Sādhanā Vol. 31, Part 2, April 2006, pp. 105–115. © Printed in India

Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to *Vibrio* genomes

SANJAY NAG[1], RAGHUNATH CHATTERJEE[1],
KEYA CHAUDHURI[1] and PROBAL CHAUDHURI[2,*]

[1]Human Genetics & Genomics Group, Indian Institute of Chemical Biology,
Jadavpur, Kolkata 700 032, India
[2]Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203,
BT Road, Kolkata 700 108, India
e-mail: probal@isical.ac.in

Abstract. *Vibrio cholerae, Vibrio vulnificus, Vibrio parahaemolyticus* and several other related *Vibrio* species show distinctly similar two-chromosomal genome organization. However, the modes of pathogenicity are very different among these species, and this is largely attributed to externally acquired genetic elements. We develop some statistical methods to determine these external genetic elements or genomic islands in genomes based on their differential oligonucleotide usage patterns compared to the rest of the genome. Genomic islands identified by these unsupervised statistical methods include integron and pathogenicity islands. After statistical determination of the genomic islands, we investigate their gene contents and their possible association with the pathogenic behaviour of the corresponding *Vibrio* species. These investigations lead to observations that are of evolutionary and biological significance.

Keywords. Dendrogram; hierarchical clustering; horizontally acquired genes; oligonucleotide distributions; statistical test of significance; transposons.

1. Introduction

Identification of genomic islands in prokaryotic genomes has received considerable attention in the literature due to possible connections of those genomic islands to pathogenic virulence and antibiotic resistance of micro-organisms. Sometimes genomic islands contain horizontally transferred genetic materials that are of critical importance in the evolution of prokaryotic organisms and pathogenic behaviour.

The family Vibrionaceae of bacteria includes several pathogens of human and fish. The most notable member of this family is *Vibrio cholerae* (Heidelberg *et al* 2000), the etiological agent of epidemic cholera, a severe and sometimes lethal diarrheal disease. Other important

[*]For correspondence

members of this family are *Vibrio parahaemolyticus* (Makino *et al* 2003) and *Vibrio vulnificus* (Chen *et al* 2003). It is well known that the genome organization of these three *vibrios* and several related *Vibrio* species are generally very similar (Tagomori *et al* 2002). The genomes are distributed unequally between two circular chromosomes (Trucksis *et al* 1998). The larger chromosome (Chr-I) is of comparable size in all these three *vibrios* (Tagomori *et al* 2002; Iida 2003). Furthermore, the origin of replication of the small chromosome (Chr-II) is conserved among at least three other *vibrio* species (Thompson *et al* 2004). Similarity in the genome structure and sharing of the same characteristics for origin of replication suggest that these *vibrios* might have descended from the same common ancestor (Egan & Waldor 2003).

However, the modes of pathogenicity are very different among these three species. *V. parahaemolyticus* gastroenteritis is associated with the synthesis and secretion of thermostable direct haemolysin (Fabbri *et al* 1999; Makino *et al* 2003). *V. vulnificus* is an etiological agent for severe human infection acquired through wounds or contaminated seafood. A hallmark of *V. vulnificus* infection is fulminant reaction caused by the invading bacteria in connective tissues displayed as blisters and hemorrhagic necrosis (Chen *et al* 2003; Chiang & Chuang 2003). Biochemical and genetic studies suggest that the extra-cellular proteins released by the invading bacteria mediate the pathogenetic process of penetrating cellular barriers, vascular dissemation, and local destruction of affected tissues (Linkous & Oliver 1999). *V. cholerae* infection, on the other hand, is noninvasive in nature (Iida 2003). The mode of invasion and pathogenicity of *V. cholerae* is dependent on external phage-mediated toxins, which are externally acquired in the *Vibrio* genome (Boyd & Waldor 1999). In fact, there is a pathogenicity island, coding for genes required for pathogenicity, which is horizontally acquired (Li *et al* 2003). So it appears that the externally acquired parts of the genome are closely involved with the pathogenicity of different *Vibrio* species.

Karlin (2001) attempted to detect gene clusters and pathogenicity islands in different bacterial genomes based on statistical differences in codon usage. There are other recent attempts to identify genomic islands in bacterial genomes based on GC contents, di-nucleotide frequencies, codon usage bias and amino acid usage bias (Hsiao *et al* 2003; Tu & Ding 2003; Zhang & Zhang 2004). The present study aims at determining segments of *Vibrio* chromosomes each of which exhibits statistically significant differences in oligonucleotide usage patterns compared to the rest of the chromosome to which it belongs and also the other chromosome. Such segments are likely to contain horizontally acquired genetic material and might have critical relations to the cause and nature of pathogenicity of different *Vibrio* species. Unlike the supervised statistical methods like discriminant analysis used by some of the earlier authors (Tu & Ding 2003) that require training data sets, our method is fully unsupervised in nature. Such an unsupervised clustering approach has been used in a different context, where the population dynamics of pathogenic clones of *Staphylococcus aureus* has been analysed by high-throughput amplified fragment length polymorphism (AFLP) (Melles *et al* 2004). A genuine difficulty in formation of training datasets required for any supervised statistical methods for identifying genomic islands is that the genomic islands that are known *a priori* and available for an organism would typically be very few. Earlier authors (Tu & Ding 2003) using supervised statistical techniques tried to cope with this problem by using a training dataset formed by the aggregation of known genomic islands from different organisms. Using such an aggregated training dataset may not be appropriate unless there are several organisms with some statistical similarities in their genome sequences as well as in their known genomic islands, and clearly this may not often be achievable in practice. Further, since our method is based only on the distributions of oligonucleotides and does not depend on codon or amino acid usage biases that have been used by some of the above mentioned authors, it does not require knowledge of annotation of genes, making it a much simpler procedure, which is applicable to identify even those genomic islands that contain very little or no protein coding gene sequences. However, after the method is applied to some *Vibrio* genomes and some genomic islands are identified using our statistical techniques, the gene contents of these segments are carefully examined for the presence of possible pathogenic elements. This leads to useful insights into the evolution of the chromosomes and pathogenic behaviour of different *Vibrio* species.

2. Description of data and statistical methodology

2.1 *Genome sequences*

The genome sequences of *V. cholerae* N16961 (El Tor, O1, Str^{R}), *V. parahaemolyticus, V. vulnificus* CMCP6 and *V. vulnificus* YJ016 and related information are available in the website, www.tigr.org. From that web site the complete genome sequences of the two chromosomes of the four *Vibrios* were downloaded and used for further analysis.

2.2 *Subdivision of the genome and statistical determination of genomic islands*

In order to determine genomic islands in a chromosome, each chromosome of a *Vibrio* species needs to be taken separately and divided into smaller sections. As transposons are known to be involved in horizontal acquirement into the genome (Beaber *et al* 2002), we chose to divide the genome according to the presence of transposable elements. In other words, the stretch of a chromosome from the start of a transposon sequence to the start of the next transposon sequence was considered as a fragment to begin with. The transposon sequences inserted in the genome of the *Vibrio* were determined from the website www.tigr.org. Earlier authors (Hsiao *et al* 2003; Tu & Ding 2003) used window-based segmentation of the genome for the identification of genomic islands, where the window sizes were chosen subjectively. If we attempt to use a range of possible window sizes and compile the results, the method becomes computationally very expensive while the choice of the range of possible window sizes still remains arbitrary. Instead of these, we prefer to use a method of segmentation that is motivated by biological considerations. Other methods of segmentation guided by biological considerations could be to use flanking repeats and proximal tRNAs instead of transposons but we do not intend to pursue that in this paper.

Since a DNA sequence is formed using an alphabet of four letters denoting four DNA bases: adenine (A), thymine (T), cytosine (C) and guanine (G), the simplest form of statistical analysis can be based on various frequencies of DNA *k*-words, which are *k*-tuples ($k \geq 1$) formed using these four letters. For example, if a sequence runs like TTTGCGCGTGCGT . . . , the first 4-word is TTTG, the second is TTGC, the third is TGCG and so on (Chaudhuri & Das 2001). The relative frequencies of different *k*-words for a specific *k* can be determined for the chromosomal fragments of the *Vibrios*. These word frequencies for the segments of *Vibrio* chromosomes, which were obtained using transposons as mentioned above, were determined by the software SWORDS (Chaudhuri & Das 2002). Standard hierarchical average linkage cluster analysis was carried out among the chosen segments of a specific chromosome of a *Vibrio* species. The distance matrix required for such cluster analysis was computed from the absolute differences in the word frequencies between a pair of segments under comparison (Basu *et al* 2003). The software SWORDS (Chaudhuri & Das 2002) performed the cluster analysis and constructed the dendrogram tree (Duda *et al* 2001; Everitt *et al* 2001). Those fragments having very different word usage compared to the other fragments of the same chromosome then branch out and show up as genomic island fragments in the dendrogram tree.

Analysis of short oligonucleotide combinations is recognized to be affected by the biases in nucleotide composition and organization in prokaryotic genomes. For instance, selective pressures as a result of di-nucleotide stacking, DNA conformational tendencies, DNA replication and repair mechanisms, or selection by restriction endonucleases (Gelfand & Koonin 1997) may influence di-nucleotide frequencies represented as genomic signature of the organism (Karlin 1998). Codon usage, which affects translational efficiency, also influences genomic trinucleotide usage (Pride *et al* 2003). However, only through analysis of longer oligonucleotides, biases beyond di-nucleotide frequencies and codon usage preferences, which are also of biological importance, can be identified (Pride & Blaser 2002). When we considered di-nucleotide and tri-nucleotide frequencies in *Vibrio* genomes they did not lead to any interesting results. This is why we had to go beyond di- and tri-nucleotides. Similar advantages of tetra-nucleotide frequencies over mono-, di- and tri-nucleotide frequencies have been observed by earlier authors (Chaudhuri & Das 2001) in the evolutionary study of different species using markers such as *r*RNA and mitochondrial DNA. We have looked into oligonucleotides as large as octa-nucleotides and found that the results obtained from tetra- through octa-nucleotides are fairly similar, and they are in conformity with one another.

To check for the statistical significance of the genomic islands thus generated, up to 1000 random genome segments with the same size as that of an identified genomic island were cut out from the same chromosome excluding its genomic island(s). Then a set of distances between the identified genomic islands and these 1000 random segments were determined based on their oligonucleotide usage patterns in the same way as in the cluster analysis described above. Further, a second set of pair-wise distances between those randomly picked genome segments was also formed in the same way. Then some standard statistical tests, namely Fisher's t-test and Wilcoxon–Mann–Whitney rank tests were carried out to compare those two sets of distances. All statistical tests were performed using publicly available *R* software (http://www.r-project.org/). Note that the statistical testing procedure used here is completely model-free in nature as it does not require any model for the genome sequence, and consequently is applicable to genome sequence in any organism. This is an advantage as the modelling of the genome sequence of an organism is not an easy task if at all possible. Besides even if there are reasonable models for genome sequences, such models are likely to be different for different organisms leading to different model-based statistical tests if there are at all such feasible tests. It is appropriate to note here that some of the earlier methods for identifying genomic islands such as those based on GC contents (Hsiao *et al* 2003; Tu & Ding 2003; Zhang & Zhang 2004) are not based on any formal statistical test and thresholding is done somewhat arbitrarily there.

3. Results from statistical analysis

Using the cluster analysis of the chromosomal fragments, as described in the preceding section, the dendrogram for each *Vibrio* species has been drawn (figures 1–4) based on distances computed using tetra-nucleotide (i.e., $k = 4$) frequencies. In the dendrogram corresponding to *V. vulnificus* CMCP6, six fragments stretching from locus VV12451 and ending in VV12551 in Chr-I form a clear genomic island cluster (figure 1). The dendrogram for *V. vulnificus*

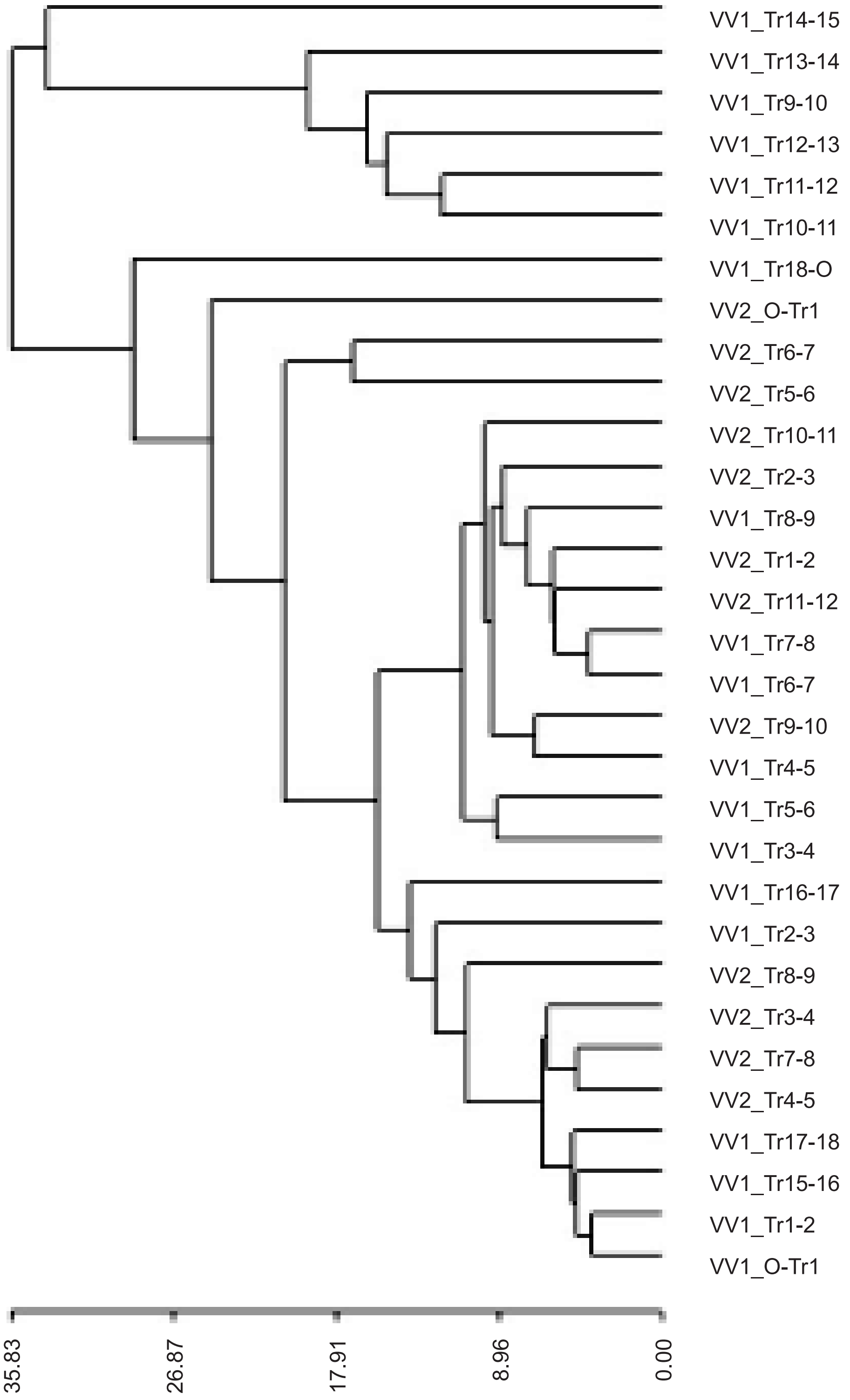

Figure 1. *V. vulnificus* CMCP6.

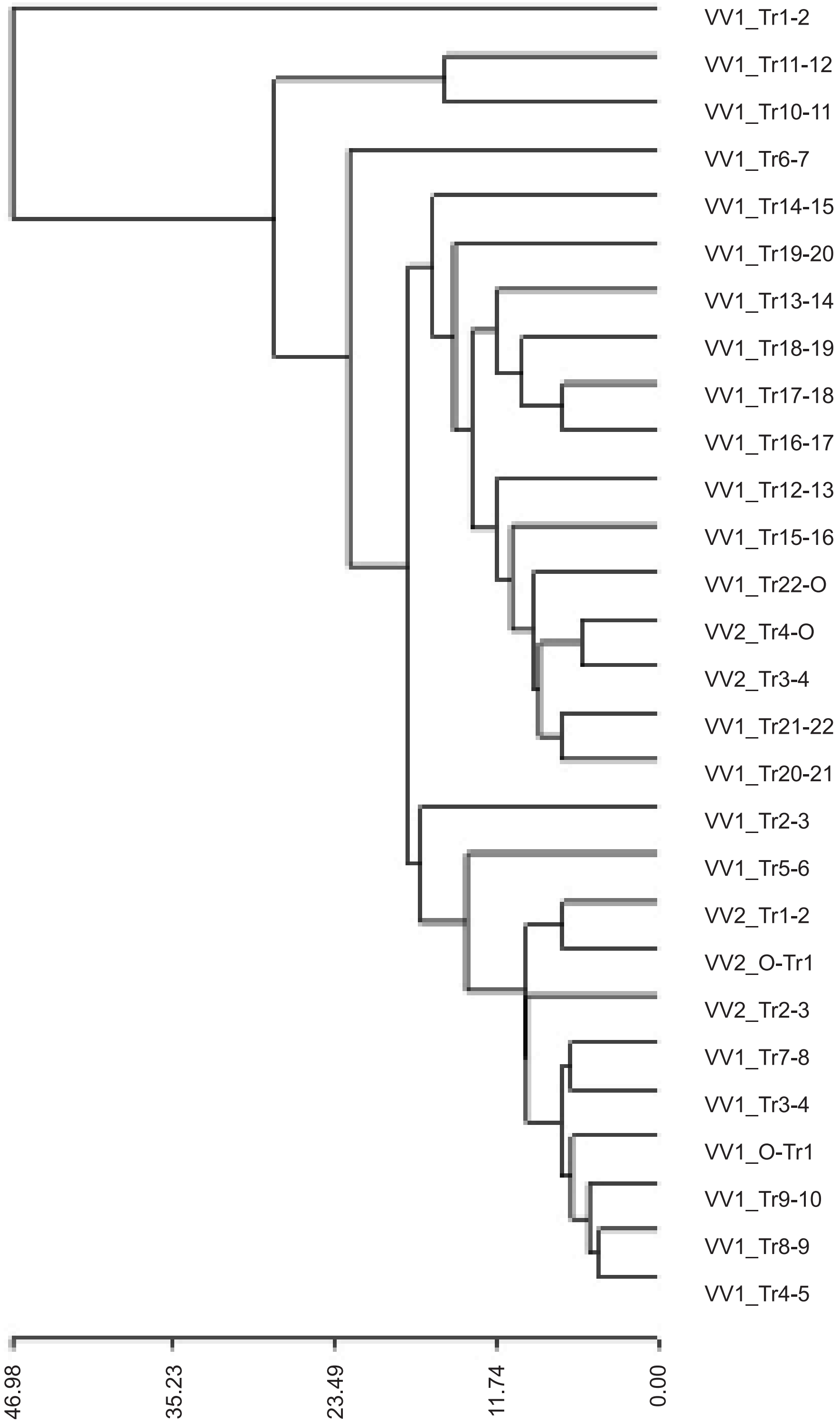

Figure 2. *V. vulnificus* YJ016.

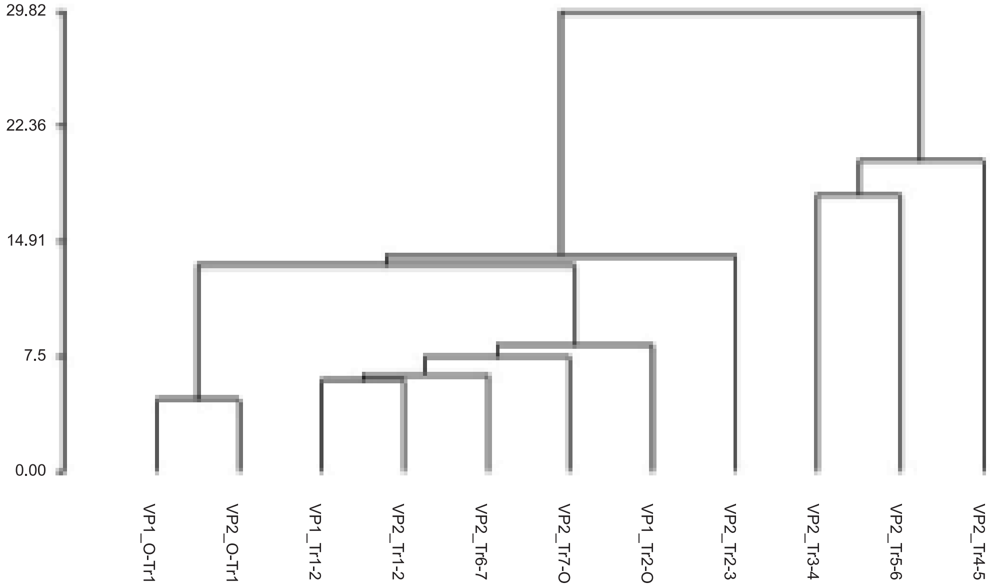

Figure 3. *V. parahaemolyticus.*

YJ016 has two genomic islands, all in Chr-I. The first genomic island has one fragment and stretches from locus VV0152 to VV0160 and the second has two fragments and stretches from VV1704 to VV1947 (figure 2). In the case of *V. parahaemolyticus*, the dendrogram shows a distinct genomic island formed by a cluster of three fragments from Chr-II stretching from the locus VPA1311 to the locus VPA1396 (figure 3). Finally, in the case of *V. cholerae* two genomic islands are visible in the dendrogram. One segment, consisting of four fragments, is located in Chr-II, and it stretches from the locus VCA0283 to the locus VCA0508. The other genomic island of *V. cholerae* genome is located in Chr-I, and it stretches from the locus VC0818 to the locus VC870 (figure 4). As for our analysis, we have split the whole genome based on transposon locations, in all the figures, instead of locus names, we have labeled the fragments using transposon (Tr) and origin of replication (O), e.g., VC1_Tr3-4, represent a fragment of *Vibrio cholerae* Chr-I from the third transposon (VC0818) through the fourth transposon (VC0870).

All the genomic islands identified in the dendrograms (figures 1–4) were subject to the statistical tests described at the end of the preceding section. The P-values turned out to be almost zero for all such tests corresponding to different identified genomic island. In all cases, the distances of randomly picked segments with a genomic island were significantly higher than the distances among randomly picked segments. This confirms significantly different oligonucleotide usage patterns in each of the identified genomic island compared to the rest of the chromosome. The identified genomic islands, their locations and relative sizes are summarized in table 1.

In the next step of our analysis, we carefully investigated the gene contents of each of the genomic islands identified above as well as the gene contents of the fragment that appears in the dendrogram as the neighbour of the cluster forming the genomic island. Our findings are summarized in the next section.

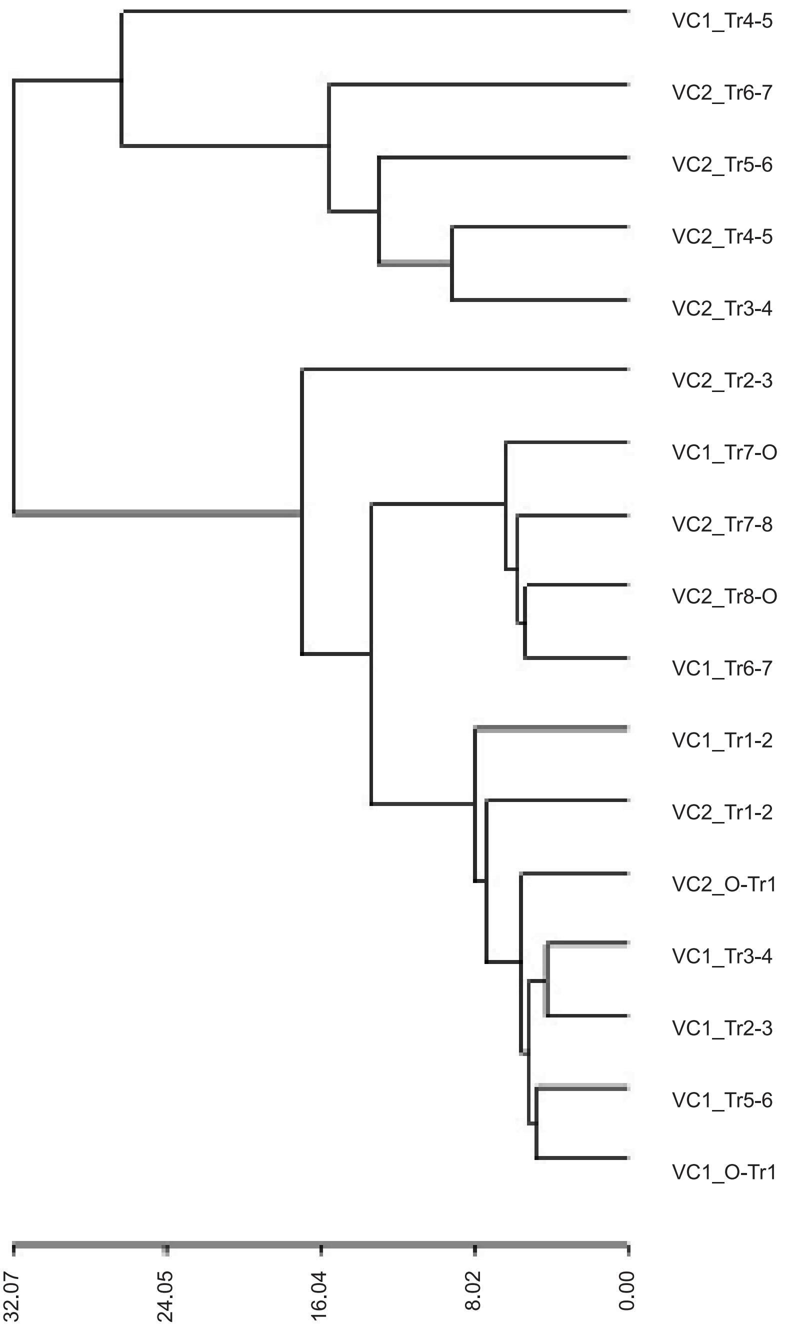

Figure 4. *V. cholerae* N16961.

Table 1. Positions of each identified genomic island in *Vibrio* genomes and its relative size (%) compared to the whole genome.

Chromosome	*V. vulnificus* CMCP6	*V. vulnificus* YJ016	*V. parahaemolyticus*	*V. cholerae* N16961
Chr-I	2490347–2590685 (1.96%)	159732–167647 (0.15%) 1757361–1936159 (3.41%)	874931–936351 (1.52%)	-
Chr-II	-	-	1389397–1467005 (1.50%)	302647–436629 (3.32%)

4. Analysis of the gene contents of identified islands and their neighbours in the dendrogram

In the genome of *Vibrio vulnificus CMCP6*, the identified genomic island is composed of the genome fragments VV1_Tr9-15. The island has a few acetyltransferase genes (VV12454, VV12475, VV12479, VV12480, VV12482, VV12491, VV12496, VV12506, VV12510, VV12514, VV12524, VV12537, VV12540, VV12544) (see for example, Doublet *et al* 2004). Interestingly, these genes are proposed to be acquired by horizontal gene transfer and are included in the HGT-database (Garcia-Vallve *et al* 2003; http://www.fut.es/∼debb/HGT/). A few more genes with external origin located here are as follows: lactoglutathione lyase (VV12461, VV12494, VV12549), prophage antirepressor (VV12522), plasmid stabilizing agent ParE (VV12525). This genomic island has not been studied much in the existing literature so far though a genomic island, which has substantial overlap with it, was identified by the method of Zhang & Zhang (2004). The immediate neighbouring genomic fragment to this cluster in the dendrogram (figure 1) is VV1_Tr18-O, and it includes the origin of replication of Chr-I. As the origin is conserved in a genome, and any change near the origin can be detrimental to the cell, we can rule out the possibility of it being part of a genomic island.

In the genome of *Vibrio vulnificus YJ106*, two genomic islands VVY1_Tr1-2 and VVY1_Tr10-12 have been identified. The first island has lactoglutathione lyase (VV0155). The second island represents the super integron containing plasmid stabilizing system protein ParE (VV1867), and acetyl transferase (VV1879, VV1886, VV1893, VV1904, VV1907, VV1910, VV1918, VV1928), lactoylglutathione lyase (VV1937), super-integron integrase IntIA (VV1941) (see Chen *et al* 2003; Garcia-Vallve *et al* 2003). In the dendrogram (figure 2), the genomic fragment that appears next to these two islands is VVT_Tr6-7. It contains genes like transcriptional activators (VV0682 and VV0683) and also ribosomal protein S20 (VV0684). So, the chance of this fragment being an externally acquired fragment is quite remote.

In the genome of *Vibrio parahaemolyticus*, the genomic island is located in the region VP2_Tr3-6. The genomic island contains potentially acquired genes like thermostable direct hemolysin (VPA1314, VPA1378), cytotoxic necrotizing factor (VPA1321), components of type III secretion system (VPA1335, VPA1339, VPA1342, VPA1349, VPA1354, VPA1355, VPA1367) (see Makino *et al* 2003), and a large number of hypothetical proteins. In the dendrogram (figure 3), the immediate neighbouring fragment of the genomic island is VP2_Tr2-3. This fragment represents 6% of the total genome, and it is a fairly big chunk, which is unlikely to be acquired totally from outside. The region mainly consists of hypothetical proteins and regulators like the cold-shock transcriptional regulator CspA (VPA1289) etc. There is no known gene in this region, which can be confirmed to be externally acquired.

The genomic islands detected in the *Vibrio cholerae* genome represent the pathogenicity island (VC1_Tr4-5) and the integron island (VC2_Tr3-7). The integron island has potentially acquired genes like chloramphenicol acetyltransferase (VCA0300), putative killer protein and antidote protein (VCA0391-VCA0392), haemagglutinin (VCA0446-VCA0447), as well as a few copies of acetyltransferase. Besides, there are many hypothetical and conserved hypothetical proteins. The immediate neighbouring fragment of the genomic islands in the dendrogram (figure 4) is VC2_Tr2-3. This fragment consists of genes coding for enzymes involved in amino acid biosynthesis and these are mostly constitutive enzymes. It is very unlikely that such genes would be acquired and these genes are not listed in the HGT-database.

By comparing the gene contents of genomic islands in different *Vibrio* genomes, it appears that these genes are involved in species-specific virulence and survival.

This research was supported in part by grants from the Council of Scientific and Industrial Research and Department of Biotechnology, Government of India. Authors thank an anonymous referee whose comments led to significant improvement over an earlier version of the paper.

References

Basu S, Burma D P, Chaudhuri P 2003 Words in DNA sequences: some case studies based on their frequency statistics. *J. Math. Biol.* 46: 479–503

Beaber J W, Hochhut B, Waldor M K 2002 Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae. J. Bacteriol.* 184: 4259–4269

Boyd E F, Waldor M K 1999 Alternative mechanism of cholera toxin acquisition by *Vibrio cholerae*: generalized transduction of CTXPhi by bacteriophage CP-T1. *Infect. Immunol.* 67: 5898–5905

Chaudhuri P, Das S 2001 Statistical analysis of large DNA sequences using distribution of DNA words. *Curr. Sci.* 80: 1161–1166

Chaudhuri P, Das S 2002 SWORDS: a statistical tool for analysing large DNA sequences. *J. Biosci.* 27: 1–6

Chen C Y *et al* 2003 Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res.* 13: 2577–2587

Chiang S R, Chuang Y C 2003 *Vibrio vulnificus* infection: clinical manifestations, pathogenesis, and antimicrobial therapy. *J. Microbiol. Immunol. Infect.* 36: 81–88

Doublet B, Weill F X, Fabre L, Chaslus-Dancla E, Cloeckaert A 2004 Variant Salmonella genomic island 1 antibiotic resistance gene cluster containing a novel 3'-N-aminoglycoside acetyltransferase gene cassette, aac(3)-Id, in Salmonella enterica serovar newport. *Antimicrob Agents Chemother.* 48: 3806–3812

Duda R O, Hart P E, Stork D G 2001 *Pattern classification* (New York: John Wiley)

Egan E S, Waldor M K 2003 Distinct replication requirements for the two *Vibrio cholerae* chromosomes. *Cell* 114: 521–530

Everitt B S, Landau S, Leese M 2001 *Cluster analysis* (New York: Oxford University Press)

Fabbri A, Falzano L, Frank C, Donelli G, Matarrese P, Raimondi F, Fasano A, Fiorentini C 1999 *Vibrio parahaemolyticus* thermostable direct hemolysin modulates cytoskeletal organization and calcium homeostasis in intestinal cultured cells. *Infect. Immunol.* 67: 1139–1148

Garcia-Vallve S, Guzman E, Montero M A, Romeu A 2003 HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.* 31: 187–189

Gelfand M S, Koonin E V 1997 Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25: 2430–2439

Heidelberg J F *et al* 2000 DNA sequence of both chromosomes of the cholera pathogen, *Vibrio cholerae*. *Nature (London)* 406: 477–483

Hsiao W, Wan I, Jones S J, Brinkman F S 2003 IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 19: 418–420

Iida T 2003 [*Vibrios* (*Vibrio cholerae*, *V. parahaemolyticus*, *V. vulnificus*).]. *Nippon Rinsho* 61 (Suppl 3): 722–726

Karlin S 1998 Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1: 598–610

Karlin S 2001 Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.* 9: 335–343

Li M, Kotetishvili M, Chen Y, Sozhamannan S 2003 Comparative genomic analyses of the *vibrio* pathogenicity island and cholera toxin prophage regions in nonepidemic serogroup strains of *Vibrio cholerae*. *Appl. Environ. Microbiol.* 69: 1728–1738

Linkous D A, Oliver J D 1999 Pathogenesis of *Vibrio vulnificus*. *FEMS Microbiol. Lett.* 174: 207–214

Makino K *et al* 2003 Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. *Lancet* 361: 743–749

Melles D C *et al* 2004 Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. *J. Clin. Invest.* 114: 1732–1740

Pride D T, Blaser M J 2002 Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. *J. Mol. Biol.* 316: 629–642

Pride D T, Meinersmann R J, Wassenaar T M, Blaser M J 2003 Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* 13: 145–158

Tagomori K, Iida T, Honda T 2002 Comparison of genome structures of *vibrios*, bacteria possessing two chromosomes. *J. Bacteriol.* 184: 4351–4358

Thompson F L, Iida T, Swings J 2004 Biodiversity of *vibrios*. *Microbiol. Mol. Biol. Rev.* 68: 403–431, table of contents

Trucksis M, Michalski J, Deng Y K, Kaper J B 1998 The *Vibrio* cholerae genome contains two unique circular chromosomes. *Proc. Natl. Acad. Sci. USA* 95: 14464–14469

Tu Q, Ding D 2003 Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol. Lett.* 221: 269–275

Zhang R, Zhang C T 2004 A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics* 20: 612–622