speed.

3. Complex, user-friendly search options

The SPGDB database provides a powerful and user-friendly search engine. All annotations may be explored utilizing either a simple or advanced Boolean-based search tools. In simple search, the user can browse for different strains, genes and proteins of entire *Pneumococci* by entering strain/gene/protein name in the text box, respectively. The advance search has options to return list of proteins, localizing to a particular chamber. To serve downstream system level analysis, SPGDB enables searching of proteins by COG category and pattern/profile. Further, it facilitates the user to retrieve proteins based on status and virulent genes, which was manually curated from various PUBMED literatures.

4. Facilitating sequence based DNA motif and BLAST searches

The DNA sequence motifs with biological function have become progressively necessary for the analysis of gene regulation [10] and they are found non-randomly in the genome [11]. We have provided a search tool in SPGDB that can be used to identify user-specified DNA motifs like putative transcription factor binding sites and other interesting motifs within the pneumococcal strains. This search tool accepts an IUPAC formatted stretch of DNA sequence with varying lengths and converts the input sequence into a regular expression. Another tool, BLAST [12,13] is also interfaced in SPGDB with which sequence similarity searches can be performed for both protein and nucleotide sequences against a specific or entire pneumococcal strains. The BLAST tool allows users to set parameters like word size, gap open, extension penalty and substitution matrix. The results produced by both DNA motif and BLAST tool can

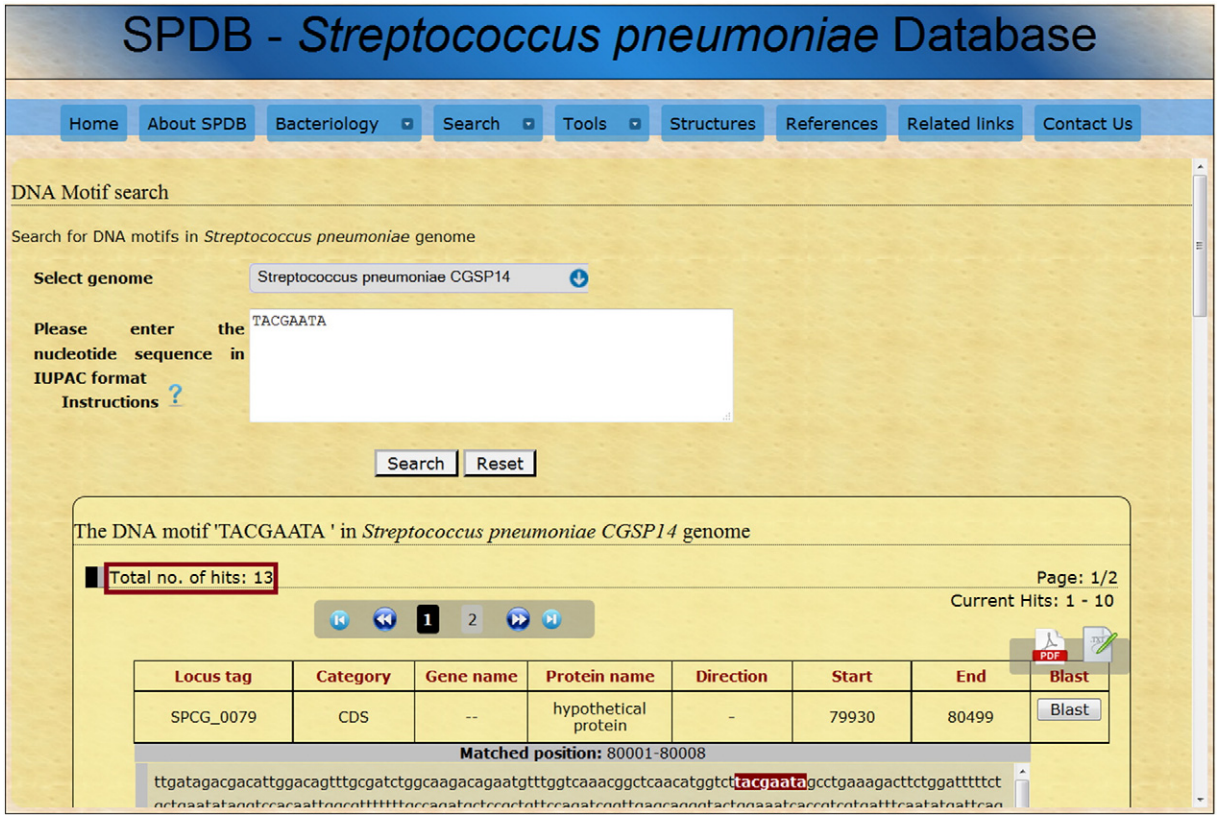

Fig. 1. The results of a com-box (TACGA{2}TA) searched in 'DNA motif search tool' against *Streptococcus pneumoniae* CGSP14.

Fig. 2. A. The genes, proteins, GC content, 3-frame translation and 6-frame translation tracks of *Streptococcus pneumoniae* TIGR from 1,002,689 to 1,072,420 in GBrowse. B: The PsaA gene in *Streptococcus pneumoniae* G54 visualized in GBrowse.

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

The com-box (also referred to as cin-box) is a sequence comprising of eight bases, TACGAATA. This sequence acts as a typical binding site for sigma factor, a bacterial transcription initiation factor [14]. The sequence of interest has been searched through the SPGDB DNA motif search tool against all pneumococcal strains and the results are shown in Table 1. The *S. pneumoniae* CGSP14 genome has the highest number of occurrences (13) of com-box compared to all complete genome of pneumococcal strains (Fig. 1). This is just one example of how integration of this tool can led to new insights through the pneumococcal genome analysis.

5. Genome sequences utilizing GBrowse

In recent years, the genome content of *Pneumococci* has increased drastically. It has to be accessible to researchers for easy interpretation with the help of feasible and interactive viewer. To expedite this, a platform-independent web based application, Generic Genome Browser (GBrowse) has been incorporated in SPGDB. GBrowse was developed by Stein et al. [15] of the Generic Model Organism System Database Project

(GMOD). The browser has the 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. It displays 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 on SPGDB database or NCBI [9]. Thus, the SPGDB GBrowse makes the end-user to easily view the genomic content of different strains of *Pneumococci*.

5.1. Case study

Among the different pneumococcal strains, capsular serotype 4 clinical isolate of *S. pneumoniae* designated as TIGR4 (TIGR — The Institute of Genome Research) is highly virulent and invasive. The studies reported that the genome comprises of 2,160,837 base pairs with 39.7% of GC content [16]. The genome of this strain is visualized using GBrowse. Fig. 2A shows the various track of genome from the position 1,002,689 to 1,072,420 where the GC content is notably high.

The gene PsaA (Pneumococcal Surface Adhesin) was recognized as a potential adhesin and virulence factor in *Pneumococci*. It is a member of lipoprotein receptor-associated antigen I (LraI) family [17]. In GBrowse, the PsaA gene (locus tag: SPG_1560) of *S. pneumoniae G54* is searched and it is found to be positioned from 123,363 to 125,492. It is observed that the transcription of the gene occurs in 5′–3′ direction. In addition, the page displays landmark of its corresponding protein, manganese ABC transporter substrate-binding lipoprotein (Fig. 2B). When the track is zoom out, the adjacent genes are found to be PsaC (locus tag: SPG_1559) and tpx (locus tag: SPG_1561).

6. Other utilities of SPGDB

As of August 20, 2014, 464 three dimensional structures of pneumococcal proteins are available in Protein Data Bank (PDB) [8]. Due to increasing number of protein structures day by day, it is also necessary to include these structures in SPGDB database. This can be useful to the scientific community working on *Pneumococci* for the analysis of functionally active proteins. These structures are visualized using the interactive graphics JAVA based plug-in Jmol. Fig. 3 shows an example of Jmol viewer displaying the three-dimensional structure of an enzyme, phosphomevalonate kinase (PDB id: 1K47) [18].

The genome map, a pictorial representation of genomic sequence data and its bioinformatics analysis, are downloaded for available pneumococcal strains from the Genome Atlas Database [19]. The option to

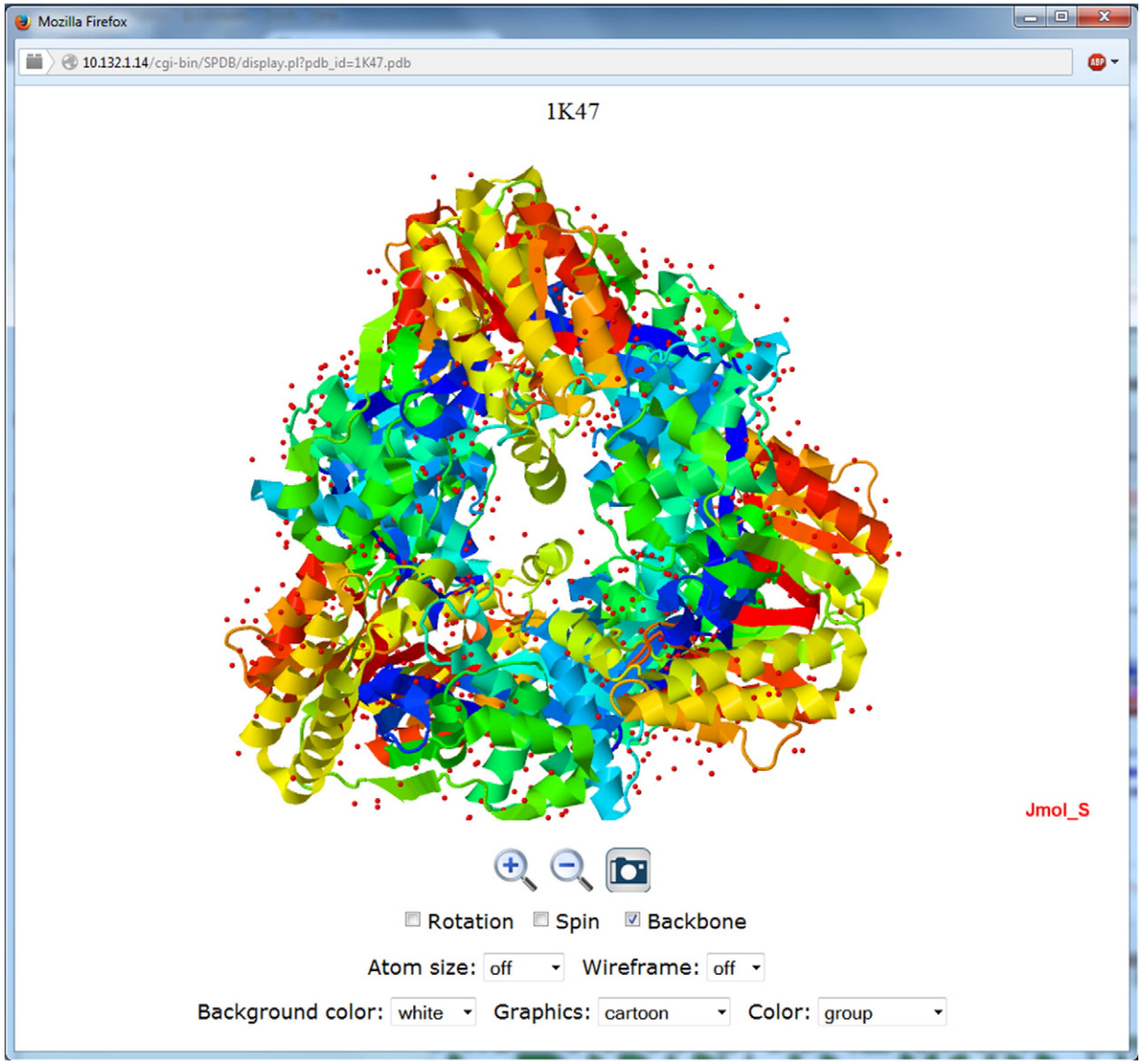

Fig. 3. The Jmol view of a three-dimensional crystal structure of phosphomevalonate kinase (PDB id: 1K47).

view genome maps is incorporated in the database under 'Search' menu. Additionally, the information on cultural characteristics, virulence factors, pathogenesis and laboratory diagnosis of *Pneumococci* has been provided under 'Bacteriology' menu to render preliminary knowledge about the bacterium. The links for different resources related to *Pneumococci* are provided in the 'Related links' menu.

7. Conclusion

In summary, SPGDB database has been developed to deliver complete genomic and proteomic information of *Pneumococci* to the research community. By providing the dynamic search and browse tools, the database aims to act as not only an integrated resource for *Pneumococci*, but also as a versatile application platform for the genomic and comparative study of *Pneumococci* strains. The database will be updated periodically.

Acknowledgments

SR and AA gratefully acknowledge the Indian Council of Medical Research (ICMR) for the research grant IRIS ID: 2014-0099. RGS thank ICMR for the Senior Research Fellowship. The authors also thank the management of VIT University for their support. DKKS, EDD, ZZA and KS acknowledge the facilities offered by the Indian Institute of Science, Bangalore.

References

[1] K.L. O'Brien, et al., Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates, Lancet 374 (2009) 893–902.
[2] T. Barichello, et al., Pathophysiology of acute meningitis caused by *Streptococcus pneumoniae* and adjunctive therapy approaches, Arq. Neuropsiquiatr. 70 (2012) 366–372.
[3] R. Pallares, et al., The epidemiology of antibiotic resistance in *Streptococcus pneumoniae* and the clinical relevance of resistance to cephalosporins, macrolides and quinolones, Int. J. Antimicrob. Agents 22 (2003) S15–S24.
[4] F. Blasi, et al., Understanding the burden of pneumococcal disease in adults, Clin. Microbiol. Infect. 18 (2012) 7–14.
[5] The World Health Organization, Measuring Impact of *Streptococcus pneumoniae* and *Haemophilus influenzae* Type b Conjugate Vaccination, WHO/IVB/12.0, 2012. (www.who.int/vaccines-documents/).
[6] R.N. Jones, et al., Evolving trends in *Streptococcus pneumoniae* resistance: implications for therapy of community-acquired bacterial pneumonia, Int. J. Antimicrob. Agents 36 (2010) 197–204.
[7] E.S. Honsa, et al., The roles of transition metals in the physiology and pathogenesis of *Streptococcus pneumoniae*, Front. Cell. Infect. Microbiol. 3 (2013) 92.
[8] H.M. Berman, et al., The Protein Data Bank, Acta Crystallogr. D Biol. Crystallogr. 58 (2002) 899–907.
[9] D.L. Wheeler, et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 35 (2007) D5–D12.
[10] P. D'haeseleer, What are DNA sequence motifs? Nat. Biotechnol. 24 (2006) 423–425.
[11] D. Halpern, et al., Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling, PLoS Genet. 3 (2007) 1614–1621.
[12] S.F. Altschul, et al., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[13] M. Uthayakumar, et al., BSSB: BLAST Server for Structural Biologists, J. Appl. Crystallogr. 44 (2011) 651–654.
[14] A. Dagkessamanskaia, et al., Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of ciaR mutant cells, Mol. Microbiol. 51 (2004) 1071–1086.
[15] L.D. Stein, et al., The generic genome browser: a building block for a model organism system database, Genome Res. 12 (2002) 1599–1610.
[16] H. Tettelin, Complete genome sequence of a virulent isolate of Streptococcus pneumoniae, Science 293 (2001) 498–506.
[17] J.W. Johnston, et al., Lipoprotein PsaA in virulence of *Streptococcus pneumoniae*: surface accessibility and role in protection from superoxide, Infect. Immun. 72 (2004) 5858–5867.
[18] M.J. Romanowski, et al., Crystal structure of the *Streptococcus pneumoniae* phosphomevalonate kinase, a member of the GHMP kinase superfamily, Proteins 47 (2002) 568–571.
[19] P.F. Hallin, D.W. Ussery, Genome Atlas Database: a dynamic storage for bioinformatics results and sequence data, Bioinformatics 20 (2004) 3682–3686.