

Database and Comparative Identification of Prophages

K.V. Srividhya¹, Geeta V Rao¹, Raghavenderan L¹, Preeti Mehta¹, Jaime Prilusky²,
Sankarnarayanan Manicka¹, Joel L. Sussman³, and S Krishnaswamy¹

¹ Bioinformatics centre, School of Biotechnology, Madurai Kamaraj University, Madurai
krishna@mrna.tn.nic.in

² Biological Services, Weizmann Institute of Science Rehovot 76100, Israel

³ Department of Structural Biology, Weizmann Institute of Science Rehovot 76100, Israel

Abstract. Prophages are integrated viral genomes in bacteria. Prophages are distinct from other genomic segments encoding virulence factors that have been acquired by horizontal gene transfer events. A database for prophages (<http://bicmku.in:8082/prophagedb> <http://ispc.weizmann.ac.il/prophagedb>) has been constructed with data available from literature reports. To date other than bacteriophage corner stone genes based iterative searches, no other exhaustive approach unique for identifying prophage elements is available. Here we report detection of prophages based on proteomic signature comparison using a prophage proteome as reference set. This method was tested with using the database and then extended over newly sequenced bacterial genomes with no reported prophages. The approach of using similarity of proteins over a given region helped identify twenty putative prophage regions in nine different bacterial genomes.

1 Introduction

Bacteriophages are viruses infecting bacteria. Bacteriophages take up two life cycles, one being lytic infects, multiplies, and lyses host bacterium during progeny release [1] whereas in the other temperate mode the phage DNA integrates with the bacterial genome and is termed as prophage. Prophages range from fully viable to cryptic. Cryptic prophages harbor mutational decay and do not result in lytic growth. Prophages can constitute as much as 10-20% of a bacterium's genome (*Escherichia coli* O157:H7 strain Sakai contains 18 prophage elements constituting 16% of the genome) [2]. Prophage sequences contribute to interstrain variability [3]. At present 230 prophages have been reported in 82 bacterial genomes [4]. In addition, prophages are important vehicles for horizontal gene exchange between different bacterial species. Virulence factors in many pathogenic bacteria are observed to be located on prophage locus, indicating the possible role played by prophages in conferring pathogenicity to host bacterium [5][6][7]. The impact of prophages on bacterial evolution has been reviewed extensively [8]. Prophages do not seem to be a homogenous group and show mosaic nature. Their diverse nature is also reflected by the diversity of genome sizes ranging from 549 kb (Flex9 prophage of *Shigella flexneri* 2a301) to 139449kb (Bh1 prophage from *Bacillus halodurans*). Analysis of e14 prophage of *E coli* K-12 revealed the modular nature of the element [9] and provided the basis for the approach of using similarity of proteins over a given region [10].

Identifying and understanding prophage elements is medically important as some phage genes are known to increase the survival fitness of lysogens [3]. Unambiguous detection of cryptic prophages is difficult as these defective prophages may be devoid of any corner stone genes. A prophage database has been initiated with the information available from the integrated prophages of sequenced bacterial genomes. At present, the database contains 227 prophage sequences from 49 organisms and details on the twenty putative prophage regions identified in 9 bacterial genomes by protein similarity approach.

2 Methods

The prophage database was constructed using PostgreSQL server at backend and PHP in front end. All genomes, proteomes and protein table files were downloaded from <ftp://ftp.ncbi.nih.gov/genomes/Bacteria>. Currently the database covers prophages reported by Casjens [4]. In order to identify e14 homologs, similarity searches at the protein level were done taking the twenty-three e14 proteins as query and the bacterial proteomes as target. Similarity searches were done using blastp, (local version of the WWW-BLAST) [11] using an e-value of 0.01 and Blosum62 as the scoring matrix.

3 Results and Discussion

3.1 Database for Prophage Elements

Prophage database covers genome data (GC content, integration site, location, genome size) and protein data (location, protein annotation, related COG information, PDB homologs and Unfoldability index by FoldIndex[®]). Fig 1 represents the screenshot of prophage database homepage.

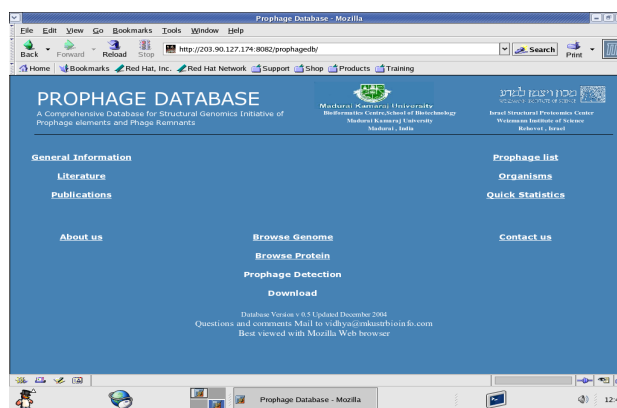


Fig. 1. Screenshot of Prophage Database home page

3.2 Identifying Prophage Elements in Bacterial Genomes

The e14 element is a very well characterized prophage element [9], which contains all the highly conserved prophage genes like the integrase, excisionase, phage portal, cro type regulator, repressor and terminase genes elements. Genes encoding the BLAST hits for the different e14 proteins, which were within a particular distance (this distance varies from one organism to another; it is the size of the longest prophage in the organism's genome) were then clubbed together. Any regions with two or more genes in this cluster were considered as putative prophage elements and further analyzed [10]. Fig 2 summarizes the method.

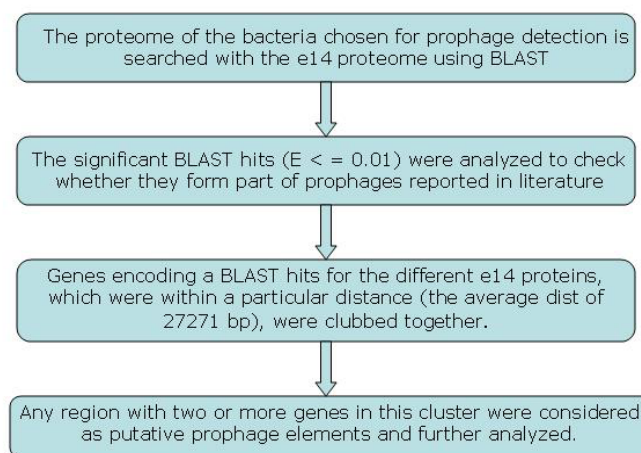


Fig. 2. Schematic representation of protein similarity method

A set of 61 bacterial genomes was chosen for prophage detection. To confirm the sensitivity of the approach bacterial genomes with prophage incorporated in prophage database data was used for comparison (BLAST hits ($e \leq 0.01$)). Out of 174 reported prophages, 87 loci were identified by protein similarity approach. Among them include 27 from *E coli* (*K-12* O157 VT-2 Sakai, CFT073, EDL933) out of 61 reported in literature. With *M. tuberculosis* 3 loci were detected among the reported 6 prophages. With *S.aureus* strains 5 could be identified among 7 reported, 10 amongst 19 in *Salmonella*, 4 out of 10 in *Yersinia*, 9 among 15 in *S.pyogenes*. Samples results reflecting 14 genome sequences is presented in Table 1.

Hence the method was further extended on to genomes with no reports of prophage. Out of 9 genomes, 24 loci were identified among which 9 were found to be highly probable prophage locations. Table 2 details the prophages identified and respective organisms. For the former, prophage regions were delimited using data from the prophage database and from literature [4]. It was observed that most putative prophage regions encode hypothetical proteins suggesting that these regions need to be characterized further. Interestingly among the newly identified prophages, five are

Table 1. Bacterial genomes and probable prophages identified using the Protein Similarity Approach method in comparison with literature reports

Organism	Literature reported	PSA detected	Prophages Detected
<i>B. subtilis</i> 168	5	2	PBSX, SKIN
<i>C. tetani</i> E88	3	2	Cpt2, Cpt3
<i>E. coli</i> K12	11	4	DLP12 ,QIN , Rac, KpLE1
<i>E. coli</i> O157 VT-2 Sakai	24	16	Sp8 , Sp9, Sp6 , Sp4, Sp14, Sp3, Sp15, Sp1,Sp5, Sp12, Sp11 Sp7,Sp10,Sp18,Sp16, Sp17
<i>H. influenzae</i> Rd KW20	3	1	FluMu
<i>M. loti</i> MAFF303099	3	1	Meso2
<i>M. tuberculosis</i> H37Rv	3	1	phiRv1
<i>N. meningitidis</i> Z2491	3	2	Pnm2 ,Pnm1
<i>S. aureus</i> N315	1	1	phiN315
<i>S. enterica</i> CT18 (serovar Typhi)	12	5	Sti4b, Sti8 , Sti3, Sti1,Sti7
<i>S. flexneri</i> 2a301	11	3	Flex9, Flex5, Flex2
<i>S. pyogenes</i> M1 SF370	4	2	370.2 , 370.1
<i>X. fastidiosa</i> 9a5c	9	1	XfP4
<i>Y. pestis</i> KIM	5	2	Yers3, Yers1

Table 2. Prophage regions detected using the PSA approach from six bacterial genomes. Indicated in # are genomes with no report of prophages.

Organism	Prophage	Gene products
<i>S. enterica</i> LT2 (serovar Typhimurium)	St1	Transposase, cytoplasmic proteins, phoQP
<i>S. flexneri</i> 2457T	Sf1	Integrase, replication protein, helicase , mating formation
<i>S. pyogenes</i> M18 MGAS8232	Sp1	Efflux, phage portal protein, transposase
<i>S. pyogenes</i> M3 MGAS315	Spy1	Transposase, antiterminator drug resistance protein
<i>P. luminescens</i> subsp. laumondii T101#	P11,P12, P13 P14, 16,P17	DNA Invertase HIN, Mostly hypothetical proteins
<i>Mycobacterium bovis</i> AF2122/97 #	4 putative prophages	Integrase, transposase. Transcriptional regulatory

located near dehydrogenase genes. *A priori* there seems to be no attributable reason to this tendency for the putative lambdoid phages to get integrated near a dehydrogenase gene in the bacterial genome. However, it must be noted that the search template e14 is also integrated at the isocitrate dehydrogenase gene in the *E. coli* K12 genome.

3 Conclusion

A prophage database has been constructed and used to devise a prophage identification approach using similarity of proteins over a given region. Prediction of prophage related areas in genomes is problematic due to low similarity between prophage genes and the mosaic nature. Five bacterial genomes for which no prophage has been reported in the literature were analyzed in detail. It was observed that most putative prophage regions encode hypothetical proteins suggesting that these regions need to be characterized further. The database provides information on prophages, cryptic prophages and phage remnants, providing effective and efficient way to access the prophage genomes. Prophage identification can be further extended over newly sequenced genomes and incorporated into the database.

Acknowledgments

We acknowledge the use of Bioinformatics centre facility funded by DBT, Govt of India, DBT and the Israel Ministry of Science and Technology (MOST) for INDO-ISRAEL project support, MOST's support for the Israel Structural Proteomics Center.

References

1. Brussow, H., Hendrix, R.: Phage Genomics: Small Is Beautiful. *Cell* 108 (2002) 13-16
2. Canchaya, C., Proux, C., Fournous, G., Bruttin, A., Brussow, H. : Prophage Genomics. *Microbiol Mol Biol Rev.* 67 (2003) 238–276
3. Brussow, H., Canchaya, C., Hardt, W.D.: Phages And The Evolution Of Bacterial Pathogens: From Genomic Rearrangements To Lysogenic Conversion. *Microbiol Mol Biol Rev.* 68 (2004) 560-602
4. Casjens, S.: Prophages And Bacterial Genomics: What Have We Learned So Far? *Mol. Microbiol.* 49 (2003) 277-300
5. Waldor, M, K.: Bacteriophage Biology And Bacterial Virulence. *Trends Microbiol*, 6 (1998) 295-297
6. Davis, B.M., Waldor, and M.K.: Filamentous Phages Linked To Virulence Of *Vibrio cholerae*. *Curr Opin Microbiol*, 36 (2000) 35-42
7. Boyd, E.F., Brussow, H.: Common Themes Among Bacteriophage-Encoded Virulence Factors And Diversity Among The Bacteriophages Involved. *Trends Microbiol*, 10 (2002) 521-529
8. Canchaya, C., Fournous, G., Brussow, H.: The Impact Of Prophages On Bacterial Chromosomes. *Mol Microbiol.* 53 (2004) 9-18

9. Mehta, P., Casjens, S., Krishnaswamy, S.: Analysis Of The Lambdoid Prophage Element *e14* In The *E.Coli* K12 Genome. *BMC Microbiol*, 4 (2004) 1
10. Rao, G.V., Mehta, P., Srividhya, K.V., Krishnaswamy, S.: A Protein Similarity Approach For Detecting Prophage Regions In Bacterial Genomes. *Genome Biology*, 6, (2005) P11 (<http://genomebiology.com/2005/6/10/P11>)
11. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (1990) 403-410