

Analyzing Pathogen DNA Sequences

Thomas S. Brettin

DNA-based experimental techniques allow rapid detection and identification of pathogens for medical treatment, criminal forensics, and possibly attribution (that is, finding the source of an outbreak). Bioinformatics provides the computational tool for that identification process. Loosely defined as the merger of computers and biology, bioinformatics evolved significantly during the Human Genome Project in response to the biologists' need to assemble a complete genomic sequence from DNA fragments. In the area of bioterror reduction, we are primarily concerned with analyzing the information contained in a pathogen's genome. Bioterror reduction, therefore, requires a collection of computational tools and databases that are different from those used to determine the genomic sequence.

The initial step in creating an identification tool for reducing the biological threat is to identify the complete set of genes contained in the pathogen's genome. This task is generally accomplished by use of hidden Markov-based techniques, which enable us to calculate the probability of finding the next base in a DNA sequence, given the bases directly preceding it. It turns out that in regions of DNA that encode genes, the next nucleotide in a sequence can be predicted with high probability based on the previous four to seven nucleotides. In genomic sequences that do not code for genes, this predictability is significantly less. DNA sequences that have high predictability are therefore identified as gene sequences. For organisms such as the common bacterium *Escherichia coli*, this approach has shown better than 95 percent accuracy. We routinely apply this

approach to pathogenic bacteria that represent a threat to our safety, such as *Bacillus anthracis* and *Yersinia pestis*, the causative agents of anthrax and plague, respectively.

With the gene predictions in hand, we are in a position to interpret the functions of these genes in the cell, as well as identify genes that are unique to a pathogen. At its simplest, this procedure is known as functional annotation. It is a difficult process, in part because genes from two organisms that perform the same function rarely have the same DNA sequence. We use various statistical techniques to assess the degree to which a gene sequence is similar to anything in a database of known gene sequences. Even in a single-cell organism like *B. anthracis*,

which contains between 5000 and 6000 genes, about 40 percent have no statistically significant similarity to anything in the database. Recent studies have shown, however, that the functions of the protein products of two genes can be remarkably similar, despite any discernible sequence similarity between the genes. Thus, our techniques look not only for overall similarities to specific genes but also for similarities to conserved functional domains within the genes.

Current technology allows us to complete 95 percent of the DNA sequence of a bacterium of five million bases in a matter of days. Because decreasing that time even further is expected, studies that compare the genomes of closely related

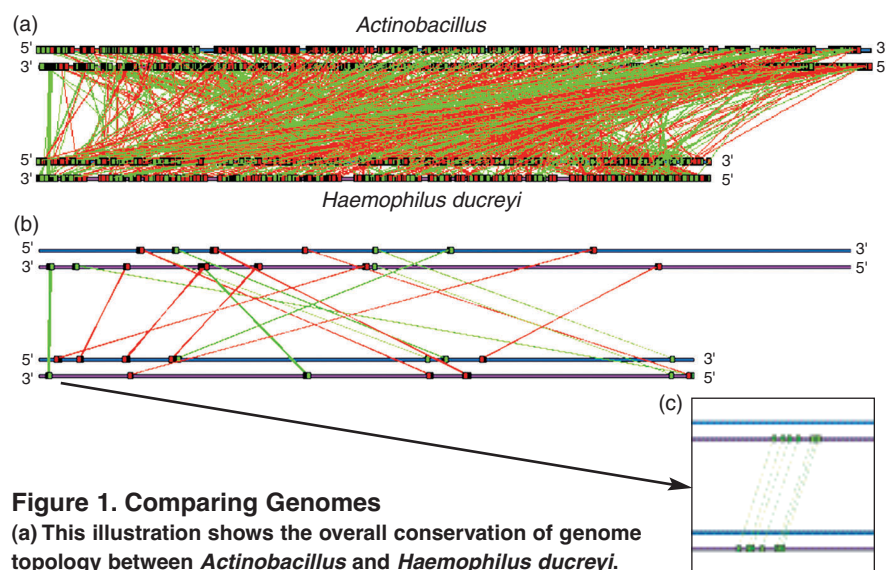


Figure 1. Comparing Genomes

(a) This illustration shows the overall conservation of genome topology between *Actinobacillus* and *Haemophilus ducreyi*.

Both bacteria have only one chromosome, which is represented by the two heavy lines (one for the forward strand and one for the reverse). Genes appear on both strands. The green lines connect similar genes that appear on the same strand in both genomes, whereas red lines connect genes appearing on opposite strands. Most genes are present in both genomes, yet overall, the gene order, or the location of the gene, is not conserved. (b) There are, however, contiguous stretches of six or more genes that are conserved in gene order. (c) The detail shows a stretch of six genes that are involved in cell envelope biosynthesis and antibiotic susceptibility. These genes are conserved in both gene order and function and therefore point to an important evolutionary link between the two bacteria.

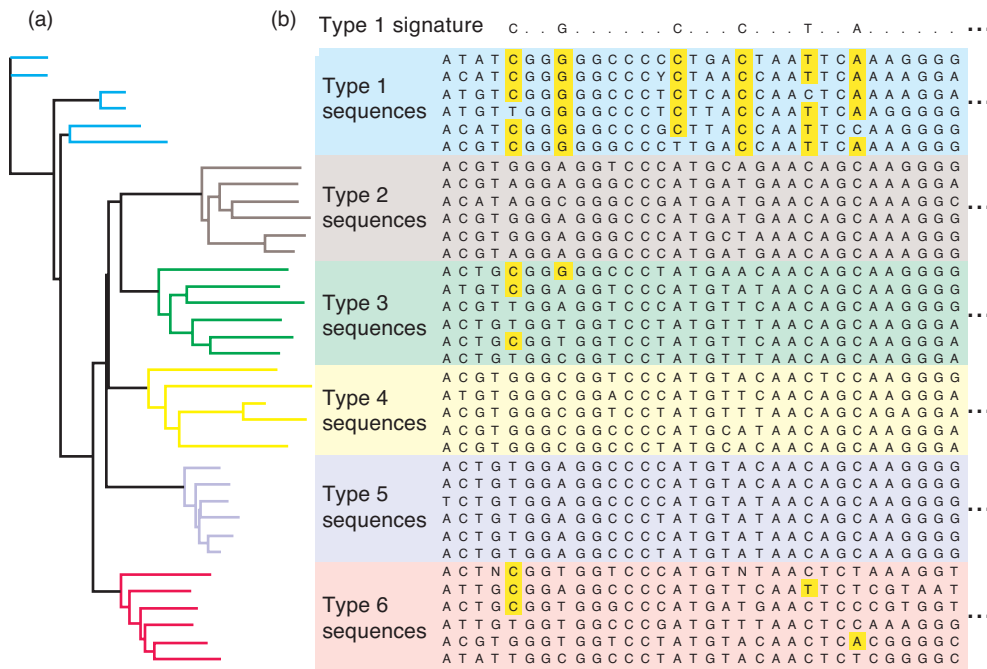


Figure 2. DNA Signature
 (a) This phylogenetic tree shows the evolutionary relationships of a subset of hepatitis C viruses. (b) Given the tree, we can align partial DNA sequences and identify individual bases that can be used to distinguish a particular clade. The signature for Type 1 hepatitis C viruses is shown below the Type 1 sequences. It can be used to identify such viruses, or when used in conjunction with other signatures, to help characterize unknown viral strains.

bacteria will become more common. Such studies may help us understand the remarkable, possible physiological differences between closely related organisms. For example, although the genomes of *B. anthracis* and its near neighbor *B. thuringiensis* are greater than 99 percent similar, the former can kill humans, whereas the latter is harmless to humans and is widely used as a pesticide. Our first examination of these two genomes has revealed small sections of unique DNA, scattered throughout each genome, that range from small mobile DNA elements (insertion elements, transposons, and phages) to regions of around 25 genes of unknown function. The unknown genes presumably relate to existing physiological differences.

Other comparisons such as those between *Actinobacillus* and *Haemophilus ducreyi* show small regions of conserved gene order in an otherwise nonconserved genome topology (see Figure 1). In addition to their potential evolutionary significance, those conserved regions may also serve as gene targets for detection assays and disease mitigation strategies.

The article “Reducing the Biological Threat” on page 168 describes how variable number tandem repeats and single nucleotide polymorphisms can be used as “signatures” to differentiate among different strains of a single organism. Given the genomic sequence, we can easily locate these simple repeats and supply laboratory personnel with the necessary information to design and test an assay. But we are also developing techniques to construct new signature sets that will identify pathogen strains.

The DNA sequences from thousands of related pathogens are first used to construct a phylogenetic tree, as seen in Figure 2(a). (Note that the tree seen in the figure is “pruned” to show only a small subset of branches. Each branch end therefore corresponds to a specific DNA sequence that is representative of many similar sequences.) Overall, the branches cluster into groupings, or clades. We examine the DNA sequences that make up each clade and identify a set of individual bases that is highly likely to be seen in the chosen, but not in other sequences—see Figure 2(b).

This set of bases—the signature—provides a statistically powerful means to discriminate among clades and allows us to place new strains/DNA sequences into their genetic context.

Genome sequencing has opened a new era in the study of pathogens. Soon, hundreds of genomes from closely related pathogens will be available, and understanding what makes pathogens both similar and different will start at the DNA level. The looming challenge will be to develop methods for rapid detection and identification of pathogens, as well as new treatments. Making good use of genomic sequence data is a significant step forward in our goal of meeting that new challenge. ■

Tom Brettin received a bachelor’s degree in biochemistry and a master’s degree in plant genetics from Michigan State University. He is a technical staff member in the Bioscience Division at Los Alamos, where he works in the field of bioinformatics. Tom has authored a number of genetic databases.