Computers in Biology and Medicine 55 (2014) 86–91

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/

[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

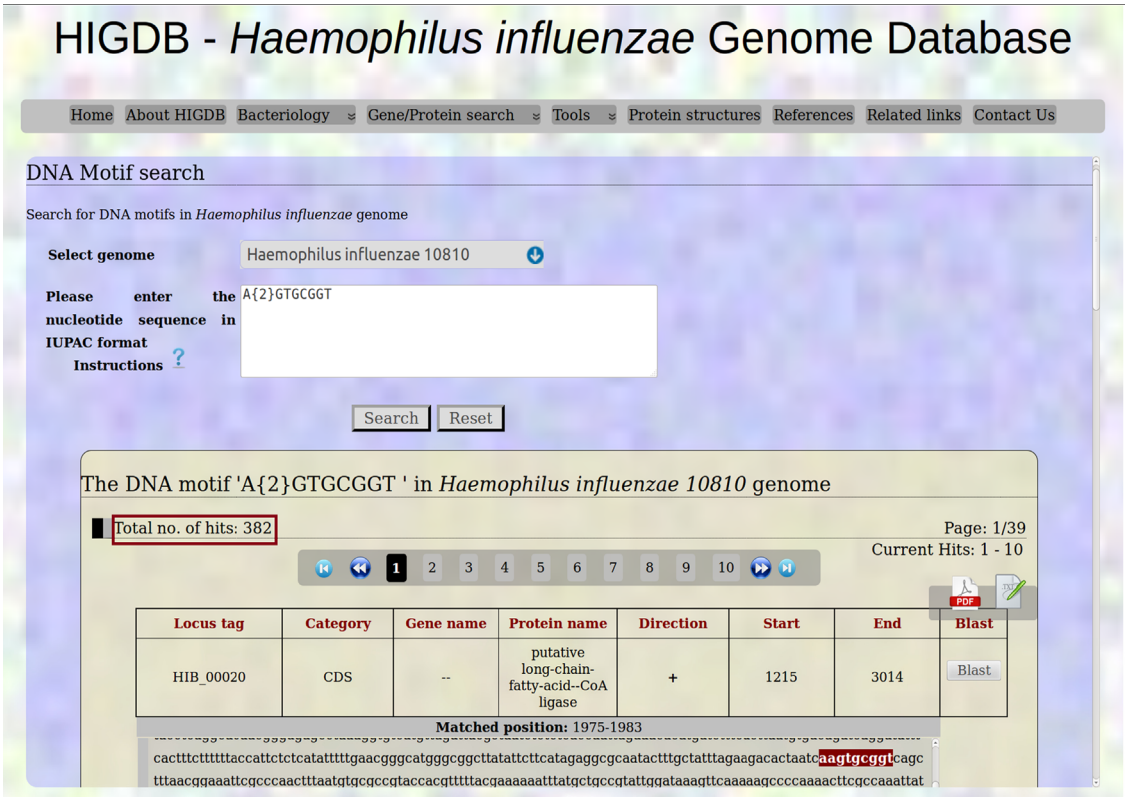

Fig. 1. The results of a motif "A{2}GTGCGGT" searched in 'DNA motif search tool' against *Haemophilus influenzae* 10810.

capsule. The genome comprises of 1830,138 base pair with 38.2% of GC content [26,27]. The genome of this strain is visualized in GBrowse. Fig. 2A shows the various tracks in the genome browser from the position 937,849 to 1036,694 where the GC content is notably high.

H. influenzae secretes immunoglobulin A1 (IgA1) proteases which is recognised as an important virulence factor in causing meningitis. IgA1 protease is an extracellular bacterial enzymes and it cleaves the hinge region of human IgA1's heavy chain [28]. In GBrowse, the IgA1 (locus tag: HI0990) of *H. influenzae* Rd KW20 has been searched and it is found to be positioned from 1048,666 to 1053,241. Interestingly, it is observed that the IgA1 gene is encoded on the reverse strand. When the gene track is zoomed out, the adjacent genes are found to be leuD (locus tag: HI0989) and recF (locus tag: HI0991). In addition, GBrowse displays a track of its corresponding protein (Fig. 2B).

6. Other utilities of HIGDB

As of September 2, 2014, 284 three dimensional structures of *H. influenzae* proteins were available in PDB [14]. Due to the increasing number of protein structures in PDB and to study the functionally active proteins structure of *H. influenzae*, it is necessary to include these structures in HIGDB database. This can be helpful to the scientific community working on *H. influenzae* for the structural analysis of proteins. These structures can be 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, aspartate-semialdehyde dehydrogenase (PDB ID: 1Q2X) [29].

The genome map, a pictorial representation of genomic sequence data and its bioinformatics analysis, were downloaded for available *H. influenzae* strains from the Genome Atlas Database [30]. These genome maps were incorporated in HIGDB under 'Search' menu. The links for varied resources related to *H. influenzae* are provided in the 'Related links' menu. Additionally, from various literatures, the information on cultural characteristics, virulence factors, pathogenesis and laboratory diagnosis of *H. influenzae* has been collected and included under 'Bacteriology' menu. This information helps the users to acquire the preliminary knowledge on *H. influenzae* and the several diseases caused by the organism.

7. Conclusion

The comparative study of *H. influenzae* genomes provides insights into strain specific characteristics that may act as a significant role in virulence and antimicrobial resistance. Hence, we developed a database exclusively for *H. influenzae*. The HIGDB database provides flexible, user-friendly interface and tools such as BLAST and DNA motif search to facilitate the comparative study of *H. influenzae* genomes. These analyses can be exploited for determining putative essential and core *H. influenzae* genes. The database also delivers the complete genomic and proteomic information of *H. influenzae* to the research community. The database aims to act as an integrated resource for *H. influenzae* and we strongly believe that HIGDB will be useful for the researchers studying this species. The HIGDB will be updated on a periodic basis.

Summary

Haemophilus influenzae is a respiratory tract commensal and it is considered as an invasive pathogen. It causes respiratory tract infections, community-acquired pneumonia, bacteremia and meningitis. *H. influenzae* associated diseases remain a serious problem in both developed and developing countries and cause significant mortality and morbidity worldwide. The organism is well-known for their inherent resistance to antibiotics and it is necessary to

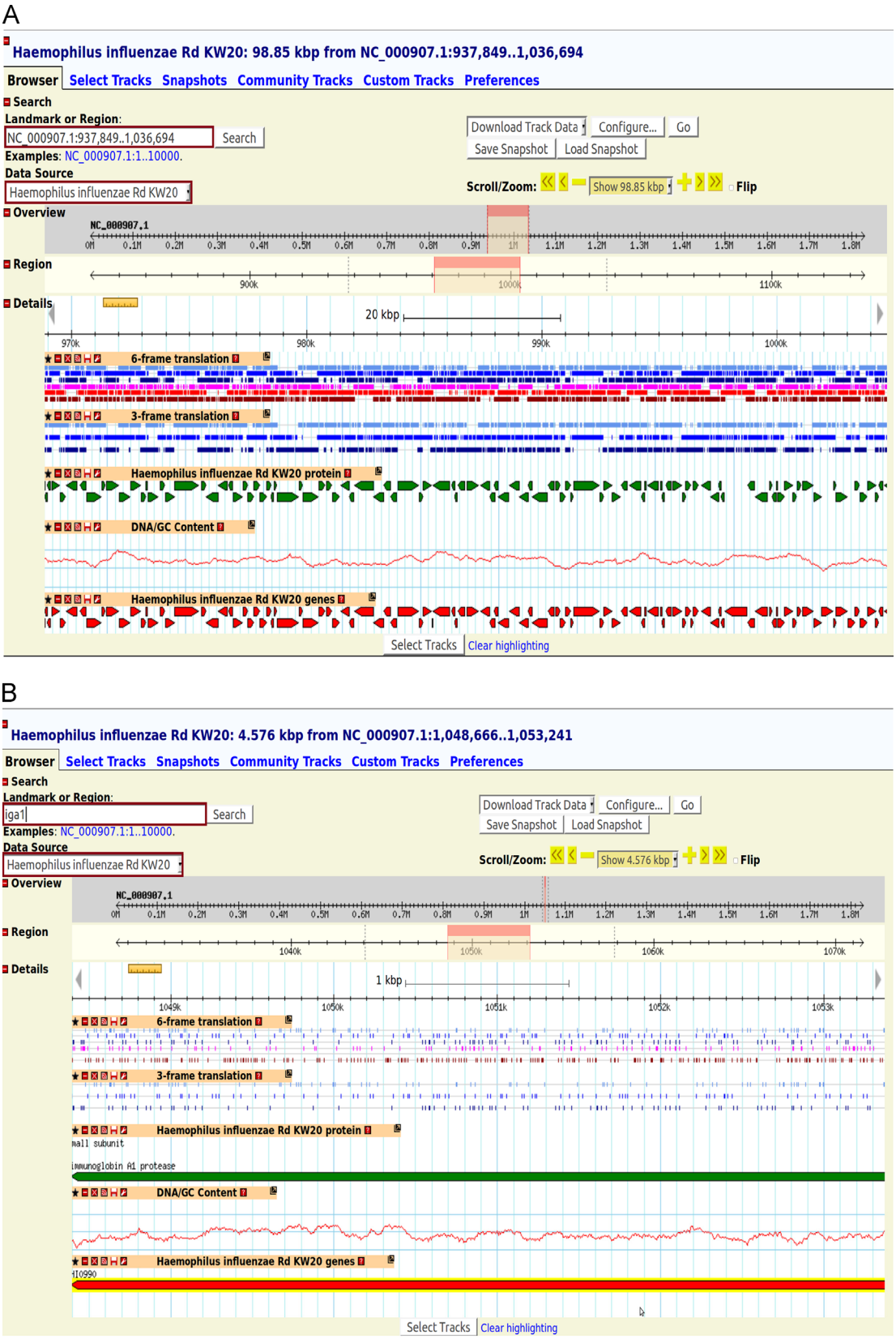

Fig. 2. (A) The genes, proteins, GC content, 3-frame translation and 6-frame translation tracks of *Haemophilus influenzae* Rd KW20 from 937,849 to 1036,694 in GBrowse. (B) The IgA1 gene of *Haemophilus influenzae* Rd KW20 visualized in G Browse.

study the organism at the genome level. A better understanding of the genes responsible for virulence in these strains and comparing *H. influenzae* genomes with each other can help researchers to determine immunogen targets and drug candidates. In view of above, we have developed a database, *Haemophilus influenzae* Genome Database (HIGDB). This database provides a powerful,

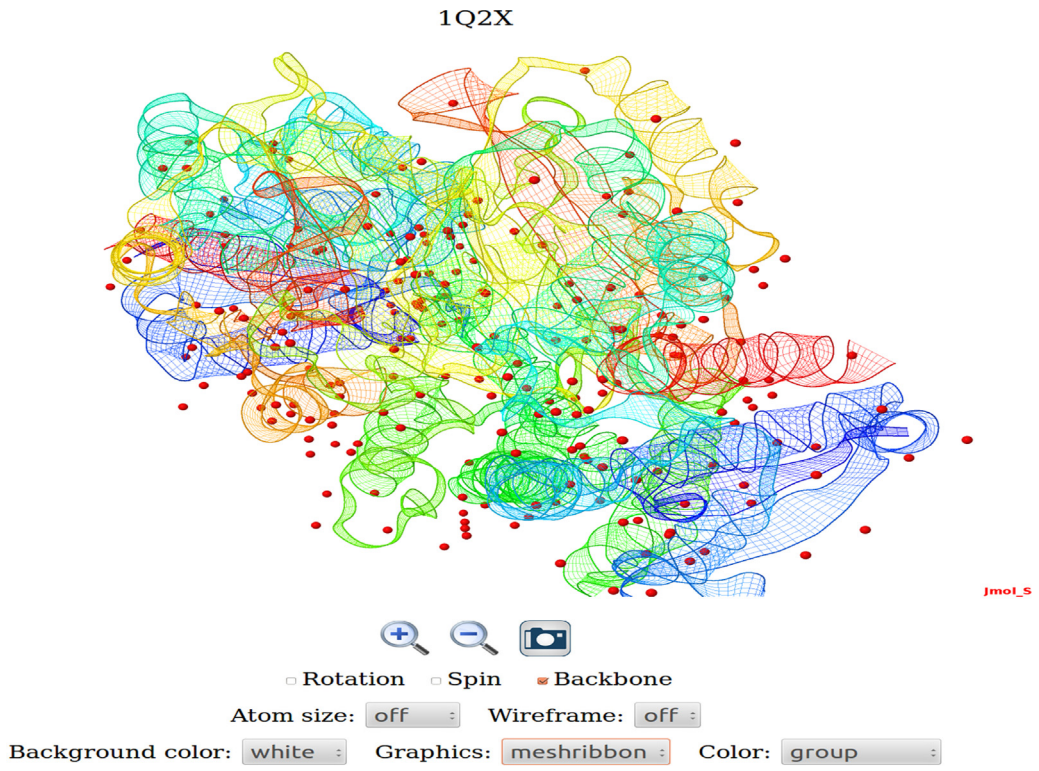

Fig. 3. The Jmol view of a three-dimensional crystal structure of aspartate-semialdehyde dehydrogenase (PDB ID: 1Q2X) in meshribbon graphics coloured by group.

user-friendly search and comparative analysis interface. It caters the genomic and proteomic information of complete *H. influenzae* strains to research community. We believe that our database will be useful for the analysis of phenotypic variation as well as to perform broad evolutionary studies. The database will be updated periodically and it is available through the URL: http://bioserver1.physics.iisc.ernet.in/HIGDB/.

Conflict of interest statement

None.

Acknowledgements

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

References

[1] S. Bae, J. Lee, J. Lee, E. Kim, S. Lee, J. Yu, Y. Kang, Antimicrobial resistance in *Haemophilus influenzae* respiratory tract isolates in Korea: results of a nation-wide acute respiratory infections surveillance, Antimicrob. Agents Chemother 54 (2010) 65–71.

[2] K.M. Kumar, P. Anitha, V. Sivasakthi, S. Bag, P. Lavanya, A. Anbarasu, S. Ramaiah, *In silico* study on Penicillin derivatives and Cephalosporins for upper respiratory tract bacterial pathogens 3, Biotech 4 (2014) 241–251.

[3] D. Kofteridis, G. Samonis, E. Mantadakis, S. Maraki, G. Chrysofakis, D. Alegakis, J. Papadakis, A. Gikas, D. Bouros, Lower respiratory tract infections caused by *Haemophilus influenzae*: clinical features and predictors of outcome, Med. Sci. Monit. 15 (2009) 135–139.

[4] F. Okada, Y. Ando, S. Tanoue, R. Ishii, S. Matsushita, A. Ono, T. Maeda, H. Mori, Radiological findings in acute *Haemophilus influenzae* pulmonary infection, Br. J. Radiol. 85 (2012) 121–126.

[5] A.S. Marshall, C.I. Barker, A.S. Pulickal, E. Kibwana, S.C. Gautam, E.A. Clutterbuck, S.M. Thorson, S. Shrestha, N. Adhikari, A.J. Pollard, D.F. Kelly, The seroepidemiology of *Haemophilus influenzae* type b prior to introduction of an immunization programme in Kathmandu, Nepal, PLoS One 9 (2014) e85055.

[6] W.J. Chen, L.H. Moulton, S.K. Saha, A.A. Mahmud, S.E. Arifeen, A.H. Baqui, Estimation of the herd protection of *Haemophilus influenzae* type b conjugate vaccine against radiologically confirmed pneumonia in children under 2 years old in Dhaka, Bangladesh, Vaccine 32 (2014) 944–948.

[7] A. Agarwal, T.F. Murphy, *Haemophilus influenzae* infections in the *H. influenzae* type b conjugate vaccine era, J. Clin. Microbiol. 49 (2011) 3728–3732.

[8] P. Mangtani, K. Mulholland, S.A. Madhi, K. Edmond, R. O'Loughlin, R. Hajjeh, *Haemophilus influenzae* type b disease in HIV-infected children: a review of the disease epidemiology and effectiveness of Hib conjugate vaccines, Vaccine 28 (2010) 1677–1683.

[9] A.S. Dajani, B.I. Asmar, M.C. Thirumoorthi, Systemic *Haemophilus influenzae* disease: an overview, J. Pediatr. 94 (1979) 355–364.

[10] D.W. Hood, M.E. Deadman, M.P. Jennings, M. Bisercic, R.D. Fleischmann J.C. Venter, E.R. Moxon, DNA repeats identify novel virulence genes in *Haemophilus influenzae*, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 11121–11125.

[11] S. Tristram, M.R. Jacobs, P.C. Appelbaum, Antimicrobial resistance in *Haemophilus influenzae*, Clin. Microbiol. Rev. 20 (2007) 368–389.

[12] World Health Organization. *Haemophilus influenzae* Type B (HiB). WHO fact, 294 (2005). ⟨http://www/who.int/mediacentre/factsheets/fs294/en/index.html⟩.

[13] H. Peltola, Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates, Clin. Microbiol. Rev. 13 (2000) 302–317.

[14] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, The protein data bank, Acta Crystallogr., Sect. D: Biol. Crystallogr 58 (2002) 899–907.

[15] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 35 (2007) D5–D12.

[16] P D'haeseleer, What are DNA sequence motifs? Nat. Biotechnol 24 (2006) 423–425.

[17] D. Halpern, H. Chiapello, S. Schbath, S. Robin, C. Hennequet-Antier, A. Gruss A, M El Karoui, Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling, PLoS Genet. 3 (2007) 1614–1621.

[18] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[19] M. Uthayakumar, G. Sowmiya, R. Sabarinathan, N. Udayaprakash, M. Kirti Vaishnavi, K. Sekar, BSSB: BLAST server for structural biologists, J. Appl. Crystallogr 44 (2011) 651–654.
[20] J.N. Varela, M.C. Amstalden, R.F. Pereira, L.M. de Hollanda, H.J. Ceragioli, V. Baranauskas, M. Lancellotti, *Haemophilus influenzae* porine ompP2 gene transfer mediated by grapheme oxide nanoparticles with effects on transformation process and virulence bacterial capacity, J. Nanobiotechnol. 12 (2014) 14.
[21] J.C. Mell, I.M. Hall, R.J. Redfield, Defining the DNA uptake specificity of naturally competent *Haemophilus influenzae* cells, Nucleic Acids Res. 40 (2012) 8536–8549.
[22] S.H. Goodgal, M.A. Mitchell, Sequence and uptake specificity of cloned sonicated fragments of *Haemophilus influenzae* DNA, J. Bacteriol. 172 (1990) 5924–5928.
[23] H.O. Smith, J.F. Tomb, B.A. Dougherty, R.D. Fleischmann, J.C. Venter, Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome, Science 269 (1995) 538–540.
[24] L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson J.E. Stajich, T.W. Harris, A. Arva, S. Lewis, The generic genome browser: a building block for a model organism system database, Genome Res. 12 (2002) 1599–1610.
[25] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496–512.
[26] K.W. Wilcox, H.O. Smith, Isolation and characterization of mutants of *Haemophilus influenzae* deficient in an adenosing 5′-triphosphate-dependent deoxyribonuclease activity, J. Bacteriol. 122 (1975) 443–453.
[27] D.A. Daines, L.A. Cohn, H.N. Coleman, K.S. Kim, A.L. Smith, *Haemophilus influenzae* Rd KW20 has virulence properties, J. Med. Microbiol. 52 (2003) 277–282.
[28] K. Poulsen, J. Brandt, J.P. Hjorth, H.C. Thogersen, M. Kilian, Cloning and sequencing of the immunoglobulin A1 protease gene (iga) of *Haemophilus influenzae* serotype b, Infect. Immun. 57 (1989) 3097–3105.
[29] J. Blanco, R.A. Moore, C.R. Faehnle, D.M. Coe, R.E. Viola, The role of substrate-binding groups in the mechanism of aspartate-beta-semialdehyde dehydrogenase, Acta Crystallogr., Sect. D: Biol. Crystallogr 60 (2004) 1388–1395.
[30] P.F. Hallin, D.W. Ussery, Genome Atlas Database: a dynamic storage for bioinformatics results and sequence data, Bioinformatics 20 (2004) 3682–3686.