J. Chem. Sci., Vol. 119, No. 5, September 2007, pp. 559–563.

On the analysis of the virulence nature of TIGR4 and R6 strains of *Streptococcus pneumoniae* using genome comparison tools

R JOTHI, K MANIKANDAKUMAR, K GANESAN and S PARTHASARATHY*

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, Tamil Nadu
e-mail: bdupartha@gmail.com

MS received 26 June 2007; accepted 20 August 2007

Abstract. Comparative genome sequence analysis is a powerful technique for gaining insights into any genome of interest. *Streptococcus pneumoniae* is a human pathogen, which causes life-threatening diseases, such as pneumoniae, bacteremia, meningitis, etc. After the whole genome of two strains of *S. pneumoniae*, the virulent TIGR4 and non-pathogenic R6 were sequenced; there is a hope that comparing the genomes will allow an identification of the genes responsible for its virulence and thus the development of treatment and control. Many antimicrobial drugs have diminished the risk from pneumococcal disease because of its multi-drug resistance nature. Several pneumococcal proteins are also being investigated, as virulence factors as potential vaccine or drug targets. Structural and biochemical studies of these pneumococcal virulence factors have facilitated the development of novel antibiotics or protein antigen-based vaccines for the treatment of pneumococcal disease. Here we describe the comparison between the genomes of two strains of *S. pneumoniae* with few existing genomics databases and tools available in the public domain websites. By comparing nucleotide and protein sequences of the two strains, we investigate the existing differences and similarities. Mainly we focus on the virulence factors and its encoding genes in TIGR4 and how do they differ from R6 strain.

Keywords. Genomics; *Streptococcus pneumoniae*; virulence factors; TIGR4; R6.

1. Introduction

Sequencing of whole microbial genomes is re-shaping the fields like microbiology, biotechnology, molecular biology, biochemistry, etc. Presently, over 578 complete genome sequences have been reported with an approximate 1927 ongoing genomes (Genomes Online Database). Among them, 24 genomes of streptococcal species are completed and 45 are ongoing. *Streptococcus* becomes one of the most heavily sequenced genera of all and the isolates of *S. pneumoniae* are varying in their polysaccharide capsule and 90 different serotypes are available. Among these, only 4 genomes of strains of *S. pneumoniae* are completed including TIGR4 and R6 and nearly 14 strains are in progress (http://genome.microbio.uab.edu/strep/info, and http://www.sanger.ac.uk/Projects/S_pneumoniae/).

TIGR4: This strain is a highly virulent capsular serotype 4 clinical isolate. Many number of virulence factors are studied in this strain.[1] R6: Non-capsulated and the lack of a polysaccharide capsule in R6 renders its avirulent and a safe strain with which to work. The essential utility of the strain is its genetic malleability. Other than genes associated with capsule synthesis, the genes encoding several putative virulence functions are present in the R6 genome.[2]

1.1 *Surface components, virulence factors and multi-drug resistance of S. pneumoniae*

Three major surface layers can be distinguished in their surface: plasma membrane, cell wall (cell wall polysaccharides (CWPS) and peptidoglycan cell wall) and capsule. The cell wall consists of a triple-layered peptidoglycan backbone that anchors the capsular polysaccharide and also possible proteins. The capsule is the thickest layer, completely concealing the inner structures in exponentially growing pneumoniae. Although CWPS is common to all pnemococcal serotypes, chemical structure of the polysaccharide capsule is serotypic specific.[3] After Avery's experiment, the capsule has long been recognized as the major virulence factors of *S. pneu-*

*For correspondence

moniae. Experimental proof for this was provided by the difference in 50% lethal dose between capsulated and non-capsulated strains. Capsulated strains were found to be at least 10^5 times more virulent than strains lacking the capsule. The chemical structure of the capsular polysaccharide and, to a lesser extent, the thickness of the capsule determine the differential ability of serotypes to survive in the blood stream and possible to cause invasive disease.[3]

The preliminary identification of the pneumococcal surface proteins are done by computational analysis of the genomic sequences of *S. pneumoniae*.[1,2] Then the subsequent study[4,5] indicates that certain pneumococcal proteins including pneumococcal surface protein (pspA), autolysin (lytA), hyaluronate lyase (hyl), pneumolysin (ply), neuraminidases A and B (nanA and nanB), choline binding proteinA (cbpA) are important virulent factors and these could be used as potential vaccine candidates. Since 1990, the number of penicillin-resistant strains has increased and many strains are now resistant to commonly prescribed antibiotics such as penicillin, macrolides and fluoroquinolones.[1] Because of the multi-drug resistance nature of the pneumococcal strains, we need the deeper understanding of the virulence factors. For that, the comparative analysis of genes and proteins may provide more insight on their resistance nature and virulence factors. Therefore, the availability of sequence data for the strains TIGR4 and R6 provides a unique opportunity to compare their genes and proteins for the comparison of virulence nature between the two strains.

2. Materials and methods

In order to retrieve the complete genome sequences, annotated gene and protein sequences list of TIGR4 and R6, we have used NCBI-FTP server (ftp://ftp.ncbi.nih.gov/genomes). For the comparison of two genomes, we have used the genome tools (all genes in a genome, role category pie chart, DNA molecule information and genome properties) and comparative tools (multi-genome homology comparison and align genome – MUMmer) from Comprehensive Microbial Resource (CMR) (http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi).

3. Results and discussion

Comparative genomics and *in silico* studies have begun to reveal insights into gene and protein functions of many bacterial species and strains. In our present work, we would like to consider the comparison of genome features, the whole genome alignment and comparison of gene role category particularly virulence factors between two strains, TIGR4 and R6 of *S. pneumoniae*. Here we have made the comparative study by using the available public domain databases and tools and the results are discussed below.

3.1 *Comparison of the genome features*

The genome sequence of an organism provides information about the size of the genome, base composition, complete gene content, number of RNAs, number of direct and inverted repeats and other features. The genome features of two strains are obtained from CMR and Center for Biological Sequence analysis and the results are provided in table 1 for comparison. This will provide the understanding of general features of the two strains of *S. pneumoniae* in their genome level. Among the two strains, TIGR4 is the largest based on their genome size (table 1). As the *S. pneumoniae* has approximately 60% of AT and 40% of GC content in its genome, the frequency of the aminoacids like Leucin is very high and Alanine, Isoleucine, Glutamic acid, Valine, Lysin are also high.[6] Global repeats and local repeats are comparatively high in TIGR4 (table 1); this may be the duplicated regions of the chromosome. The larger amount of global repeats in TIGR4 reflects larger amount of transposable elements in its genome. High local repeats indicate that high rate of mutation in its genome.

3.2 *Whole genome alignment*

Generally, the genomes of prokaryotes are very dynamic, with insertions, deletions, inversions, and translocations commonly observed among related species or even between different strains of the same species.[7] Here we analyse the genome stability between TIGR4 and R6 using the genome alignment tool MUMmer and the stability of the gene order in the genomes is high and very few proteins are diversely spotted as shown in figure 1. Figure 1 shows that all the genes and protein sequences of two strains are more or less similar; this is not surprising that because all strains occupy the same niche in the human respiratory system. Then the differences might have arisen after the divergence of this strains from other

Table 1. Comparison of the genome features of two strains TIGR4 and R6 of *S. pneumoniae.*

Genome information and features	TIGR4 (capsulated)	R6 (non-capsulated)
Sequencing center	TIGR	Eli Lilly
Genbank accession	AE005672.1	AE007317
Refseq	NC_003028	NC_003098
Topology	Circular	Circular
Molecule	DsDNA	DsDNA
Contig	1	1
Genome size	2·16 Mb	2·03 Mb
Sequence length	2160842 bp	2038615 bp
Number of A	653880 (30·26%)	615270 (30·18%)
Number of T	649168 (30·04%)	613689 (30·10%)
Number of G	430998 (19·95%)	406018 (19·91%)
Number of C	426796 (19·75%)	403638 (19·79%)
No of A + T	60·30%	60·28%
No of G + C	39·69%	39·70%
Mol. weight of DNA (ss)	654868889 dlt	617827172 dlt
Mol. weight of DNA (ds)	1309788163 dlt	1235710019 dlt
Number of primary annotation coding bases	1885084 bp (87·23%)	1761157 bp (86·38%)
Number of genes	2234	2219
Number of genes assigned to role ids	1506 (67·41%)	1313 (59·17%)
Number of genes not assigned to role ids	0%	167 (7·38%)
Structural RNAs	70	73
tRNA genes	58	58
rRNA genes	12	12
scRNA	–	1
rnpB	–	1
ssrA	–	1
Pseudo genes	109	None
Global directed repeats	8·30%	5·70%
Global inverted repeats	7·00%	5·40%
Local directed repeats	6·40%	5·80%
Local inverted repeats	4·30%	4·20%

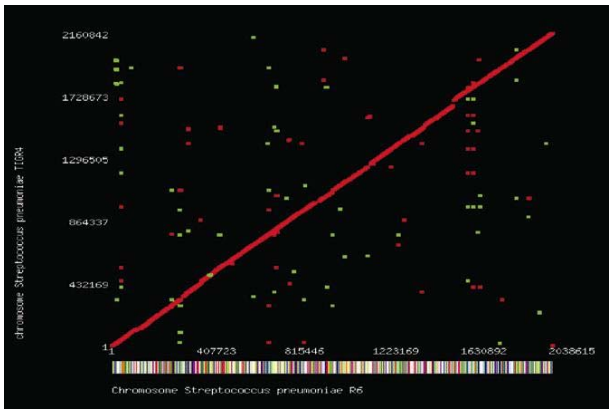

Figure 1. Whole proteome alignment based on whole genome of TIGR4 and R6 using MUMmer. Plot shows plasticity and stability in gene order between two strains.

evolutionary lineages for adaptations in their host, these increase greatly in frequency in pathogens and appear to be associated with the ability to infect eukaryotes, perhaps reflecting a mechanism for evading host immune defenses.[7]

3.3 *Gene role category comparison*

In role category, the genes responsible for biosynthesis of co-factors, prosthetic groups and carrier, fatty acid and phospholipid metabolism, protein fate, protein synthesis, purine pyrimidine synthesis, regulatory functions, transcription, transport and binding proteins of TIGR4 are nearly same as in R6, this suggests that the basic complement of proteins required for certain cellular processes in two strains (table 2). Major cellular systems and features of TIGR4 that are notably different include the genes involved in amino acid biosynthesis, cell envelope, cellular processes, central intermediary metabolism, disrupted reading frame, energy metabolism, hypothetical, conserved hypothetical, mobile and extra chromosomal element, signal transduction, unclassified, unknown and viral functions from the genome R6. This suggests that, these proteins are important

Table 2. Gene role category comparison of TIGR4 and R6 using role category pie chart of CMR. For all organisms, the numbers of pathogenic responsible genes (sl. no. 14) are given as 0 and we manually counted the genes to be 113 and 47 for TIGR4 and R6 respectively.

Sl. no.	Gene role category	TIGR4 (%)	R6 (%)
1·	Amino acid biosynthesis	53–2·37	98–4·79
2·	Biosynthesis of co-factors, prosthetic groups, and carriers	42–1·88	47–2·30
3·	Cell envelope	136–6·08	94–4·60
4·	Cellular processes	147–6·58	75–3·67
5·	Central intermediary metabolism	11–0·49	93–4·55
6·	Disrupted reading frame	92–4·11	0–0
7·	DNA metabolism	92–4·11	104–5·09
8·	Energy metabolism	143–6·40	193–9·44
9·	Fatty acid and phospholipids metabolism	23–1·02	34–1·66
10·	Hypothetical proteins	431–19·20	118–5·77
11·	Hypothetical–conserved protein	302–13·50	416–20·30
12·	Mobile and extra chromosomal element functions	134–5·99	80–3·91
13·	Protein fate	70–3·13	68–3·32
14·	Pathogen responses	113–5·06	47–2·30
15·	Protein synthesis	120–5·37	127–6·21
16·	Purines, pyrimidines nucleosides and nucleotides	54–2·41	61–2·98
17·	Regulatory functions	121–5·41	122–5·97
18·	Signal transduction	79–3·53	4–0·19
19·	Transcription	29–1·29	31–1·51
20·	Transport and binding proteins	267–11·90	235–11·50
21·	Unclassified	0–0	196–9·59
22·	Unknown function	174–7·78	51–2·49
23·	Viral functions	0–0	25–1·22

Table 3. Comparison of common virulence factors between TIGR4 and R6.

Strains	Gene ID	VF*	GC%	Protein length	Gene length	Coordinates 5′	Coordinates 3′	Identity (%)
TIGR4	gi|15900059	pspA	40·23	744	2235	118423	120657	VF* of TIGR4 are taken as
	gi|15901761	lytA	46·44	318	957	1841361	1840405	references and are aligned
	gi|15900247	hysA	40·15	1066	3201	287483	290683	with the same kind of
	gi|15901747	ply	41·83	471	1416	1833311	1831896	sequences of R6 using
	gi|15901180	nanA	35·36	740	2223	1251631	1249409	LALIGN with default
	gi|15901522	nanB	33·38	697	2094	1589236	1587143	parameters.
	gi|15901997	cbpA	41·90	693	2082	2122806	2121460	
R6	gi|15902165	pspA	42·65	619	1860	128356	130215	53·6
	gi|15903796	lytA	46·54	318	957	1723025	1722069	99·7
	gi|15902330	hysA	40·01	1078	3237	285103	288339	97·8
	gi|15903781	ply	42·04	471	1416	1715341	1713926	99·8
	gi|15903579	nanA	42·67	1035	3108	1517937	1514944	19·6
	gi|15903574	nanB	33·43	697	2094	1510307	1508214	99·1
	spr1995	cbpA	41·32	701	2106	1989649	1987544	73·7

*VF, Virulence factors

for strain uniqueness and they may be involved in variations in pathogenesis (table 2).

3.4 *Common virulence factors comparison*

We compared the virulence factors of these two strains and the results are tabulated in table 3. GC content of the individual virulence factors varies from 33 to 46%. The length of the protein and gene sequences is more or less similar and their position varies according to their genome size. Based on the identities, the protein sequence of pspA of TIGR4 has 53·6% with R6. It seems that the effect of virulence nature of pspA of TIGR4 is 50% higher than

Table 4. Comparison of capsular polysaccharide sequences between TIGR4 and R6.

Strain name	Gene ID	GC%	No of amino acids	Gene length	Coordinates	Identity
TIGR4	gi|15900046-putative (SP0103)	42·21	616	1851	104668–106518	99·8%
	gi|15900275-cps4A (SP0346)	38·32	481	1446	320077–321522	17·8%
	gi|15900276-cps4B (SP0347)	41·98	243	732	321524–322255	13·0%
	gi|15900277-cps4C (SP0348)	40·29	230	693	322264–322956	12·7%
	gi|15900278- cps4D (SP0349)	34·21	227	684	322966–323649	9·9%
	gi|15900279-cps4E (SP0350)	33·49	211	636	323990–324625	11·2%
	gi|15900280-cps4F (SP0351)	33·17	409	1230	324634–325863	15·4%
	gi|15900281-cps4G (SP0352)	27·84	358	1077	325868–326944	16·2%
	gi|15900282-cps4H (SP0353)	31·36	372	1119	326937–328055	15·7%
	gi|15900287-cps4J (SP0358)	38·46	351	1056	332875–333930	19·7%
	gi|15900288-cps4K (SP0359)	36·19	409	1230	334030–335259	15·4%
	gi|15901666-putative (SP0907)	28·79	455	1368	859370–860737	15·1%
R6	gi|15902136-capD (Spr0092)	42·26%	616	1851	99217–101067	Reference sequence

R6. Identities of lytA and hysA between TIGR4 and R6 strains have 99·7% and 97·8%, respectively. So that, the virulence effects of autolysin and hyaluronidase are same in R6 as like TIGR4. Exact sequence of ply (99·8%) exists both in TIGR4 and R6 seems that, the functions of pneumolysin of R6 are exactly similar like TIGR4. Since the percentage identity of cbpA between TIGR4 and R6 is 73·4%, the virulence effect of cbpA is higher in TIGR4 than R6. The virulence effect of nanA between two strains may vary because the existence of less percent identity (19·5%) indicates that, this large variation may be responsible for the avirulent nature of R6.

The study of capsular polysaccharide is important to know encapsulation (TIGR4) and non-encapsulation (R6) of *S. pneumoniae*. We have analysed capsular polysaccharides of the two genomes and the results are given in table 4. Since TIGR4 has 12 different genes involved in capsular polysaccharide biosynthesis and R6 has only CapD gene, showing the importance of capsules in virulence nature of TIGR4. Further, we find that CapD gene of R6 is identical (99·8%) to the putative capsular polysaccharide biosynthesis protein (gi|15900046) of TIGR4.

4. Conclusions

Based on the above comparative analysis with special reference to the virulence nature, we have found that the high global repeats reflect larger amount of transposable elements and high local repeats indicate the higher rate of mutation in the genome of TIGR4 than R6. Genome comparisons reveal that the two strains occupy the same niche and the differences might have arisen after their divergence. The differences in the major cellular systems including hypothetical sequences suggest that these proteins are important for strain uniqueness and they may be responsible for pathogenetic variations. The pspA of R6 has 50% effect of TIGR4 and nanA of TIGR4 has very low (19.5%) percent identity with R6 indicating that the difference may be one of the reasons for avirulence nature of R6. TIGR4 has 12 different genes involved in capsular polysaccharide biosynthesis but R6 has only CapD gene, showing the importance of capsules in virulence. Thus, there is an indication that comparative analysis of genome sequences will reveal the virulence nature between the two strains of *S. pneumoniae*. Additional supportive evidence for the virulence nature in TIGR4 is sought through further analysis with other available completed genomes D39 and G54 in addition to R6 strain of *S. pneumoniae*, which will be published elsewhere.

References

1. Tettelin H *et al* 2001 *Science* **293** 498
2. Hoskins J *et al* 2001 *J. Bacteriol*. **183** 5709
3. AlonsoDeVelasco E, Verheul A F, Verhoef J and Snippe H 1995 *Microbiol. Rev.* **59** 591
4. Jedrzejas M J 2001 *Microbiol. Mol. Biol. Rev.* **65** 187
5. Rigden D J, Galperin M Y and Jedrzejas M J 2003 *Crit. Rev. Biochem. Mol. Biol*. **38** 143
6. Fraser C M, Eisen J, Fleischmann R D, Ketchum K A and Peterson S 2000 *Emerg. Infect. Dis.* **6** 505
7. Hughes D 2000 *Genome Biol*. 1(6) reviews 0006.1