CALL FOR PAPERS | Comparative Genomics

Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis

Ren Zhang1 and Chun-Ting Zhang2

1Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060; and 2Department of Physics, Tianjin University, Tianjin 300072, China

Submitted 6 October 2003; accepted in final form 28 October 2003

Zhang, Ren, and Chun-Ting Zhang. Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis. Physiol Genomics 16: 19–23, 2003. First published November 4, 2003; 10.1152/physiolgenomics.00170.2003.—Horizontal gene transfer has been recognized as a universal event throughout bacterial evolution. The availability of both complete genome sequences of Bacillus cereus and B. anthracis provides the possibility to perform comparative analysis based on their genomes. By using a windowless method to display the distribution of the genomic GC content of B. cereus and B. anthracis, we have found three genomic islands in the genome of B. cereus, i.e., BCGI-1, BCGI-2, and BCGI-3, respectively, which are absent in the genome of B. anthracis. All the genomic islands have abrupt changes in GC content compared with that of surrounding regions. BCGI-1 has many conserved features of genomic islands, e.g., a Val-tRNA gene is utilized as the integration site, and a site-specific recombinase gene is located at the 3’ end. BCGI-2 has a large percentage of phosphate gene, suggesting a phage-related recombination is involved. BCGI-3 contains a ferric anguibactin transport system, which is likely to be involved in the iron transport that enables the bacterium to overcome the iron limitation in the host. In addition, BCGI-3 also contains a cluster of genes related to lantibiotics, which may play a role during the evolution of the genome. Furthermore, the integrations of the genomic islands, BCGI-1 and BCGI-3, result in deletions of DNA sequence fragments; therefore, such integrations lead to both gene gain and gene loss simultaneously.

Horizontal gene transfer has been recognized as a universal event throughout bacterial evolution (9, 14, 15). Genomic islands contain clusters of horizontally transferred genes. Obtaining foreign genes is an effective way to alter the genotype of a bacterium, which may lead to the creation of new traits or even new species (3, 4, 7, 12, 13).

The identification of genomic islands has received intense interest during the past few years. Among the methods to detect the horizontal gene transfer events in bacteria, assessing the change in GC content remains an established and effective way. Usually, as a routine procedure, the distribution of the genomic GC content is calculated by counting the frequency of G and C bases within the sliding windows that move along genomes. However, in this method the window size is difficult to adjust, i.e., large window size leads to low resolution, whereas small window size leads to large statistical fluctuations. Recently, a windowless method to calculate GC content, the cumulative GC profile, was proposed (22). The resolution of the cumulative GC profile in displaying the genomic GC content is high since no sliding window is used. This method has been used to identify genomic islands in the genomes of Corynebacterium glutamicum and Vibrio vulnificus (24). In this brief communication, the cumulative GC profile was used to detect genomic islands in B. cereus, based on comparison with B. anthracis. Consequently, three genomic islands have been identified. One genomic island, BCGI-3, contains a cluster of genes that encode the ferric anguibactin transport system, which may play a role in enabling the bacteria to overcome iron limitation in the host. In addition, BCGI-3 also contains a cluster of genes related to lantibiotics, which may have an impact on the evolution of the genome.

MATERIALS AND METHODS

The complete genome sequences of the bacteria B. cereus ATCC 14579 (AE016877) and B. anthracis (AE016879) were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/).

Using the cumulative GC profile to display the GC content distribution. The Z-curve is a three-dimensional space curve constituting the unique representation of a given DNA sequence in the sense that each can be uniquely reconstructed given the other (23, 25). Based on the Z-curve, any DNA sequence can be uniquely described by three independent distributions, i.e., those of the bases of purine/pyrimidine (xn), amino/keto (yn), and weak/strong hydrogen bonds (zn), respectively. In particular, zn displays the distribution of bases of GC/AT types along the sequence, which is calculated as follows (23, 25)

\[ z_n = (A_n + T_n) - (C_n + G_n), \]

\[ n = 0, 1, 2, \ldots, N, \quad z_n \in [-N, N] \]
where $A_n$, $C_n$, $G_n$, and $T_n$ are the cumulative numbers of the bases $A$, $C$, $G$ and $T$, respectively, occurring in the subsequent from the first base to the $n$-th base in the DNA sequence inspected.

Based on $z_n$, GC content can be calculated using a windowless technique (22). Usually, for an AT-rich genome, $z_n$ is approximately a monotonously increasing linear function of $n$, whereas for a GC-rich genome, $z_n$ is approximately a monotonously decreasing linear function of $n$. To amplify the deviations, the curve of $z_n \sim n$ is fitted by a straight line using the least square technique

$$z = kn \quad (2)$$

where $(z, n)$ is the coordinate of a point on the straight line fitted, and $k$ is its slope. Instead of using the curve of $z_n \sim n$, we will use the $z'$ curve, or cumulative GC profile, hereafter, where

$$z'_n = z_n - kn \quad (3)$$

Therefore, the deviations of $z_n \sim n$ curve from the straight line, which corresponds to a constant GC content (see Eq. 4, below), are protruded by the $z'$ curve. A program to draw the $z'$ curve online is accessible from \texttt{http://tubic.tju.edu.cn/zcurve}. The $z'$ curve and the cumulative GC profile are used interchangeably in this paper.

Let GC denote the average GC content within a region $\Delta n$ in a sequence, then we find from Eqs. 1–3

$$\text{GC} = \frac{1}{\Delta n} \left( 1 - k \frac{\Delta z'}{\Delta n} \right) = \frac{1}{2} (1 - k - k') \quad (4)$$

where $k' = \Delta z' / \Delta n$ is the average slope of the $z'$ curve within the region $\Delta n$. Both quantities of $\Delta z'$ and $\Delta n$ can be calculated by using the $z'$ curve. The region $\Delta n$ is usually chosen to be a fragment of a natural DNA sequence, e.g., a genomic island. The above method is called the windowless technique for the GC content computation (22).

**RESULTS AND DISCUSSION**

Three genomic islands in the genome of B. cereus. Some basic characteristics of the cumulative GC profile ($z'$ curve) are: 1) an up jump (a drop) in the $z'$ curve indicates a decrease (increase) of GC content; and 2) any sharp maximum (minimum) point in the $z'$ curve indicates a turning point, where the GC content undergoes an abrupt change from a relatively GC-poor (GC-rich) region to a relatively GC-rich (GC-poor) region.

The horizontal transferred elements, such as genomic islands, are usually absent in the genomes of close relatives of the host genome. By comparing the cumulative GC profiles of B. cereus and B. anthracis, it is obvious that most parts of the genomes overlap. However, there are three regions in the genome of B. cereus that have a sharp change in GC content, reflected by the fact that the $z'$ curves associated with these regions have sharp jumps. In addition, these three regions are absent in the genome of B. anthracis, suggesting a possibility that these three regions are genomic islands, which are designated the names BCGI-1, BCGI-2, and BCGI-3, respectively (Fig. 1).

BCGI-1, a 15.9-kb genomic island, has a GC content of 0.30, much lower than 0.35, the GC content of the surrounding regions. Although the length of BCGI-1 is relatively short, it has many conserved features of genomic islands. The tRNA genes have been frequently found to be the integration sites of genomic islands. Indeed, BCGI-1 utilizes a Val-tRNA gene (BC1273) as the integration site. In addition, it is also frequently found that a gene coding for an integration protein is close to the site of integration. At the 3' end of BCGI-1 is located a gene coding for a DNA integration protein (BC1272).

BCGI-2, a 62.2-kb genomic island, has a GC content of 0.38, much higher than 0.34, the GC content of the surrounding regions. At the 3' end, there is also a gene coding for site-specific recombinase (BC1921). There are totally 77 genes in this genomic island. Among these genes, 52 code for phage proteins (67.5%). There are totally 81 phage proteins in the genome. This high percentage of phage proteins also indicates that a phage-related recombination event is involved in this genomic island.

BCGI-3, a 50.3-kb genomic island, has a GC content of 0.30, much lower than 0.36, the GC content of the surrounding regions. Among the 54 genes in this genomic island, 6 are transposase genes. BCGI-3 contains an open-reading frame (ORF) (BC5092) coding for a bleomycin resistance protein, suggesting that this genomic island may play a role in its antibiotic resistance.

BCGI-3 contains a cluster of genes for a ferric anguibactin transport system. Four genes related to ferric anguibactin were found, which are ferric anguibactin transport ATP-binding protein (BC5103), ferric anguibactin transport system permease protein fatC (BC5104), ferric anguibactin transport system permease protein fatD (BC5105), and ferric anguibactin-binding protein (BC5106).
In the vertebrate host, iron is not freely available, and it is mostly found in red cells. In addition, iron in the vertebrate host is bound by the host protein transferring in blood and lactoferrin in secretions. Consequently, bacteria need to overcome the iron limitation to survive in the host and establish an infection (1). *B. cereus* is an opportunistic pathogen that causes food poisoning. Therefore, *B. cereus* should also have its own mechanism to transport the iron across the cytoplasmic membrane.

The system that transports the ferric anguibactin complex usually has an outer membrane receptor FatA, which binds the ferric anguibactin and shuttles it to the periplasm (1, 21). Among this cluster of genes, FatA gene is absent; however, indeed, there is a gene coding for ferric anguibactin-binding protein (BC5106). Although we did not detect high homology of this protein with FatA, there is still a possibility that this protein may function in the place of FatA. The ferric anguibactin transport system permease proteins FatC and FatD are inner membrane proteins that catalyze the transport of ferric anguibactin from the periplasm to the cytosol where the ferric iron is released. The ferric anguibactin transport ATP-binding protein may be involved in the energy supply in this process.

The ferric anguibactin transport system in BCGI-3 is the only ferric anguibactin transport system in the genome of *B. cereus*. No other genes, including ferric anguibactin transport ATP-binding protein and ferric anguibactin transport system permease proteins fatA, fatB, fatC, and fatD, were found in the genome. Therefore, the ferric anguibactin transport system in BCGI-3 is very likely to be involved in the iron transport for *B. cereus* that enables the bacterium to overcome the iron limitation in the host.

BCGI-3 also contains a cluster of genes related to lantibiotics. Lantibiotics are a class of bactericidal peptides that are produced by and mainly act against gram-positive bacteria. Lantibiotic peptides are characterized by the presence of thioether bridges termed lantionines, and hence the name lantibiotics (lantionines-containing antibiotics). The thioether bridges are generated by dehydration of serine and threonine followed by addition of cysteine residues. In recent years, the interest in these lantibiotics has continuously increased, mainly because of their potential to serve as natural food preservatives because of their potential to serve as natural food preservatives. We then searched the deduced protein sequences of these two ORFs against the Conserved Domain Database. We found that the domains of both the NH₂ terminus and COOH terminus of these two ORFs against the Conserved Domain Database.

The ORFs BC5083 and BC5084 encode a lantibiotic biosynthesis protein and lantionine biosynthesis protein, respectively. We then searched the deduced protein sequences of these two ORFs against the Conserved Domain Database. Indeed, the ORF BC5083 has a domain of the COOH terminus of lantibiotic dehydratase, whereas the ORF BC5084 has the domains of both the NH₂ terminus and COOH terminus of lantibiotic dehydratase. In addition, the ORF BC5086 encodes a putative lantibiotic biosynthesis protein, although no conserved domain was found. Furthermore, from the ORF BC5087 to BC5090, there are four consecutive ORFs encoding putative lantibiotic precursor peptides.

The presence of these lantibiotics in the genome of *B. cereus* poses many questions. A natural question is: what mechanisms does *B. cereus* use to protect itself from the toxicity of these bactericidal peptides? Generally, proteins conferring immunity to the producer strains antagonize specifically the lantibiotics (5). For *B. cereus*, now it is not clear which proteins have the above functions. Another question is: what advantages does *B. cereus* have by possessing these lantibiotics over other bacteria during evolution? All these questions remain to be answered.

We have also found that genes and gene orders are highly conserved between the regions around genomic island integration sites of *B. cereus* and the corresponding regions in the genome of *B. anthracis*. At the 5’ junction of BCGI-1, for instance, the ORFs of *B. cereus*, BC1254, BC1255, BC1256, and BC1257, are homologs of the ORFs of *B. anthracis*, BA1272, BA1273, BA1274, and BA1275, respectively. At the 3’ junction of BCGI-1, the ORFs BC1274, BC1275, BC1276, and BC1277 are homologs of the ORFs BA1281, BA1282, BA1284, and BA1286, respectively (Fig. 2A). The ORF BA1283 encodes a short polypeptide (34 residues) that does not have a homolog in public databases, based on the BLAST search. In addition, ZCURVE, a new system for protein-coding gene prediction, which has been shown to have low false-positive prediction rate (6), does not predict this segment as a protein-coding gene. Therefore, it is likely that the annotation of BA1283 is due to the false-positive prediction. In the GenBank file for *B. anthracis*, there is no record of the ORF BA1285. Therefore, the ORFs BA1283 and BA1285 are skipped. It is interesting to point out that the segment of DNA sequence (from ORF BA1276 to BA1280) that is between the conserved regions of the *B. anthracis* genome is absent in the genome of *B. cereus*. Therefore, it is likely that the integration of BCGI-1 causes a deletion of a segment of DNA sequence. Similar gene-loss process applies to BCGI-3, in which the ORFs BA5324–BA5331, is deleted in the genome of *B. cereus*. This segment is between the conserved regions, i.e., at the 5’ end, BC5069 and BC5070 are homologs of BA5321 and BA5322; at the 3’ end, BC5128 and BC5129 are homologs of BA5332 and BA5334, respectively. The process containing both gene gain and gene loss apparently has a more profound impact on the genome evolution than the process of gene gain only.

Likewise, the regions around BCGI-2 are also highly conserved between the two genomes. At the 5’ junction of BCGI-2, the ORFs of *B. cereus*, BC1841, BC1842, BC1843, and BC1844, are homologs of the ORFs of *B. anthracis*, BA1916, BA1917, BA1918, and BA1919, respectively. At the 3’ junction, the ORFs BC1922, BC1923, BC1924, and BC1925 are homologs of the ORFs BA1921, BA1922, BA1923, and BA1924, respectively (Fig. 2B). However, there is almost no gene loss for the integration of BCGI-2.

Comparison between the GC content distributions obtained based on windowless and window method. As a routine procedure in analyzing genome sequencing results, the distribution of GC content is displayed by the GC content within the windows that move along genomes. Although this method is intuitive, i.e., it directly shows the GC content in each particular window, a drawback is that it only displays the local GC content along genomes. On the contrary, the GC content computed without windows is a cumulative GC content; therefore, it displays a global distribution of GC content. For instance, the cumulative GC profile shown in Fig. 1 clearly shows that the genome can be roughly divided into three domains, i.e., from 1.8 to 3.5 Mb is a GC-low region; from 3.5 to 0.8 Mb is a GC-rich region; and from 0.8 to 1.8 Mb has a GC content in between. This is consistent with the result reported by the authors of the published sequence (10). By using the
windowless method, it is easily detected (compare with Fig. 3, which is based on the window method).

Another drawback of the window method is that the resolution is low. The size of window is hard to adjust, i.e., large window size leads to low resolution, whereas small window size leads to large statistical fluctuations. On the contrary, the resolution of the windowless method is high, e.g., in an extreme case, the GC content can be computed at a point (one single base), which does not have definition at all based on the window method. Therefore, by using the cumulative GC profile, the precise boundaries of the regions that have a change in GC content can be determined (Fig. 1); such boundaries, however, are hard to determine based on the window method (Fig. 3). In addition, the plots based on the window method are different when the window size is changed, but those based on the windowless method are unique. Furthermore, due to the special subtraction procedure, i.e., Eq. 3, which amplifies the variation of GC content, the cumulative GC profile has high sensitivity in detecting the changes in GC content, which is useful when the difference between the GC content of horizontally transferred elements and that of the host genome is small.

In summary, by using the cumulative GC profile to display the distribution of genomic GC content of B. cereus, based on comparison with that of B. anthracis, we have found three genomic islands in the genome of B. cereus, BCGI-1, BCGI-2 and BCGI-3, respectively. All the genomic islands have abrupt changes in GC content compared with that of surrounding regions. BCGI-1 has a typical structure of genomic islands, i.e., a Val-tRNA gene is utilized as the integration site, and a site-specific recombinase gene is located at the 3' end.
has a large percentage of phage protein, suggesting a phage-related recombination is involved. BCGI-3 contains a ferric anguibactin transport system, which is very likely to be involved in the iron transport that enables the bacterium to overcome the iron limitation in the host. In addition, BCGI-3 also contains a cluster of genes related to lantibiotics, which may play a role during the evolution of the genome. Furthermore, the integrations of the genomic islands, BCGI-1 and BCGI-3, result in deletions of DNA sequence fragments; therefore, such integrations lead to both gene gain and gene loss simultaneously.

GRANTS
The present study was supported in part by the 973 Project Grant G1999075606 of China.

REFERENCES