Comparison of the Virulence Factors and Analysis of Hypothetical Sequences of the Strains TIGR4, D39, G54 and R6 of \textit{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 \textit{Streptococcus Pneumoniae}. J Comput Sci Syst Biol 1: 103-118. doi:10.4172/jcsb.1000010

Copyright: © 2008 Jothi R, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Whole genome sequences of the four strains of \textit{Streptococcus 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 \textit{S. pneumoniae}.

Keywords: Comparative genomics; \textit{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. \textit{Streptococcus pneumoniae}, commonly \textit{pneumococcus} (Dowson, 2004; Gregory and DeSalle, 2005), a human pathogen, causes life threatening diseases like pneumonia, bacteremia, meningitis, sepsis, and otitis media. Genome sequencing of four \textit{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-
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 10^4 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 Shanmugavel, 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://
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 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.
| Sl. No. | Genome Information and Features | TIGR4 | D39 | G54 | R6 |
|--------|--------------------------------|-------|-----|-----|----|
|        | Sl. No. | Sequencing center | GenBank accession | Topology | Contig | Completed date | Genome size (sequence length) | Number of A | Number of T | Number of G | Number of C | No. of A+T (%) | No. of G+C (%) | Total size of DNA molecule | Number of coding bases | Number of genes assigned to role ids | Structural RNAs | tRNA genes | rRNA genes | % global direct repeats | % global inverted repeats | % local direct repeats | % local inverted repeats | CBS tool does not have the whole genome data of D39 and G54 | NBC tool does not have the whole genome data of D39 and G54 | NA – Not Available |
| 1      | 1       | AE005672.1 | CP000410.1 | AE007317.1 | CP000410.1 | 2001/10/03 | 2.16 Mb | 653880 (30.26%) | 649168 (30.04%) | 430998 (19.95%) | 426796 (19.75%) | 60.30 | 39.69 | 2160842 bp | 1885091 bp (87.23%) | 1506 (67.41%) | 70 | 58 | 12 | 8.30 | 7.00 | 6.40 | 4.30 |
| 2      | 2       | TIGR     | TIGR     | TIGR     | TIGR     | 2006/10/24 | 2.07 Mb | 677717 (30.19%) | 664046 (30.10%) | 476486 (19.92%) | 474844 (20.00%) | 60.39 | 39.57 | 2074072 bp | 1761157 bp (86.38%) | 1343 (65.60%) | 73 | 58 | 12 | 5.70 | 5.40 | 5.80 | 4.20 |
| 3      | 3       | NA       | NA       | NA       | NA       | Not yet included in NCBI | 2.03 Mb | 612570 (30.18%) | 613689 (30.11%) | 406018 (19.91%) | 403638 (19.79%) | 60.28 | 39.71 | 2038615 bp | 1761157 bp (86.38%) | 1313 (64.26%) | 73 | 58 | 12 | NA | NA | NA | NA |
| 4      | 4       | NA       | NA       | Geneva Biomedical Research Institute | Circular dsDNA | 2006/10/24 | 2.0 Mb | 563663 (30.19%) | 561408 (30.04%) | 402426 (19.92%) | 401044 (19.78%) | 60.29 | 39.71 | 2046115 bp | 1761157 bp (86.38%) | 1343 (65.60%) | 73 | 58 | 12 | 5.70 | 5.40 | 5.80 | 4.20 |

**Table 1:** Comparison of the genome features of the strains: encapsulated TIGR4, D39 & G54 and non-encapsulated 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 (gi15900275-cps4A, gi15900276-cps4B, gi15900278-cps4D, gi15900046-cps-ptv & gi15901666-cps-ptv) are related to that of D39 (gi116516963-cps2A, gi116517023-cps2D, gi116517199-cps and gi116516120-cps-ptv of D39 with NT05SP0190-cps4A, NT05SP0191-cps4B, NT05SP0192-cps4C, NT05SP0193-cps4D, NT05SP0198-cps9E & NT05SP0202-cps7G of G54), but gi116516773-cps2E and gi116516341-cps-ptv of D39 are not present in G54. Similarly, it is interesting to note that the only cps gene of R6 (gi15902136-capD), has 99.8 % identity with the gene gi15900046-cps-ptv of TIGR4, 100 % identity with the gene gi116517199-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 pathogenicity.

From similar analysis, we have also noted that the genes, gi15900279-cps4E, gi15900280-cps4F, gi15900281-cps4G, gi15900282-cps4H, gi15900286-cps4I, gi15900287-cps4J, gi15900288-cps4K, gi15900289-cps4L and gi15900788-cps-ptv are unique to TIGR4. Similarly, the genes gi116516773-cps2E and gi116516341-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 gi15902136-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.
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.
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.

| 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(%) |
|--------|--------------------------------------------------------|-------------------------------------|----------------------------------|----------------------------------|
| 1.     | Similar proteins (common proteins)                     |                                      |                                  |                                 |
| 1.     | Biosynthesis of cofactors, prosthetic groups, and carriers | 42 (1.88%)                          | 48 (2.34%)                       | 47 (2.11%)                      |
| 2.     | DNA 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 and nucleotides       | 120 (5.37%)                         | 129 (6.3%)                       | 128 (5.76%)                     |
| 6.     | Regulatory functions                                   | 54 (2.41%)                          | 58 (2.83%)                       | 61 (2.74%)                      |
| 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%)                     |
| 10.    | 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 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).

| 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     | 174   | 167 | 201 |
|                    | Unknown          | Nil   | 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 | -- | -- |
|                    | Type 2 capsule locus | Sp_0352-cps4G | -- | -- |
|                    | Cell wall surface anchor family protein | Sp_0359-cps4K | -- | Spr0315 |
|                    | PspC             | --     | Sp0462 | Spr0317 |
|                    | NanA, authentic frameshift | --     | Sp0463 | Spr0319 |
|                    | IgG1 protease, degenerate | --     | Sp0464 | -- |
|                    | Spr0315          | --     | Sp1417 | -- |
|                    | Spr0317          | --     | Sp1693 | -- |
|                    | Spr0319          | --     | Sp2155 | -- |

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

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |

% Identity with G54

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |

% Identity with R6

% Identity with D39

% Identity with G54

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |

% Identity with G54

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |

% Identity with G54

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |

% Identity with G54

| Strain | Gene ID | G+C (%) | Protein length (aa) | G value | Virulence factors | Gene Coordinates 5' | Role category |
|--------|---------|---------|---------------------|---------|------------------|---------------------|--------------|
| TIGR4  | 915900059 | 74.03   | 1527 | 1.00 | PpSA | 911150576-911152066 | | |
| D39    | 915900167 | 80.77   | 1260 | 1.00 | LVA | 911165576-911167280 | 1 | |
| G54    | 915900178 | 84.38   | 2004 | 1.00 | N.lyase-plv | 911165576-911167280 | 2 | |
| R6     | 915900189 | 79.02   | 1530 | 1.00 | IgA1 | 911165576-911167280 | 3 | |
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 IgA1 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 (gil15901572) is predicted in R6 as a hypothetical protein (gil15903627) 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 (gil15902372, gil15903388, gil15903446, gil15902652, gil15902781, gil15903694, gil15903627 and gil15903771) 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 gil15903771 are compared in Table 7.

The hypothetical protein sequence, gil15903771 of R6 has 71 amino acids and its functional region is predicted as pu-
Table 6: List of predicted 4 and 22 hypothetical protein sequences as virulence factors from Tigr4 and R6 respectively.

| 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            |

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 | Length | Domain position | Name of Virulence Factors | Functional region | Length Domain position | ID of known Virulence Factors of TIGR4 and R6 | Domain position | ID of known Virulence Factors of TIGR4 and R6 |
|----------------------|--------|-----------------|---------------------------|-------------------|------------------------|---------------------------------------------|----------------|---------------------------------------------|
| PF07679              | 5-84   | 1727 – 1766     | Hyl                       | Hypothetical      | 84                     | g|15903627-VirPredTIGR4                        | 15902332 – R6 | g|15903627-VirPredTIGR4                        |
| PF00746              | 202    | 156-199         |cps putative              | CapD              | 3-237                  | gi|15902136 – R6                               | 156-199        | gi|15902136 – R6                               |
| G3DSA: PF0746        | 337    | 2-129           | cps putative              | Ply               | 2551                   | gi|15900040 – TIGR4                            | 2-129          | gi|15900040 – TIGR4                            |
| PF01289              | 316    | 63-168          | Ply                       | Ply               | 2551                   | gi|15900040 – TIGR4                            | 63-168         | gi|15900040 – TIGR4                            |
| PF04650              | 2551   | 15-41           | Ply                       | Ply               | 2551                   | gi|15900040 – TIGR4                            | 15-41          | gi|15900040 – TIGR4                            |
| PF07501              | 2551   | 315-393         | Ply                       | Ply               | 2551                   | gi|15900040 – TIGR4                            | 315-393        | gi|15900040 – TIGR4                            |
| PF07679              | 84     | 5-84            | Hyl                       | Hypothetical      | 84                     | g|15903627-VirPredTIGR4                        | 5-84           | g|15903627-VirPredTIGR4                        |

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, Caldana 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, Blaszcak 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.

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.

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.

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.

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.

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

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

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.

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.