Computer Note

GenScalpel: An Application for Sequence Retrieval and Extraction from the GenBank flatfile
Yong-Hua Yin*, Lian-Ming Du*, and Bi-Song Yue

From the Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, College of Life Science, Sichuan University, Chengdu, PR.China.

Address correspondence to Bi-Song Yue at the address above, or e-mail: bsyue@scu.edu.cn

*These authors contributed equally to this work.

Abstract
GenScalpel is a program designed for the retrieval and extraction of specified sequences from large-scale sequence sets in NCBI GenBank flatfile format. This routine task in bioinformatics analysis is a pressing need for laboratory biologists. Another objective of application development is to respond to the new form of the NCBI Nucleotide Sequence Database, which was updated in November 2011. In addition to a powerful sequence refinement application, GenScalpel provides convenient functions for web-based sequence downloading or multiple files batch processing. This note discusses major applications of the program and includes example data sets to demonstrate its performance. The program is written in PERL. GenScalpel, including installation packages for Windows and Linux systems as well as the accompanying documentation, are available free of charge at http://genscalpel.biosv.com/.

Key words: format parsing, GenBank flatfile, GenScalpel application, sequence retrieval and extraction

The GenBank flatfile (GBF) format is the standard unit of the GenBank database, and the most popular sequence file format in the world. As part of the International Nucleotide Sequence Database Collaboration with the EMBL and DDBJ, GenBank and its collaborators update sequences submitted from more than 100 000 distinct organisms every day (McEntyre and Ostell 2002). From this process, the problem of how to parse and utilize the burgeoning molecular sequence data (in GBF format mostly) has arisen for laboratory scientists. Huge endeavors were made to reach these goals in the past decade, such as Bioperl (Stajich et al. 2002) and EMBOSS (Rice et al. 2000). An example is GBParsy, a powerful and highly functional library developed by Lee et al. (2008) to parse and extract annotated information from GBF files. But all software packages failed to effectively accomplish gene sequence retrieval and extraction, which is an urgent need for biologists.

For example, the rates of synonymous (Ks) and nonsynonymous (Kn) substitution are commonly used to detect natural selection in genes and genomes. As we know, natural selection mainly acts on the protein level (Kimura 1977; Yang 2007). However, the genome sequence data in GBF format include all the features, such as genes, transfer RNAs, contigs or others (e.g., noncoding regions in mtDNA), which means researchers need to extract protein-coding genes from the whole genome scale. Generally, the task performed by manual cutting and pasting, and reformattting is tedious, time-consuming, multistep, and error prone. The GenScalpel program aims to solve problems mentioned above and develops a fully function for sequence retrieval and gene extraction from large datasets. This is done by using a novel approach to format parsing.

NCBI upgraded the Entrez Genome database in a more secure and sophisticated format in November 2011. This makes most previous software fail to retrieve data dynamically from a remote database via the Internet (Stajich et al. 2002; Kumar and Dudley 2007). Facing the new interface standard of GenBank, GenScalpel provides a user-friendly tool for rapid internet connectivity, sequence data accession, and web-based downloading.

Batch-processing applications allow the user to process multiple files at one time, rather than processing them one by one in a single-file format. All these two applications of GenScalpel are new features in the GBF parsing software.

Program Applications and Development Methods

The GenScalpel functions illustrated in Figure 1 are organized by implementation modules. These applications allow biologists to use integrated methods to access a best-suited dataset. To develop its functions, two key strategies were used in designing GenScalpel project.

First, accessing and downloading GenBank sequence records are performed by E-entity calls, which is a universal interface provided by the NCBI Entrez query and database system (Sayers 2010). Through converting search terms from user input to an Entrez uniform resource locators (e.g., Gene_id), the GenScalpel program constructs an internet-based pipeline to access the Entrez database and retrieve sequences.
When manipulating sequence data, we followed certain biological principles. For example, GenScalpel supports the merge operation against a set of sequence data (e.g., all protein-coding genes in a genome). But in some viral and mitochondrial genomes, there are a series of overlaps between the genes (Lewin 2004). In such a situation, the same sequence of DNA is shared by two adjacent genes. Considering the different open reading frames of the genes, the overlapping sequences are copied and split by GenScalpel and then assigned into each of the two genes. This application is effectively kept the completeness of the triplet sets of genes.

Performance Test

To demonstrate the functioning of GenScalpel, four types of genome sequences were processed for accuracy test (Table 1). Raw NCBI reference sequences were downloaded in TXT format (www.ncbi.nlm.nih.gov/genome/).

In this test, different types of gene sets (gene, protein coding genes, ribosomal RNA and transfer RNA) were extracted from their complete genomes by GenScalpel, respectively. Meanwhile, the operations with manual cutting-and-pasting were performed to provide a comparable result. All the output sequences were aligned by MEGA sequence alignment editor (Tamura et al. 2007).

As the results show, the gene sequences output from both GenScalpel and manual handling shared identical base compositions and lengths (bp, 4839; 10734; 2115; 3005). It demonstrates gene sequences can be accurately extracted by using the GenScalpel program.
Input and Output Files

Four input data file formats can be read and automatically recognized by GenScalpel: (1) Accession number of the NCBI Reference Sequence; (2) the GenBank sequence format (.gb); (3) standard TXT format (easily produced from a WINDOWS Notepad file); and (4) a folder that contains multiple TXT files. Nucleotide sequences in IUPAC single-letter codes will be written in TXT or Fasta format, which allows the direct reading by many sequence-based programs, such as Primer Premier (Lalitha 2000) and MEGA.

The instruction manual accompanying the GenScalpel application provides a brief tutorial for easy learning and usage, details on how to produce input data files in standard TXT format, and tips on potential pitfalls. Several examples that explicitly explain the performance of the program were provided with the documentation as well.

Availability

GenScalpel is a program written in PERL 5.12.4 with a stable graphical user interface (GUI). It provides the user with a friendly window and menus environment for sequence assembly and analysis. Executables are compiled for Linux and Windows respectively. The installation packages, documentation and the source code are distributed free of charge at its web site: http://genscalpel.biosv.com/

Funding

National Natural Science Foundation of China (31172118).

References


Table 1: Accuracy tests applied to four types of genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Organism</th>
<th>Length (bp)</th>
<th>Feature</th>
<th>Accession</th>
<th>Data set reference</th>
<th>Fragments</th>
<th>Sequence output (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hepatitis B virus (HBV) subtype ayr</td>
<td>Drosophila yakuba</td>
<td>3,215</td>
<td>the S, C, P and X genes</td>
<td>NC_003977</td>
<td>Okamoto et al. 1986</td>
<td>4</td>
<td>4,839</td>
</tr>
<tr>
<td>HIV-1</td>
<td>Drosophila yakuba</td>
<td>9,181</td>
<td>protein coding genes</td>
<td>NC_001802</td>
<td>Fu et al. 2009</td>
<td>9</td>
<td>10,734</td>
</tr>
<tr>
<td>mitochondrial DNA</td>
<td>Drosophila yakuba</td>
<td>16,019</td>
<td>rRNA</td>
<td>NC_001032</td>
<td>Clary and Wolstenholme 1985</td>
<td>2</td>
<td>2,115</td>
</tr>
<tr>
<td>chloroplast DNA</td>
<td>Arabidopsis thaliana</td>
<td>154,478</td>
<td>transfer RNA</td>
<td>NC_000932</td>
<td>Sato et al. 1999</td>
<td>37</td>
<td>3,005</td>
</tr>
</tbody>
</table>

Feature: the region of DNA that annotated with a key/type in GBF. Fragments, the number of genes involved in present extracting; Sequence output, the length of extracted sequences using GenScalpel; Accession, the accession number of the NCBI Reference Sequence; the S, C, P, and X genes, the HBV genes that code for the large S protein, the precore/core protein, the polymerase, and the X protein respectively.

910


Received April 16, 2012; Revised June 1, 2012; Accepted June 4, 2012

Corresponding Editor: Adalgisa Caccione