Journal of Computer Science & Systems Biology - Open Access
Research Article JCSB/Vol.1 2008

Comparison of the Virulence Factors and Analysis of Hypothetical Sequences of the Strains TIGR4, D39, G54 and R6 of *Streptococcus Pneumoniae*

R. Jothi, S. Parthasarathy* and K. Ganesan

Department of Bioinformatics, School of Life Sciences
Bharathidasan University, Tiruchirappalli 620 024,
Tamil Nadu, India

*Corresponding author : S. Parthasarathy, Department of Bioinformatics,
School of Life Sciences, Bharathidasan University,
Tiruchirappalli 620 024,Tamil Nadu, India,
Tel: +91 94435 33095; Fax: +91 431 2407045; E-mail: bdupartha@gmail.com

Received October 21, 2008; Accepted November 19, 2008; Published December 26, 2008

Citation: Jothi R, Parthasarathy S, Ganesan K (2008) Comparison of the Virulence Factors and Analysis of Hypothetical Sequences of the Strains TIGR4, D39, G54 and R6 of *Streptococcus Pneumoniae* . J Comput Sci Syst Biol 1: 103-118. doi:10.4172/jcsb.1000010

Abstract

Whole genome sequences of the four strains of S*treptococcus pneumoniae,* encapsulated TIGR4, D39, G54 and nonencapsulated R6 are considered for the comparative study on genome features, whole genome pairwise alignment, gene role category, and virulence factors using relevant comparative genomics tools. The study of capsular polysaccharide synthesizing genes reveals that many cps genes are unique to TIGR4, which shows the high virulence nature of TIGR4. Further, the study on the other virulence factors such as pneumococcal surface protein A, autolysin, hyaluronate lyase, pneumolysin, neuraminidase B, and pneumococcal surface antigen A of TIGR4 are much related to those of the other three strains, and hence the virulence nature due to these factors among four strains seems to be similar. But it differs from neuraminidase A, choline binding protein A and immunoglobulin A1 protease. Also in the present study, 4 and 22 hypothetical protein sequences of TIGR4 and R6 respectively are predicted as virulence factors. Among those sequences, it is found that 8 hypothetical protein sequences with 7 different functional regions of R6 are related to other previously known virulence factors of TIGR4 and R6 of *S. pneumoniae.*

Keywords: Comparative genomics; *Streptococcus pneumoniae*; TIGR4; D39; G54; R6; Virulence factors; Hypothetical protein sequences

Abbreviations: CMR: Comprehensive Microbial Resource; CPS: Capsular Polysaccharide; PspA: Pneumococcal surface protein A; LytA: autolysin; Hyl: Hyaluronate lyase; Ply: Pneumolysin; NanA and NanB; Neuraminidases A and B; CbpA: Choline binding protein A; PsaA: Pneumococcal surface antigen A; IgA1: Immunoglobulin A1 protease

Introduction

The whole genome sequences of bacteria of closely related species or strains are providing new avenues of investigation for the further understanding of microbial diversity, pathogenesis, host-parasite interaction, evolution, etc. through a comparative analysis of their genomes. *Streptococcus pneumoniae*, commonly *pneumococcus* (Dowson, 2004; Gregory and DeSalle, 2005), a human pathogen, causes life threatening diseases like pneumoniae, bacteremia, meningitis, sepsis, and otitis media. Genome sequencing of four *S. pneumoniae* strains, namely, TIGR4, D39, G54 and R6 have been completed and genome sequencing of other 14 strains are ongoing. G54 genome sequence is not yet added in GenBank but it is inbuilt in Comprehensive Microbial Resource (CMR) and D39 genome sequence is available in GenBank but not in CMR. TIGR4, a clinical isolate, is encapsulated and highly virulent and many of its virulence fac-

tors have been studied (Tettelin et al., 2001). D39, the encapsulated and virulent strain (Lanie et al., 2007), was used by Avery, Macleod, and McCarty (Avery et al., 1979) in their landmark study on the role of DNA as the genetic material. G54 is an encapsulated clinical strain type 19F (Dopazo et al., 2001). R6, a derivative of the serotype 2 clinical isolate D39, is nonencapsulated and avirulent. The genes encoding many virulence factors are present in R6 genome in addition to the genes of capsular biosynthesis (Hoskins et al., 2001).

Many types of comparative studies (Tettelin et al., 2001; Lanie et al., 2007; Hoskins et al., 2001; AlonsoDeVelasco et al., 1995; Brückner et al., 2004; Ferretti et al., 2004; Silva et al., 2006) have already been carried out in *Streptococcus* strains on various aspects. The preliminary comparative analysis (Jothi et al., 2007) of the whole genomes of both the encapsulated TIGR4 and nonencapsulated R6 strains of *S. pneumoniae* provided some insights into the high virulence nature of TIGR4. This present study summarizes specifically how the whole genomes of the four strains, namely, TIGR4, D39, G54 and R6 of *S. pneumoniae* differ from each other by their genome features, genome diversity, gene role category and virulence factors. Comparison of the virulence factors among these strains can provide further insight into any strain uniqueness with relevance to virulence nature and can stimulate new approaches into disease prevention and treatment.

S. pneumoniae has two surface layers outside the plasma membrane, namely, cell wall and capsule. The cell wall has triple-layered peptidoglycan that holds the capsular and cell wall polysaccharides, and also few proteins. The capsule completely covers the inner structure of *S. pneumoniae*. The cell wall polysaccharide is common to all serotypes of *S. pneumoniae*, but the chemical structure of the capsular polysaccharide is serotype-specific (AlonsoDeVelasco et al., 1995). After Avery's experiment (Avery et al., 1979), the capsule has long been recognized as the major virulence factor of *S. pneumoniae*. Experimental proof for this was provided by the difference in 50% lethal dose between encapsulated and nonencapsulated strains. Encapsulated strains were found (AlonsoDeVelasco et al., 1995) to be at least 10^5 times more virulent than strains lacking the capsule. Certain proteins in *S. pneumoniae* like pneumococcal surface protein A (PspA), autolysin (LytA), hyaluronate lyase (Hyl), pneumolysin (Ply), neuraminidases A and B (NanA and NanB), choline binding protein A (CbpA), pneumococcal surface antigen A (PsaA) and immunoglobulin A1 (IgA1) protease are important virulence factors (AlonsoDeVelasco et al., 1995; Jedrzejas, 2001; Rigden et al., 2003) and these could be used as potential vaccine candidates. The preliminary identification of the surface proteins and virulence factors of *S. pneumoniae* were done by computational analysis of its genome sequences (Tettelin and Hollingshead, 2004; Gregory and DeSalle, 2005; Tettelin et al., 2001; Hoskins et al., 2001) and continued in several subsequent studies (Brückner et al., 2004; Polissi et al., 1998; Wizemann et al., 2001). Strains of *S. pneumoniae* are now resistant to commonly prescribed antibiotics, such as, penicillin, macrolides and fluoroquinolones (Tettelin et al., 2001). Because of the multidrug resistance nature of the *S. pneumoniae* strains, we need a deeper understanding of the virulence factors, for that the comparative genomics approach may provide more insight.

At present, only 70 % of the genes in any given genome can be predicted with reasonable confidence (Bork, 2000). The remaining genes are either hypothetical (do not have any known homolog) or conserved hypothetical (homologous to genes of unknown function), because it is unclear whether they encode actual proteins. The large quantity of hypothetical protein sequences in completely sequenced genomes of organisms makes their study an enormous task. Characterization of these genes or proteins of unknown function is generally recognized as an essential step towards fully understanding the biology of the pathogenic organism and for potential targets. Few studies (Galperin and Koonin, 2004; Brown, 2005; Sivashankari and Shanmughavel, 2006) have already been carried out on hypothetical sequences. In the present study, hypothetical protein sequences of the strains TIGR4 and R6 of *S. pneumoniae* are analyzed to find their virulence nature using VirulentPred. Among those sequences, it is also analyzed how far the hypothetical protein sequences are related to other previously known virulence factors of TIGR4 and R6 of *S. pneumoniae.*

Materials and methods

Various analysis of the whole genomes of the four strains, namely, TIGR4, D39, G54 and R6 of *S. pneumoniae* like the whole genome alignment, comparison of gene role categories, finding the location of the virulence factors in the genome and comparison of virulence regions are carried out using the appropriate bioinformatics software tools.

Sequence Retrieval and Whole Genome Pairwise Alignment

The complete genome sequences and the list of annotated gene and protein sequences of TIGR4, D39 and R6 are retrieved from the NCBI – FTP server (ftp://ftp.ncbi.nih.gov/genomes). We used the run-mummer3 program available in the standalone MUMmer 3.20 (http://

mummer.sourceforge.net/) and its built-in mummerplot for obtaining the whole genome pairwise alignment of *S. pneumoniae* strains TIGR4, D39, and R6 in different combinations. MUMmer at Comprehensive Microbial Resource (CMR) is used for the whole genome pairwise alignment of the strains TIGR4, G54 and R6 in different combinations.

Comparison of the Role Category of Genes and Sequence Analysis

The tool in CMR database (http://cmr.tigr.org/tigr-scripts/CMR/ CmrHomePage.cgi), the role category piechart is used for the genome features and functional role category comparison of the whole genomes of TIGR4, G54 and R6. Bacterial Annotation System (BASys - http://wishart.biology.ualberta.ca/basys) - A web server for automated bacterial genome annotation is used to know the role category for three strains TIGR4, D39 and R6, whose whole genomes are already available in it. From the prediction server of the Center for Biological Sequence Analysis (CBS - http://www.cbs.dtu.dk/services), the Genome Atlas is used for the analysis of repeats of *S. pneumoniae*. The sequences of various virulence factors, which are taken for our study, have been verified by using the virulence factors database (http://www.mgc.ac.cn/VFs). BioEdit (http://www.mbio.ncsu.edu/ BioEdit/bioedit.html) is used to compute sequence composition of the genomes and genes. Further, LALIGN (http://www.ch.embnet.org/software/LALIGN_form.html) is used for the pairwise global alignment of the gene sequences of the strains of *S. pneumoniae*.

Functional Annotation of Hypothetical Sequences

VirulentPred (http://bioinfo.icgeb.res.in/virulent) is a SVM (Support Vector Machine) based method to predict bacterial virulent protein sequences, which can be used to screen virulent proteins in proteomes. In the present study the above tool is used to analyse the hypothetical sequences of the strains TIGR4 and R6 of *S. pneumoniae*. From the proteome of TIRG4 and R6 of *S. pneumoniae,* all unannotated hypothetical protein sequences are retrieved using PERL script and those sequences are used as data set for virulence factor prediction.

Results and Discussion

Comparative genomics and *in silico* studies have begun to reveal insights into gene and protein functions of many organisms. Here, we compare the genomes of the strains TIGR4, D39, G54 and R6 of *S. pneumoniae* using the appropriate tools for whole genome comparison and the results are discussed below.

Comparison of the Genome Features of Four Strains of *S. Pneumoniae*

Table 1 summarizes the general information about the genomes including statistics of genes of these four strains, obtained and compiled from CMR and NCBI web servers. The genome sizes of these four strains range between 2 Mb and 2.16 Mb (c.f. Sl.No.2 of Table1). Among these four strains, D39 is the smallest and TIGR4 is the largest based on genome size. The nucleotide base (A, T, G, C, AT and GC) compositions of four strains show that the strains have low GC (~40%) genomes. The number of genes encoding for proteins of these four strains ranges between 1914 and 2234 (c.f. Sl.No.3 of Table1). Of the total base pairs of four genomes, approximately 85 - 87% of base pairs (bps) are involved in coding and the remaining are non-coding or junk DNA. The number of genes involved in RNA synthesis (structural RNA, tRNA, and rRNA) is more or less similar in all strains. Finally, by comparing the global and local repeats of TIGR4 and R6 using CBS web server, it is evident that both the repeats are high in TIGR4 than in R6 (c.f. Sl.No.4 of Table1) and this may be related to the duplicated regions of the chromosome (Gregory and DeSalle, 2005).

Comparison of Whole Genome Pairwise Alignments

The whole genome pairwise alignments of the strains TIGR4, D39 and R6 of *S. pneumoniae* (whose sequence data are available at NCBI) are obtained using the standalone version of MUMmer and the results are plotted using its built-in mummerplot. The whole genome pairwise alignments of the strains TIGR4, G54 and R6 are obtained using CMR, where these sequences are available, and the five possible alignments are shown in Figure 1(a) – (e). Generally, the genomes of prokaryotes are very dynamic, with insertions, deletions, inversions, and translocations being commonly observed among related species or even between different strains of the same species (Gregory and DeSalle, 2005; Hughes, 2000). The net result is that the particular complement of genes and their order along the chromosome are not typically conserved over evolutionary time. In some cases, genes that are grouped into operons in one species may be dispersed throughout the genome in others. We find similar results, while we analyzed the genomes of four strains of *S. pneumoniae*. In particular, we find that there exists a stability of the gene order in the genome pairs TIGR4 vs. D39 and TIGR4 vs. R6 and they are shown by fact that most of the points lie along the diagonal in Figures 1a and 1b. The results (Figures 1a and 1b) indicate that the stability of gene order of D39 vs. R6 must also be relatively high and it is shown in Figure 1c. This also confirms the

Sl. No.	Genome Information and Features	TIGR4	D39	G54	R6
1	Sequencing center	TIGR	TIGR	Geneva Biomedical Research Institute	Eli Lilly
	GenBank accession	AE005672.1	CP000410.1	NA	AE007317.1
	Refseq	NC_003028	NC_008533	NA	NC_003098
	Topology	Circular	Circular	Circular	Circular
	Molecule	dsDNA	dsDNA	dsDNA	dsDNA
	Contig	1	1	31 contigs	1
	Completed date	2001/10/03	2006/10/24	Not yet included in NCBI	2001/10/03
2	Genome size (sequence length)	2.16 Mb	2Mb	2.07Mb	2.03 Mb
	Number of A	653880 (30.26%)	617717 (30.19%)	628663 (30.31%)	615270 (30.18 %)
	Number of T	649168 (30.04%)	615968 (30.10%)	624751 (30.10%)	613689 (30.10 %)
	Number of G	430998 (19.95%)	407646 (19.92%)	404611 (19.50%)	406018 (19.91 %)
	Number of C	426796 (19.75%)	404784 (19.78%)	414824 (20.00%)	403638 (19.79 %)
	No. of A+T (%)	60.30	60.29	60.43	60.28
	No. of G+C (%)	39.69	39.71	39.50	39.71
3	Total size of DNA molecule	2160842 bp	2046115 bp	2074072 bp	2038615 bp
	Number of coding bases	1885091 bp (87.23%)	NA	1761820 bp (84.94%)	1761157 bp (86.38%)
	Number of genes	2234	1914	2047	2043
	Number of genes assigned to role ids	1506 (67.41%)	NA	1343 (65.60%)	1313 (64.26%)
	Structural RNAs	70	73	NA	73
	tRNA genes	58	58	51	58
	rRNA genes	12	12	5	12
4	% global direct repeats	8.30	CBS tool does not have the whole genome data of D39 and G54		5.70
	% global inverted repeats	7.00			5.40
	% local direct repeats	6.40			5.80
	% local inverted repeats	4.30			4.20

NA – Not Available

Table 1: Comparison of the genome features of the strains, encapsulated TIGR4, D39 & G54 and nonencapsulated R6 of *S. pneumoniae* using CMR, Bioedit and CBS tools

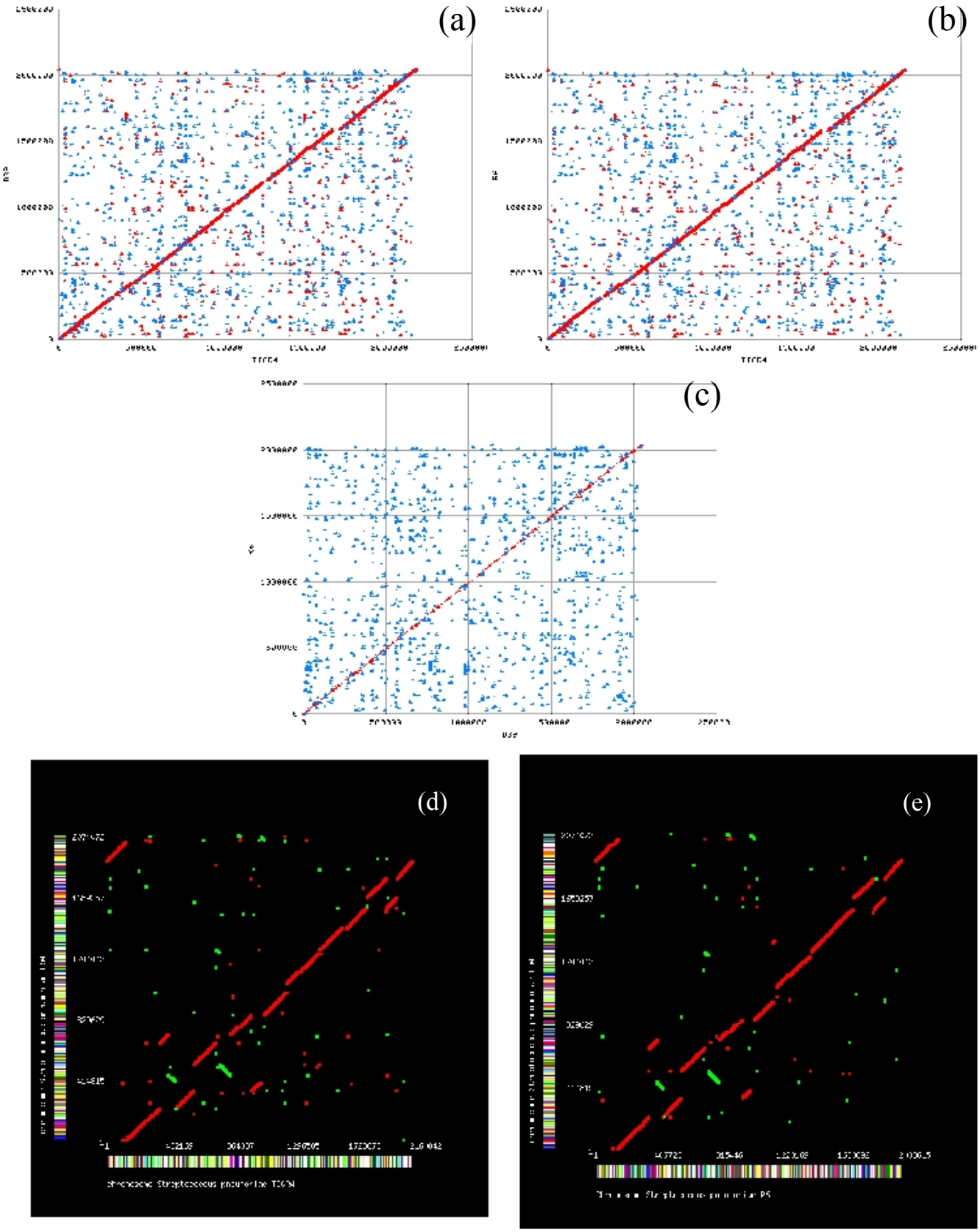

Figure 1: Whole genome alignment of a) TIGR4 vs. D39; b) TIGR4 vs. R6; c) D39 vs. R6 using stand-alone MUMmer; Whole genome alignment of d) TIGR4 vs. G54 and e) R6 vs. G54 using built-in MUMmer of CMR, which show plasticity and stability in gene order between two strains.

fact that R6 is the derivative of D39. The whole genome pairwise alignments of TIGR4 vs. G54 and that of R6 vs. G54 do not show such a high degree of the stability of gene order compared to the above results (for D39 strain) and are shown in Figures 1d and 1e, respectively.

Many of the gene and protein sequences among these strains are approximately the same and this is not surprising as all the strains occupy the same niche in the human respiratory system. The small differences might have arisen after the divergence of these strains from other evolutionary lineages for adaptations in their host. This increases greatly in pathogens and appears to be associated with the ability to infect eukaryotes, perhaps reflecting a mechanism for evading host immune defenses and the unique genes may be located in a plasticity zone.

Since G54 genome sequence is not available at NCBI web server and D39 genome is not available at CMR server, we could not get the whole genome alignment for D39 vs. G54. However, we are able to predict the whole genome pairwise alignment of D39 vs. G54, based on the earlier result. As the Figures 1d and 1e are similar, it indicates that the alignment of D39 vs. G54 must also possess similar structure. This prediction may be confirmed if the whole genome sequence of G54 is made available in NCBI or genome sequence of D39 is included in CMR.

Comparison of Capsular Polysaccharide Synthesizing Genes

We have compared the capsular polysaccharide (cps) synthesizing genes of the strains TIGR4, D39, G54 and R6 of *S. pneumoniae* and the results are shown in Table 2. There are 15 different cps genes in TIGR4, 7 in D39 and 9 in G54 and only one in R6. Their gene IDs, G+C percentage, protein length, gene length and gene coordinates are shown in Table 2. On comparison, it is estimated that 5 cps genes of TIGR4 (gi|15900275-cps4A, gi|15900276-cps4B, gi|15900278-cps4D, gi|15900046-cps-ptv & gi|15901666-cps-ptv) are related to that of D39 (gi|116516963-cps2A, gi|116516159-cpsB, gi|116517023-cps2D, gi|116517199-cps and gi|116516120-cps-ptv). All the cps genes of D39 are present in TIGR4 except gi|116516773-cps2E and gi|116516341-cps-ptv.

Between TIGR4 and G54, 6 cps genes are related (gi|15900275-cps4A, gi|15900276-cps4B, gi|15900277-cps4C, gi|15900278-cps4D, gi|15900046-cps-ptv & gi|15901666-cps-ptv of TIGR4 with NT05SP0190-cps4A, NT05SP0191-cps4B, NT05SP0192-cps4C, NT05SP0193-cps4D, NT05SP2185-cps9E & NT05SP1650-cps7G of G54). Likewise, between D39 and G54, 5 cps genes are related (gi|116516963-cps2A, gi|116516159-cpsB, gi|116517023-cps2D, gi|116517199-cps and gi|116516120-cps-ptv of D39 with NT05SP0190-cps4A, NT05SP0191-cps4B, NT05SP0192-cps4C, NT05SP2185-cps9E & NT05SP1650-cps7G of G54), but gi|116516773-cps2E and gi|116516341-cps-ptv of D39 are not present in G54. Similarly, it is interesting to note that the only cps gene of R6 (gi|15902136-capD), has 99.8 % identity with the gene gi|15900046-cps-ptv of TIGR4, 100 % identity with the gene gi|116517199-cps of D39 and 99.5 % identity with the gene NT05SP2185 of G54. All the above results are in support of the Avery's statement (Avery et al., 1979) that the capsule is responsible for pathogenecity.

From similar analysis, we have also noted that the genes, gi|15900279-cps4E, gi|15900280-cps4F, gi|15900281-cps4G, gi|15900282-cps4H, gi|15900286-cps4I, gi|15900287-cps4J, gi|15900288-cps4K, gi|15900289-cps4L and gi|15900788-cps-ptv are unique to TIGR4. Similarly, the genes gi|116516773-cps2E and gi|116516341-cps-ptv are unique to D39 strain. In the same way, the genes NT05SP0198, NT05SP0202 and NT05SP1909 are unique to the strain G54. But in R6, the only cps gene gi|15902136-capD is common to all other strains (Table 2). As the TIGR4 strain has more number of cps genes than other strains it indicates the high virulence nature of TIGR4. Further, the results also explain that the virulence nature is lesser in D39 and G54 strains, and very less in R6 compared to TIGR4.

Though all the cps genes of TIGR4 are not present in D39, G54 and R6 strains, they are also pathogenic. Therefore, to know the other virulence factors in addition to cps genes, we consider the other genes of the strains from the gene role category aspect.

Comparison of the Role Category of Genes

Role category of genes of the different strains are compared by using the two different tools, namely, i. CMR – role category pie chart for TIGR4, G54 and R6 (Table 3) and ii. Bacterial Annotation System (BASys) for the strains TIGR4, D39 and R6, based on the availability of genome sequences. The genes responsible for biosynthesis of various proteins (Sl. Nos. 1-9 of Table 3) of TIGR4 are nearly same as in G54 and R6, which suggests the basic complement of proteins required for certain cellular processes. But the genes responsible for the biosynthesis of some other proteins (Sl.Nos.10-23 of Table 3) of TIGR4 are notably different from that of G54 and R6. This suggests that, these proteins are important for strain uniqueness and they may be involved in variations in pathogenesis among the strains

Strain name	Gene ID and Name	G+C (%)	Protein length (aa)	Gene length (bp)	Gene coordinates	Comparison with cps of other strains in %Identity
TIGR4	gi|15900275-cps4A	38.32	481	1446	320077 - 321522	96.0 - D39-gi|116516963-cps2A 94.0 - G54-NT05SP0190-cps4A
	gi|15900276-cps4B	41.98	243	732	321524 - 322255	97.9 - D39-gi|116516159-cpsB 86.4 - G54-NT05SP0191-cps4B
	gi|15900277-cps4C	40.29	230	693	322264 - 322956	85.7 - G54-NT05SP0192-cps4C
	gi|15900278- cps4D	34.21	227	684	322966 - 323649	79.6 - D39-gi|116517023-cps2D 93.8 - G54-NT05SP0193-cps4D
	gi|15900279-cps4E	33.49	211	636	323990 - 324625	--
	gi|15900280-cps4F	33.17	409	1230	324634 - 325863	--
	gi|15900281-cps4G	27.84	358	1077	325868 - 326944	--
	gi|15900282-cps4H	31.36	372	1119	326937 - 328055	--
	gi|15900286-cps4I	36.70	365	1098	331774 - 332871	--
	gi|15900287-cps4J	38.46	351	1056	332875 - 333930	--
	gi|15900288-cps4K	36.19	409	1230	334030 - 335259	--
	gi|15900289-cps4L	35.02	394	1185	335260 - 336444	--
	gi|15900046-cps-ptv*	42.21	616	1851	104668 - 106518	99.8 - D39-gi|116517199-cps 99.7 - G54-NT05SP2185-cps9E 99.8 - R6-gi|15902136-capD
	gi|15900788-cps-ptv	28.79	455	1368	859370 - 860737	--
	gi|15901666-cps-ptv	43.93	408	1227	1746322 - 1747548	99.0 - D39-gi|116516120-cps-ptv 96.6 - G54-NT05SP1650-cps7G
D39	gi|116516963-cps2A	38.45	481	1446	313744 - 315189	96.3 - G54-NT05SP0190-cps4A
	gi|116516159-cpsB	41.53	243	732	315191 - 315922	85.2 - G54-NT05SP0191-cps4B
	gi|116517023-cps2D	39.06	226	681	316633 - 317313	79.3 - G54-NT05SP0192-cps4C
	gi|116516773-cps2E	37.79	455	1368	317328 - 318695	--
	gi|116517199-cps	42.19	616	1851	99217 - 101067	99.5 - G54-NT05SP2185-cps9E 100 - R6-gi|15902136-capD
	gi|116516341-cps-ptv	30.28	119	360	815811 - 816170	--
	gi|116516120-cps-ptv	44.09	408	1227	1633887 - 1635113	97.1 - G54- NT05SP1650-cps7G
G54	NT05SP0190-cps4A	38.28	484	1455	165975 - 167429	--
	NT05SP0191-cps4B	37.56	243	732	167431 - 168162	--
	NT05SP0192-cps4C	38.09	230	693	168171 - 168863	--
	NT05SP0193-cps4D	34.64	227	684	168873 - 169556	--
	NT05SP0198-cps19AI	29.82	445	1338	173388 - 174725	--
	NT05SP0202-cps23FP	41.70	198	597	178230 - 178826	--
	NT05SP1650-cps7G	43.77	417	1254	1493392 - 1492139	--
	NT05SP1909-cps3E	43.63	436	1311	1726013 - 1727323	--
	NT05SP2185-cps9E	42.46	616	1851	1999333 - 2001183	99.5 - R6-gi|15902136-capD
R6	gi|15902136-capD	42.26	616	1851	99217 - 101067	--

Table 2: Comparison of capsular polysaccharide (cps) synthesizing genes of four strains of *S. pneumoniae*. Each cps is compared with all cps sequences of other three strains using LALIGN; all the cps sequences considered fall under the Role Category 11 (Cell Envelope) of CMR.

of *S. pneumoniae*. The percentage values given for a particular role category in Table 3 is specific to the gene involved in that category only and does not represent the overall gene percentage. For example, autolysin (SP1937) of TIGR4 is categorized into two role categories such as cell envelope and cellular processes (Sl.Nos.11 and 12 of Table 3) and the percentage given is specific to the respective categories.

S. No.	Gene Role Category	TIGR4 No of genes - out of 2234(%)	G54 No of genes - out of 2047(%)	R6 No of genes - out of 2219(%)
	Similar proteins (common proteins)			
1.	Biosynthesis of cofactors, prosthetic groups, and carriers	42 (1.88%)	48 (2.34%)	47 (2.11%)
	DNA metabolism			
2.	Fatty acid and phospholipids metabolism	92 (4.11%)	98 (4.78%)	104 (4.68%)
3.	Protein fate	23 (1.02%)	37 (1.8%)	34 (1.53%)
4.	Protein synthesis	70 (3.13%)	76 (3.71%)	69 (3.10%)
5.	Purines, pyrimidines nucleosides	120 (5.37%)	129 (6.3%)	128 (5.76%)
6.	and nucleotides	54 (2.41%)	58 (2.83%)	61 (2.74%)
	Regulatory functions			
7.	Transcription	121 (5.41%)	117 (5.71%)	122 (5.49%)
8.	Transport and binding proteins	29 (1.29%)	29 (1.41%)	31 (1.39%)
9.		267 (11.90%)	218 (10.6%)	236 (10.6%)
	Dissimilar proteins (unique proteins)			
10.	Amino acid biosynthesis	53 (2.37%)	95 (4.64%)	100 (4.50%)
11.	Cell envelope	136 (6.08%)	131 (6.39%)	96 (4.32%)
12.	Cellular processes	147 (6.58%)	91 (4.44%)	76 (3.42%)
13.	Central intermediary metabolism	11 (0.49%)	87 (4.25%)	93 (4.19%)
14.	Disrupted reading frame	92 (4.11%)	0 (0%)	0 (0%)
15.	Energy metabolism	143 (6.40%)	185 (9.03%)	197 (8.87%)
16.	Hypothetical proteins	431 (19.20%)	236 (11.5%)	171 (7.70%)
17.	Conserved hypothetical proteins	302 (13.50%)	301 (14.7%)	519 (23.3%)
18.	Mobile and extra chromosomal Element functions	134 (5.99%)	71 (3.46%)	86 (3.87%)
19.	Pathogen responses*	101 (4.52 %)	47 (2.30%)	42 (1.89%)
20.	Signal transduction	79 (3.53%)	4 (0.19%)	4 (0.18%)
21.	Unclassified	0 (0%)	167 (8.15%)	201 (9.05%)
22.	Unknown function	174 (7.78%)	72 (3.51%)	51 (2.29%)
23.	Viral functions	0 (0%)	23 (1.12%)	26 (1.17%)

(* Manually counted).

Table 3: Distribution of genes in the whole genomes of TIGR4, G54 and R6 strains of *S. pneumoniae* based on their gene role category. These gene role category data are retrieved and compiled from CMR using its Gene Role Category Pie-chart.

Types of sequences	Role categories	TIGR4	G54	R6
Whole genome	Total no. of genes	2234	2047	2219
	Hypothetical	431	236	171
	Conserved hypothetical	302	301	519
	Unclassified	Nil	167	201
	Unknown	174	72	51
	Total	**907 – 40.6%**	**776 – 37.9%**	**942–42.5%**
Number of common and unique genes.	Number of sequences Present in all comparison molecules	1792	1824	1810
	Number of Present in at least one comparison molecule	1946	1943	1965
	Not present in any of the comparison molecule (unique genes)	288	104	78
Unique genes	Hypothetical	158	47	68
	Conserved hypothetical	13	11	Nil
	Unclassified	Nil	Nil	Nil
	Unknown	Nil	Nil	Nil
	Total	**189 (65.63%)**	**58 (74.36%)**	**68 (65.39%)**
Virulence factors among unique genes	Capsular polysaccharide biosynthesis protein	Sp_0351-cps4F	--	--
		Sp_0352-cps4G	--	--
		Sp_0359-cps4K	--	--
	Type 2 capsule locus	--	--	Spr0315
		--	--	Spr0317
		--	--	Spr0319
	Cell wall surface anchor family protein	Sp_0462	--	--
		Sp_0463	--	--
		Sp_0464	--	--
		Sp_1772	--	--
	PspC	Sp_1417	--	--
	NanA, authentic frameshift	Sp_1693	--	--
	IgA1 protease, degenerate	Sp_2155	--	--

Table 4: Details of the number of hypothetical sequences in whole genomes, unique genes and virulence factors in unique genes of the strains TIGR4, G54 and R6 of *S. pneumoniae*. (D39 data are not included due to the non-availability of the genome sequence information of D39 strain in CMR tool).

The number of genes which are responsible for pathogenesis in the strains TIGR4, G54 and R6 are manually counted from CMR gene role category (sub role categories pathogenesis, toxin production and resistance) and found to be 101 (4.52 %), 47 (2.30 %) and 42 (1.89 %) respectively (Sl.No.19 of Table 3). TIGR4 has many pathogenic factors and it is highly virulent and G54 and R6 strains have approximately 50% of the pathogenic factors of TIGR4. Mobile and extra chromosomal elements comprise a significant fraction of the genome as with the 134 genes (5.99 %) in TIGR4, 71 (3.46 %) in G54 and 86 genes (3.87 %) in R6 (Sl.No.18 of Table 3). Generally transposons encode genes for antibiotic resistance (Gregory and DeSalle, 2005); therefore from our results, it is evident that the antibiotic resistance may be relatively higher in TIGR4 than the strains G54 and R6.

From the results of the comparative study on TIGR4, D39 and R6, using BASys server, we find that most of the values are more or less similar. But, there is a higher percentage for unknown functions in the strains TIGR4, D39 and G54, which indicates that the reason for the differences may also be hidden in the unknown genes or proteins (data not shown).

From Table 3, the number of hypothetical, conserved hypothetical, unclassified and unknown genes of whole genomes of the strains TIGR4, G54 and R6 are noted and is shown in Table 4. Nearly 37 - 42 % of genes are of unknown type and it shows that these sequences have to be annotated and assigned functions of which some of them may be responsible for the virulence nature. Using the multigenome homology comparison tool, which is available at CMR, the numbers of unique genes in TIGR4, G54 and R6 are found to be 288, 104 and 78, respectively (Table 4).

The unique genes of the strains TIGR4, G54 and R6 themselves have many hypothetical, conserved hypothetical, unknown and unclassified sequences and their percentage ranges from 65 to 74, thus the other possible differences among the strains may be known by studying the above said gene sequences. As far as the virulence factors are concerned, in the unique genes of the strain TIGR4, 3 capsular polysaccharide biosynthesis proteins (Sp_0351 (cps4F), Sp_0352 (cps4G) and Sp_0359 (cps4K)), 4 cell wall surface anchor family proteins (Sp_0462, Sp_0463, Sp_0464 and Sp_1772), a PspC protein (Sp_1417), a NanA protein (SP_1693) and a IgA1 protease (SP_2155) are there. In the case of R6, it has three proteins of type 2 capsule locus (Spr0315, Spr0317 and Spr0319) in its unique genes. But the strain G54 does not have such virulence factors in its unique genes (Table 4). The above result shows the high virulence nature of TIGR4 and it also suggests that those virulence factors are specific to TIGR4 and R6. The above differences might have arisen because of the species-specific adaptation to their host particularly in the sake of defense mechanism.

Comparison of Virulence Factors Other than Capsular Polysaccharide Synthesizing Genes

In *S. pneumoniae*, the surface and cytoplasmic proteins such as pneumococcal surface protein A (PspA), autolysin (LytA), hyaluronate lyase (Hyl), pneumolysin (Ply), two neuraminidases (NanA and NanB), choline binding protein A (CbpA), pneumococcal surface antigen A (PsaA) and immunoglobulin A1 (IgA1) protease are already stated as the virulence factors (Jedrzejas, 2001; Rigden et al., 2003). The comparative results of the above mentioned sequences obtained from CMR, are given in Table 5. It provides more insight into the virulence factors of the strains TIGR4, D39, G54 and R6 of *S. pneumoniae*.

The virulence factors of TIGR4 are taken as reference and are compared with all other related sequences of the strains such as D39, G54 and R6, likewise the virulence factors of D39 are taken as reference and are compared with all the related sequences of the strains G54 and R6. Similarly the virulence factors of G54 are taken as reference and are compared with all the related sequences of the remaining strain R6 using the pairwise sequence alignment tool LALIGN, with default parameters (Alignment: Global; Scoring matrix: BLOSUM50, Gap opening penalty: -14 and extension penalty: -4), and all the results are comparatively shown in Table 5.

PspA is located in the cell wall of *pneumococci* and present in all *S. pneumoniae* strains (Jedrzejas, 2001). PspA of TIGR4 has ~53-63% identities with D39, G54 and R6 (Table 5). When we compare PspA in D39 vs. G54 and G54 vs. R6, the identities between those strains are nearly 63%. The above results indicate that nearly 50-60% virulence nature of PspA of TIGR4 exist in other strains D39, G54 and R6. But it is interesting to note that there is 100% identity between the PspA sequences of D39 and R6, thus the virulence nature of PspA is exactly the same.

Regarding LytA, Hyl, Ply, NanB and PsaA, all the four strains of *S. pneumoniae* have above 90% identities, thus the effect of the above mentioned five virulence factors is also similar and it also reflects on G+C percentage, protein length and gene length, but the location in their genomes varies and the similarities and differences can be noticed from the Table 5.

Strain	Gene ID	Virulence factors	G+C (%)	Protein length (aa)	Gene length (bp)	Gene Coordinates 5' 3'	Role category ***	% Identity with D39	% Identity with G54	% Identity with R6
TIGR4	gi|15900059	PspA	40.23	744	2235	118423 120657	1	53.6	62.5	53.6
	gi|15901761	LytA	46.44	318	957	1841361 1840405	3	99.7	100	99.7
	gi|15900247	Hyl	40.15	1066	3201	287483 290683	5	98.8	97.5	97.8
	gi|15901747	Ply	41.83	471	1416	1833311 1831896	6	99.8	100	99.8
	gi|15901180	Nan-ptv*	35.36	740	2223	1251631 1249409	3	10.5	20.4	19.6
	gi|15901522	NanB	33.38	697	2094	1589236 1587143	3	99.1	98.9	99.1
	gi|15901997	CbpA	41.90	693	2082	2112096 2110015	9	73.7	40.5	73.7
	gi|15901485	PsaA	37.11	309	930	1549466 1550395	14	99.7	98.1	99.7
	gi|15901019	IgA1	38.06	2004	6015	1083881 1089895	15	87.3	35.9	87.3
D39	gi|116515876	PspA	42.63	619	1860	128356 130215		--	63.4	100
	gi|116516777	LytA	46.39	318	957	1729601 1730557		--	99.7	100
	gi|116515977	Hyl	40.14	1067	3204	285186 288389		--	97.8	99.0
	gi|116515376	Ply	42.02	471	1416	1721457 1722872		--	99.8	100
	gi|116515419	N.lyase-ptv**	43.31	243	732	1190890 1191621	NA	--	10.2	9.9
	gi|116516987	NanB	33.38	697	2094	1515745 1517838		--	98.6	100
	gi|116515359	CbpA	41.26	701	2106	1995044 1997149		--	31.0	100
	gi|116515973	PsaA	37.10	309	930	1478217 1479146		--	98.4	100
	gi|116516343	IgA1	39.09	1963	5892	1037492 1043383		--	36.4	100
G54	NT05SP2202	PspA	41.36	709	2130	2015436 2017565	1	--	--	63.4
	NT05SP1836	LytA	46.29	318	957	1656972 1656016	4	--	--	99.7
	NT05SP0158	Hyl	39.97	1078	3237	137159 140395	5	--	--	98.7
	NT05SP1746	Ply	41.94	471	1416	1577243 1575828	7	--	--	99.8
	NT05SP1517	Nan A	41.48	980	2943	1379132 1376190	8	--	--	90.8
	NT05SP1511	Nan B	33.19	697	2094	1371552 1369459	8	--	--	98.6
	NT05SP2037	CbpA	40.67	739	2220	1848760 1846541	10	--	--	31.0
	NT05SP1476	psaA	37.04	313	942	1331767 1332708	7	--	--	98.4
	NT05SP2154	IgA1	36.78	1856	5571	1969880 1975450	16	--	--	36.4
R6	gi|15902165	PspA	42.65	619	1860	128356 130215	2	--	--	--
	gi|15903796	LytA	46.54	318	957	1723025 1722069	4	--	--	--
	gi|15902330	Hyl	40.01	1078	3237	285103 288339	5	--	--	--
	gi|15903781	Ply	42.04	471	1416	1715341 1713926	7	--	--	--
	gi|15903579	NanA	42.67	1035	3108	1518051 1514944	8	--	--	--
	gi|15903574	NanB	33.43	697	2094	1510307 1508214	8	--	--	--
	gi|15904036	CbpA	41.32	701	2106	1989649 1987544	10	--	--	--
	gi|15903537	PsaA	37.22	309	930	1470686 1471615	12	--	--	--
	gi|15903086	IgA1	39.09	1963	5892	1029961 1035852	13	--	--	--

NA - Not Available

Table 5: Comparison of the common virulence factors namely, pneumococcal surface protein A (PspA), autolysin (LytA), hyaluronate lyase (Hyl), pneumolysin (Ply), neuraminidase A (NanA), neuraminidase B (NanB), choline binding protein A (CbpA), pneumococcal surface antigen A (PsaA) and immunoglobulin A1 (IgA1) protease of four strains of *S. pneumoniae*. LALIGN program is used to find identity between sequences.

* Nan-ptv: Neuraminidase, putative
** N.lyase-ptv: N-acetylneuraminate lyase, putative

*** *Role category functions*

1. Cell envelope; cellular process – pathogenesis
2. Mobile and extra chromosomal element function: transposon function
3. Cell envelope biosynthesis and degradation of surface polysaccharides and Lipopolysaccharides; Cellular processes: pathogenesis
4. Cell envelope: biosynthesis and degradation of murine sacculus and peptidoglycan
5. Cellular processes: pathogenesis
6. Cellular processes: toxin production and resistance; Cellular processes: pathogenesis
7. Unclassified: role category not yet assigned
8. Viral function: general
9. Cell envelope; cellular process – pathogenesis cellular process: cell adhesion
10. Cellular processes toxin production and resistance; Fatty acid and phospholipid metabolism: degradation
11. Cell envelope biosynthesis and degradation of surface polysaccharides and Lipopolysaccharides
12. Unclassified – role category not yet assigned
13. protein fate: Degradation of proteins, peptides and glycopeptides
14. Transport and binding proteins: Cations and iron carrying compounds; Cellular processes: pathogenesis; cellular processes:cell adhesion
15. protein fate: Degradation of proteins, peptides and glycopeptides; Cellular processes: pathogenesis
16. protein fate: Degradation of proteins, peptides and glycopeptides

All strains have different neuraminidase sequences except G54 and R6 (~90% identity). In the case of CbpA and IgA1 of the strain TIGR4, high percent identities (~73 and 87%) exist with D39 and R6 respectively, exactly identical (100%) between D39 and R6. But very less identities (~40 and 35%) exist with G54 combinations. It seems that the virulence nature based on cbpA and IgaA are similar among the strains TIGR4, D39 and R6 and differs in G54.

From Table 5, it is interesting to note that all the virulence factors of D39 are very similar to R6 (above 99% identities except NanA), and it confirms the fact that the avirulent strain R6 is the derivative of the strain D39 (Lanie et al., 2007). Based on the role category, all TIGR4 virulence factors come under pathogenesis related functions and it also says that TIGR4 has high virulence nature.

Functional Annotation of Hypothetical Sequences Relevant to the Virulence Factors

Prediction of virulence factors from the hypothetical sequences of *S. pneumoniae* has implications on the identification and characterization of the virulence mechanism. The present study predicted using VirulentPred (Garg and Gupta, 2008) that 4 hypothetical sequences of TIGR4 and 22 of R6, respectively, are virulence factors. All these sequences are listed in Table 6. The prediction is based on protein features, such as, amino acid composition, di-peptide composition, similarity search, higher order di-peptide composition, PSSM and cascaded SVM module of the tool VirulentPred. However, similar predictions are not possible at present with D39 and G54 as the sequence information of the latter is not fully available.

Among the 4 predicted virulence factors of TIGR4, only one sequence (gi|15901572) is predicted in R6 as a hypothetical protein (gi|15903627) and the functional region is predicted as Plasmid_Txe (PF06769). This family contains many hypothetical proteins and there is no homolog with other mentioned virulence factors. But in R6, it is interesting to note that among the 22 predicted virulence factors of hypothetical protein sequences, 8 different sequences (gi|15902372, gi|15903388, gi|15903446, gi|15902652, gi|15902781, gi|15903694, gi|15903627 and gi|15903771) with 7 different functional regions which are related to the already mentioned virulence factors of the strains R6 and TIGR4. Those virulence factors are hyaluronidase, Immunoglobulin A1 protease, capsular polysaccharide synthesis, pneumolysin, neuraminidase and choline binding protein. The above mentioned related sequences of TIGR4 and R6 except gi|15903771 are compared in Table 7.

The hypothetical protein sequence, gi|15903771 of R6 has 71 amino acids and its functional region is predicted as pu-

S. No.	Protein ID	Protein Length
TIGR4		
1	gi|15900762	177
2	gi|15900877	1039
3	gi|15901572	84
4	gi|15902036	255
R6		
1	gi|15902135	385
2	gi|15902152	450
3	gi|15902269	65
4	gi|15902355	57
5	gi|15902369	149
6	gi|15902372	1767
7	gi|15902511	111
8	gi|15902652	337
9	gi|15902781	170
10	gi|15902826	177
11	gi|15902850	122
12	gi|15903009	368
13	gi|15903331	330
14	gi|15903388	202
15	gi|15903446	2551
16	gi|15903447	502
17	gi|15903627	84
18	gi|15903694	719
19	gi|15903697	243
20	gi|15903771	71
21	gi|15903873	64
22	gi|15903916	380

Table 6: List of predicted 4 and 22 hypothetical protein sequences as virulence factors from Tigr4 and R6 respectively.

tative cell wall binding repeat (42-60) using Interproscan (ID - PF01473). It is also found that the same functional region is repeatedly present in the known virulence factors such as pneumococcal surface protein A, autolysin and choline binding proteins of the strains TIGR4 and R6. Since many domain regions have been identified in the above mentioned known virulence factors of TIGR4 and R6, the regions are not explicitly given. But one can easily obtain those regions using the tool Interproscan.

Conclusion

We have compared the virulence nature of the strains, encapsulated TIGR4, D39, G54 and nonencapsulated R6 of *Streptococcus pneumoniae* using comparative genomics tools. From the whole genome pairwise alignment, we found that the stability of the gene order in the genomes of TIGR4 vs. D39, TIGR4 vs. R6 and D39 vs. R6 are relatively higher than the genomes of TIGR4 vs. G54 and R6 vs. G54. We are able to predict the possible structure of whole genome pairwise alignment of D39 vs. G54 from the alignments of TIGR4 vs. G54 and R6 vs. G54.

From the comparison on the capsular polysaccharide (cps) synthesizing genes, we found that, TIGR4 strain has more number of cps genes than other strains, which may indicate the high virulence nature of TIGR4. Many cps genes are unique to TIGR4, only few are in D39 & G54 and none in R6, which shows the high virulence nature of TIGR4. Further, the study on other virulence factors such as, pneumo-

ID from Interpro scan	ID of Hypo. Pro. Seq. of TIGR4 and R6	Length	Domain position	ID of known Virulence Factors of TIGR4 and R6	Name of Virulence Factors	Length	Domain position	Functional region
				TIGR4				
PF06769	gi:|15901572	84	5-84	gi|15903627-VirPredR6	Hypothetical	84	5-84	Plasmid_Txe
				R6				
PF00746	gi|15902372 gi|15903388 gi|15903446	1767 202 2551	1727 – 1766 159-199 2513-2549	gi|15902330 – R6 gi|15903086 – R6 gi|15900247 – TIGR4 gi|15901019 – TIGR4	Hyl IgA1 Hyl IgA1	1078 1963 1066 2004	1040-1077 88-127 1028-1065 88-127	Surface protein from Gram-positive cocci, anchor region
G3DSA: 3.40.50.720	gi|15902652	337	3-237	gi|15902136 – R6 gi|15900287 – TIGR4 gi|15900288 – TIGR4 gi|15900046 – TIGR4	CapD cps4J cps4K cps putative	616 351 409 616	289-544 3-231 2-129 289-544	Ubiquitin-activating enzyme E1
PF01289	gi|15902781	170	63-168	gi|15903781 – R6 gi|15901747 – TIGR4	Ply Ply	471 471	67-84, 84-100, 142-162 63-168	Thiol-activated cytolysin
PF04650	gi|15903694 gi|15903446	719 2551	15-41 12-38	gi|15903579 – R6 gi|15904036 – R6 gi|15903086 – R6 gi|15901997 – TIGR4 gi|15901019 – TIGR4	NanA cbpA IgA1 cbpA IgA1	1035 701 1963 693 2004	21-47 1-40 6-32 6-32 6-32	YSIRK Gram-positive signal peptide
PF07501	gi|15903446	2551	473-549	gi|15903086 – R6 gi|15901019 – TIGR4	IgA1 IgA1	1963 2004	315-393 315-393	G5
PF06769	gi|15903627	84	5-84	gi|15901572 – VirPredTIGR4	Hypothetical	84	5-84	Plasmid_Txe

Table 7: Comparison of the predicted and known virulence factors of hypothetical protein sequences with already known virulence factors of TIGR4 and R6 of *S. pneumoniae*.

coccal surface protein A, autolysin, hyaluronate lyase, pneumolysin, neuraminidase B and pneumococcal surface antigen A of TIGR4 are closely related to those of the other three strains, which shows that the virulence nature due to these factors among four strains seems to be similar. But the virulence factors neuraminidase A, choline binding protein A and immunoglobulin A1 protease of TIGR4 differs from other strains of *S. pneumoniae*, which shows that these factors are responsible for the differences in virulence nature among four strains.

From the gene role category comparison, many genes of TIGR4 that are nearly same as in G54 and R6, suggests the basic complement of proteins required for certain cellular processes in the strains of *S. pneumoniae*. But many of the genes of TIGR4 which are notably different from the strains G54 and R6, suggest that these proteins are important for strain uniqueness and they may be involved in variations in pathogenesis. Since many hypothetical, conserved hypothetical, unknown and unclassified proteins exist among the dissimilar role categorized genes, it seems that many of these genes of *S. pneumoniae* have to be annotated and assigned functions of which some of them may also be responsible for the virulence nature. Further, we have also found that most of the virulence factors are same in D39 and R6 and hence also confirms the fact that R6 is the derivative of the strain D39.

In order to annotate the uncharacterized protein sequences (hypothetical and conserved hypothetical), the present study predicted 4 and 22 hypothetical sequences of the strains TIGR4 and R6 respectively of *S. pneumoniae* are of virulence factors. Among those predicted virulence factors, 1 and 8 different hypothetical sequences of TIGR4 and R6 respectively contain conserved sequences of known virulence factors such as hyaluronidase, immunoglobulin A1 protease, capsular polysaccharide synthesis, pneumolysin, neuraminidase and choline binding protein. These sequences also may be considered as desirable targets for therapeutics. The effort is to narrow down the search of virulence factors from all hypothetical sequences and this conclusion will be a reality only when it is experimentally proved.

References

1. AlonsoDeVelasco E, Verheul AF, Verhoef J, Snippe H (1995) *Streptococcus pneumoniae*: virulence factors, pathogenesis, and vaccines. Microbiol Rev 59: 591-603. »CrossRef » Pubmed » Google Scholar
2. Avery OT, MacLeod CM, McCarty M (1979) Studies on the chemical nature of the substance inducing transformation of *pneumococcal* types. Inductions of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 149: 297-326. »CrossRef » Pubmed » Google Scholar
3. Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10: 398-400. »CrossRef » Pubmed » Google Scholar
4. Brown TA Jr, Ahn SJ, Frank RN, Chen YY, et al. (2005) A hypothetical protein of *Streptococcus mutans* is critical for biofilm formation. Infect Immun 73: 3147-3151. »CrossRef » Pubmed » Google Scholar
5. Brückner R, Nuhn M, Reichmann P, Weber B, Hakenbeck R (2004) Mosaic genes and mosaic chromosomes - genomic variation in *Streptococcus pneumoniae*. Int J Med Microbiol 294: 157-168. »CrossRef » Pubmed » Google Scholar
6. Dopazo J, Mendoza A, Herrero J, Caldara F, Humbert Y, et al. (2001) Annotated draft genomic sequence from a *Streptococcus pneumoniae* type 19F clinical isolate. Microb Drug Resist 7: 99-125. »CrossRef » Pubmed » Google Scholar
7. Dowson CG (2004) What is a Pneumococcus? In Tuomanen et al. (eds)The Pneumococcus ASM press Washington pp 3-14.
8. Ferretti JJ, Ajdic D, McShan WM (2004) Comparative genomics of streptococcal species. Indian J Med Res 119: 1-6. »CrossRef » Pubmed » Google Scholar
9. Galperin MY, Koonin EV (2004) Conserved hypothetical proteins: prioritization of targets for experimental study. Nucleic Acids Res 32: 5452-5463. »CrossRef » Pubmed » Google Scholar
10. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 28: 9-62. »CrossRef » Pubmed » Google Scholar
11. Gregory TR, DeSalle R (2005) Comparative genomics in prokaryotes. In Gregory (ed.) The evolution of the genome, Elsevier/Academic Press. London pp 585-660.
12. Hoskins J, Alborn WE Jr, Arnold J, Blaszczak LC, et al. (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 183: 5709-5717. »CrossRef » Pubmed » Google Scholar
13. Hughes D (2000) Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biol 1: reviews 0006.1-0006.8. »CrossRef » Pubmed » Google Scholar
14. Jedrzejas MJ (2001) Pneumococcal virulence factors: structure and function. Microbiol Mol Biol Rev 65: 187-207. »CrossRef » Pubmed » Google Scholar
15. Jothi R, Manikandakumar K, Ganesan K, Parthasarathy S (2007) On the analysis of the virulence nature of TIGR4

and R6 strains of *Streptococcus pneumoniae* using genome comparison tools. J Chem Sci 119: 559-563. »CrossRef » Google Scholar

16. Lanie JA, Wai LNG, Kazmierczak KM, Andrzejewski TM, Davidsen TM, et al. (2007) Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. J Bacteriol 189: 38-51. »CrossRef » Pubmed » Google Scholar

17. Polissi A, Pontiggia A, Feger G, Altieri M, Mottl H, et al. (1998) Large-scale identification of virulence genes from *Streptococcus pneumoniae*. Infect Immun 66: 5620-5629. »CrossRef » Pubmed » Google Scholar

18. Rigden DJ, Galperin MY, Jedrzejas MJ (2003) Analysis of structure and function of putative surface-exposed proteins encoded in the *Streptococcus pneumoniae* genome: A Bioinformatics-based approach to vaccine and drug design. Crit Rev Biochem Mol Biol 38: 143-168. » Pubmed » Google Scholar

19. Silva NA, McCluskey J, Jefferies JM, Hinds J, Smith A, et al. (2006) Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates. Infect Immun 74: 3513-3518. »CrossRef » Pubmed » Google Scholar

20. Sivashankari S, Shanmughavel P (2006) Functional annotation of hypothetical proteins – A review. Bioinformation 1: 335 -338. »CrossRef » Pubmed » Google Scholar

21. Tettelin H, Hollingshead SK (2004) Comparative genomics of *Streptococcus pneumoniae*: Intrastrain diversity and genome plasticity. In Tuomanen etal. (eds) The Pneumococcus ASM press Washington pp 15-29. »CrossRef » Google Scholar

22. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293: 498-506. »CrossRef » Pubmed » Google Scholar

23. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, et al. (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 69: 1593-1598. »CrossRef » Pubmed » Google Scholar