Bacillus anthracis Genome Annotation and Comparative Study

Li Li, Karaoz Ulas and Zheng Yu
Boston University, Bioinformatics Program
Boston, MA, 02125
zlab.bu.edu/~leely/bc777.html

Abstract
The sporulating bacterium Bacillus anthracis, found in the soil in many parts of the world, is a key spore-forming biological threat agent as well as human and animal Gram-positive bacterial pathogen. Bacillus anthracis genome consists of an approximately 5.3-MB long chromosome and two plasmids, pXO1(182 kb) and pXO2(96 kb). In this class project, we have performed computational and comparative analysis of the B.anthracis genome using various computational tools. We identified over 6000 ORFs in the B.anthracis genome. Genome comparison to Bacillus subtilis, another Gram-positive bacterium generated several interesting conclusions. identifies 2351 orthologous pairs. We identify a region in B.anthracis spanning approximately 2000 genes unique to B.anthracis and annotated mostly by “lethal toxin secretion” or “multiple-drug resistance” keywords.
Introduction

The fact that bacterial genomes typically contain 90% coding sequence makes the process of gene finding in a newly sequenced bacterial genome quite different from gene discovery in a eukaryotic genomes. In a gene-rich bacterial genome, the difficult problem of determining true genes from two or more overlapping ORFs arises. The procedure of gene finding in new genomes generally consists of building statistical models of the genomic sequence to differentiate genomic DNA from non-genomic DNA and comparing the new genome to a previously annotated bacterial genome of sufficient evolutionary distance. The latter method is by itself not sufficient, as it cannot identify genes unique to the bacteria of interest. To overcome this, one can try to identify the start of translation or to find regulatory signals and incorporate this information into gene discovery process. In this class project, we have attempted to annotate Bacillus anthracis genome using GLIMMER bacterial gene identification software, TIGRFAM, and also comparative analysis of Bacillus anthracis to Bacillus subtilis.

Materials and methods

Genome and sequencing

The whole B. anthracis genome is retrieved from TIGR (http://www.tigr.org/) as a single contig.

Glimmer System for Bacterial ORF Prediction

The Bacillus anthracis genome consists of a circular chromosome of 5314440bp. We used GLIMMER (Gene Locator and Interpolated Markov Modeler ) [1,2] to find putative genes in B.anthracis genome. Unlike previous bacterial genome annotation tools like GeneMark [3] which used Markov models, GLIMMER uses interpolated Markov models (IMM)s to find coding regions in bacterial genomes. Instead of using a fixed order Markov model as GeneMark does, GLIMMER uses an IMM-based method that outputs
predictions based on a variable context. The local composition of the sequence guides GLIMMER in constructing its variable context. Since IMMs are more flexible and more powerful than Markov chains, GLIMMER is able to find more putative genes than previous systems.

**Glimmer Parameters**

GLIMMER system has two components. One of them is the training software, build-IMM. The input to this program is a set of sequences. Build-IMM builds and outputs the IMM for them. The input sequences may well be complete genes or partial orfs. Once training is done, GLIMMER system uses second software, called glimmer, which uses the previously identified IMMs to extract putative genes in an entire genome. Whenever there is an overlap in the predicted ORFs, Glimmer resolves these conflicts by choosing one of them.

We used 3772 genomic sequences to train GLIMMER. These sequences are the long ORFs found in B.anthracis genome. Since one can have more confidence in long ORFs than in short ones, the long ORFs are used in building the model for Glimmer2. Then we ran Glimmer2 on the whole B.anthracis genome with the following parameters:

- Minimum gene length: 80
- Minimum overlap length: 0 (no overlap)
- Minimum overlap percent: 0 (no overlap)

The start codons considered were ATG, GTG, and TTG.

**Glimmer Results**

We identify 6498 ORFs in the B.anthracis genome using GLIMMER.

**Sequence Annotation and Verification**

After generating all of the B.anthracis gene sequences, we used the software in HMMER package to parse out the domain structure using TIGRFAMs[4]. TIGRFAMs are a collection of protein families featuring curated multiple sequence alignments, Hidden Markov Models (HMMs), and associated information designed to support the automated functional identification of proteins by sequence homology. We used TIGRFAMs to get
the annotations for each gene. Classification by equivalog family, where achievable, complements classification by orthologs, superfamily, domain or motif. It provides the information best suited for automatic assignment of specific functions to proteins from large scale genome sequencing projects. The TIGRFAMs includes 1622 known annotations. When using HMMpfam in HMMER, we use E-value 0.1 as a cutoff to remove spurious hits. The descriptions for all domains are retrieved from the TIGRFAMs library. An integrated annotation should attach all of description information from all found hits. The annotation table also includes the gene, the sequence length, the start and stop position of the specific gene.

**Comparison with similar organism.**
The B.anthracis is very close to the B.subtilis in the tree of life, however the latter is not lethal to humans nor to animals while the former is rather lethal to both humans and animals. B.anthracis causes anthrax. Anthrax is an acute infectious disease caused by the spore-forming bacterium Bacillus anthracis. Anthrax most commonly occurs in wild and domestic lower vertebrates (cattle, sheep, goats, camels, antelopes, and other herbivores), but it can also occur in humans when they are exposed to infected animals or tissue from infected animals. Therefore, we need to compare them and find unique genes in B.anthracis. B.subtilis has already been sequenced and has 4100 genes; we use the same method to annotate the B.subtilis by using TIGRFAMs library and compare their annotations.

**Orthologous gene pairs**
In addition to comparing the annotations for the two organisms, it is useful to find the orthologous gene pairs between the two organisms to grasp the differences in gene content of the 2 organisms.
To detect the orthologs, we use BLAST [5]. For each gene in B.anthracis, we find the best match of this gene in B.subtilis and vice versa. We detect 2043 orthologs. The dot plot of the whole B.anthracis genome against B.subtilis is given in the next section.
Discovering genes unique to B.anthracis

We hypothesize that there are 2000 unique genes in B.anthracis from the fact that B.anthracis has 2000 more genes than B.subtilis. One may think that these 2000 unique genes in B.anthracis are what makes it lethal to human and animals. We divide 6000 genes to three groups according to their chromosomal order, namely 1-2000, 2000-4000, 4000-6000. For each group we calculate the number of orthologs shared by both organisms. Using the gene annotations, we also record the number of toxin related genes, multiple drug resistance genes and efflux genes.

Results and Discussion

Annotation of B. anthracis

The full annotation table can be found at zlab.bu.edu/~leely/unique2000_ans.table. A snapshot of the annotation table is as follows:

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Start</th>
<th>Stop</th>
<th>Length</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query 2201</td>
<td>1820209</td>
<td>1820907</td>
<td>232</td>
<td>3a0106s01: sulfate transport system permease protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cbiO: cobalt transport protein ATP-binding subunit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3a0501s02: Type II (General) Secretory Pathway (IISP)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Family protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>thiQ: ABC transporter, ATP-binding protein, ThiQ subfamily</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>drrA: daunorubicin resistance ABC transporter ATP-binding subunit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ntrCD: nitrate transport ATP-binding subunits C and D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ccmA: heme exporter protein CcmA;3a0107s01c2: phosphate transport system permease protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>potA: spermidine/putrescine ABC transporter ATP-binding subunit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nodI: nodulation ABC transporter NodI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>proV: glycine betaine/L-proline transport ATP binding subunit</td>
</tr>
<tr>
<td>Query 2202</td>
<td>1820923</td>
<td>1821789</td>
<td>288</td>
<td>No hits above threshold</td>
</tr>
<tr>
<td>Query 2204</td>
<td>1821767</td>
<td>1822792</td>
<td>341</td>
<td>No hits above threshold</td>
</tr>
<tr>
<td>Query 2205</td>
<td>1823645</td>
<td>1822860</td>
<td>261</td>
<td>HTH_fis: Helix-turn-helix domain, fis-type</td>
</tr>
</tbody>
</table>
2456 out of 6486 predicted genes in B.anthracis match to a TIGRFAM. The annotation rate is 38%. The low annotation rate is partly due to the fact that TIGRFAM has only 1622 known annotations. In running the TIGRFAM software, we used an E-value threshold of 0.1 instead of the default E-value of 100 so there are a number of hits lost due to their low fidelity.

Annotation of B.subtilis

The full annotation table for B.subtilis can be obtained from TIGR. A snapshot of the annotation table for B.subtilis using a similar procedure is as follows:

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>2632312</td>
</tr>
<tr>
<td>gi</td>
<td>2632313</td>
</tr>
<tr>
<td></td>
<td>thrB: homoserine kinase</td>
</tr>
<tr>
<td></td>
<td>mevDPdecarb: diphosphomevalonate decarboxylase</td>
</tr>
<tr>
<td></td>
<td>mevalon_kin: mevalonate kinase</td>
</tr>
</tbody>
</table>

1670 out of 4100 genes in B.subtilis match to a TIGRFAM. The annotation rate is 41%.
Comparison of B.anthracis and B.subtilis genomes

To compare the two genomes, in addition to annotation comparison a dot-plot analysis is performed using the whole genome BLAST results.

The immerging diagonal line in the dot-plot depicts a conserved region corresponding to genes from 4800 to 6400 in B.anthracis and to genes from 2300 to 3800 in B.subtilis. There are two other shorter conserved regions, one located at the region from 1 to 200 in B.anthracis and from 1 to 100 in B.subtilis, and the other from 1200 to 1500 in B.anthracis and from 1000 to 1300 in B.subtilis. The two relatively short reverse diagonals points to similar but reverted regions in both organisms.

There is a fairly large amount of genes unique to B.anthracis as it may be seen from the region corresponding to genes from 1800 to 4000 in B.anthracis and to genes from 1800 to 2300 in B.subtilis. To further investigate these unique genes in B.anthracis, we perform the following analysis:
1. Histogram of genes related to lethality factors

We count the number of annotations that match “lethal toxin secretion” or “multiple-drug resistance” keywords. A histogram of these counts is as follows:

Figure 1. Histogram of keywords

The supposedly unique genes in the region from position 2000 to 4000 have a higher count of the “multiple-drug resistance factor”, “efflux protein”, and “bacteria toxin” keywords.
2. Orthologous pair percentage vs. Gene position
We are looking at the percentage of orthologous genes in different regions along the B.anthracis genome. The percent of orthologous genes for 6 different regions is as follows:

![Bar chart showing percent orthologous pairs along the B.anthracis genome](chart.png)

**Figure 2. percent orthologous pairs along the B.anthracis genome**

As expected, the percent of orthologous genes goes down from 35-45% to 20% in the region where we expected the unique genes to B.anthracis are located. These genes may be further investigated in a future study looking at their annotation more carefully. Further improvements may include changing the initial set of sequences used to train the IMM model in GLIMMER. One can expect to get better annotation by using a curated training set composed of B.subtilis genes. There are a lot of further possible study that can be done specifically on the continuous region unique to B.anthracis and also on the whole B.anthracis genome. We can look for regions of atypical nucleotide composition using $X^2$ analysis and compare it to the composition for B.subtilis. In addition, we can search for homologs in other bacterial species.

3. G + C content rate analysis.
General features of the DNA sequence

Analysis at the replicon level. The B. anthracis chromosome has 6,969,760 base pairs (bp), with the origin of replication coinciding with the base numbering star point. The average G + C ratio is 26.5%, but it varies considerably throughout the chromosome. The B. subtilis has 4,214,810 base pairs (bp), and the average G + C ratio is 43.5%. (6) The average G + C content of 2000-4000 genes in B. anthracis is lower than 1-2000 genes and 4000-6000 genes. The base pairs of 2000-4000 of B. anthracis are from 1,653,510 to 3,251,408, which are unique. The G + C ratio in region between 1,653,510 to 3,251,408 is lower than the other two parts from this figure. This evidence indicates that the 2,000 unique genes which B. anthracis has are quite different in GC content with the other 4,000 genes. It is probably a result of large chromosomal transfer from another organism.

ACKNOWLEDGMENTS

We are grateful to Yu Zheng to help us finish this project. We thank Simon Kasif for his brilliant ideas of this project.
REFERENCES