GenScalpel

An Application for Sequence Retrieval and Extraction from the GenBank Flatfile

VERSION 1

Yong-Hua Yin[1], Lian-Ming Du[2], Bi-Song Yue[3]

College of Life Science, Sichuan University

Key Laboratory of Bio-resources and Eco-environment,

Ministry of Education

610064 Chengdu City, P.R.China

[1]yhyin.scu@gmail.com

[2]lmdu.scu@gmail.com

[3]bsyue@scu.edu.cn

Table of Contents

1 PREFACE

1.1 Preface

GenScalpel is a program designed for specified-sequence retrieval and extraction 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 last November. In addition to a powerful sequence refinement application, **GenScalpel** provides convenient functions for web-based sequence downloading and multiple files batch-processing. The program is 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.

1.2 Author contributions

Y.-H.Y. and B.-S.Y. designed the research; L.-M.D. and Y.-H.Y. developed the application.
Y.-H.Y. and L.-M.D. contributed equally to this work.

1.3 License

2 PART I: GETTING STARTED

2.1 System Requirements

GenScalpel is written in Perl. By making installation packages, executables are compiled for Linux and Windows, respectively.

We recommend a computer with at least 128 MB of RAM and 20 MB of available hard disk space. The program has been tested for cross-platform compatibility on operating systems mentioned above.

In order to process very large datasets, a faster processor (x86 or later) and larger amount of physical memory will be needed.

The installation packages, documentation and the source code are distributed free of charge at its web site: http://genscalpel.biosv.com/.

2.2 Installing GenScalpel

2.2.1 Installation on Windows

Step 1: The preferred way to install **GenScalpel** is to directly download it as a single compressed ZIP file. Then you must use a program, such as WinZip, to uncompress this ZIP file into a specified directory.

Step 2: Once downloaded and expanded on your hard drive, simply double-click on the Installer (GenScalpelv1.0-installer.exe), and the appropriate files will be installed on your computer.

***Or* Step:** If you are unable to install **GenScalpel** directly from the website, you can simply copy GenScalpel-related files from one computer to another.

2.2.2 Installation on Linux

You'll probably want to operate from inside your home directory. If your user is (for example) *username*, your home directory will be */home/username/*. For the rest of this section we will assume you have downloaded your zip file to */home/username/gs*. If you do not have a *gs* directory, you can create it with the following "mkdir" (make directory) command:

Code:

```
mkdir /home/username/gs/
```

Step 1:

Download the package “GenScalpel-1.0.tar.gz” from http://genscalpel.biosv.com/.

Step 2: Change to the /home/username/gs/ directory with the "cd" command like so:

Code:

```
cd /home/username/gs/
```

Step 3: We now need to unzip the zipped file:

Code:

```
tar -xvfz GenScalpel-1.0.tar.gz
```

Step 4: We now need to go into the new directory, so use the cd command:

Code:

```
cd GenScalpel-1.0
```

Step 5: Make sure the file is set to "executable" by running this command:

Code:

```
chmod +x GenScalpel
chmod +x batch
```

Step 6: Run the file like this:

Code:

```
./GenScalpel
```

***Or* Step 6:** Double click the “GenScalpel” directly to run the program.

2.3 A Walk Through GenScalpel

This section provides a **GenScalpel** tutorial. The data files for these examples are available online, or can be found in the **EXAMPLES** folder, located in the **GenScalpel** installation directory (example in **C:\Program Files\GenScalpel\Examples**).

The data files we performed are complete mitochondrial genome sequences from five insect species (How to create a standard TXT format file, see ***3.2 Single Standard TXT Format***). The parameter list of them is given below.

File (.TXT)	Species	Length (bp)	Web address (URL)
NC_002084	*Anopheles gambiae*	15,363	http://www.ncbi.nlm.nih.gov/nuccore/NC_002084
NC_001566	*Apis mellifera*	16,343	http://www.ncbi.nlm.nih.gov/nuccore/NC_001566
NC_001709	*Drosophila melanogaster*	19,517	http://www.ncbi.nlm.nih.gov/nuccore/NC_001709
NC_001322	*Drosophila yakuba*	16,019	http://www.ncbi.nlm.nih.gov/nuccore/NC_001322
NC_003081	*Tribolium castaneum*	15,881	http://www.ncbi.nlm.nih.gov/nuccore/NC_003081

In these example files, data are deliberately processed in different input formats. We recommend that you study the examples in the order presented because the techniques explained in the initial examples are used again in the subsequent ones.

2.3.1 Internet-based parsing

In this example, we will explain how to localize and parse sequence data based on internet. The protein-coding gene sequences, contained in the mitochondrial genome of *Anopheles gambiae*, will be retrieved from a remote database via the Internet.

Step 1: Start **GenScalpel** by Double-clicking on the **GenScalpel.exe** icon;

Step 2: Select the **F**ile|**G**et **N**CBI **D**ata menu command to open **G**et **N**CBI **D**ata dialog box;

Step 3: Type an Accession number in the text box: **NC_002084**, then click **D**ownload;

Note: In case of a spelling mistake the **C**lear button can be used to re-enter a search term.

Step 4: A new GBF file will be downloaded from GenBank and be read by **GenScalpel** program.

Step 5: Choose the feature '**gene**' in the top right corner of the interface and then click '**G**et **S**equence'.

Step 6: Click the **S**ave **S**equence button in the **S**ave **S**equence dialog box, and specify an output folder: **C:\Program Files\GenScalpel\Examples\Output**, and press **S**ave.

Note: If you want to merge all the gene sequences in NC_002084, select the **M**erge **S**equence check box.

Step 7: A new file **NC_002084_gene.txt** will be written in the specified output folder (C:\Program Files\GenScalpel\Examples\Output\).

2.3.2 Single .TXT File Processing

In this example we will extract the gene sequence of *Cytochrome b* from the mitochondrial genome of *Apis mellifera*. In the GBF format the feature for this gene is **CYTB**.

Step 1: Start **GenScalpel** by Double-clicking on the **GenScalpel.exe** icon;

Step 2: Click the **F**ile|**O**pen menu command to select file **C:\Program Files\GenScalpel\Examples\ NC_001566.txt**;

Step 3: Choose the feature '**CYTB 11004-12155**' in the top right corner of the interface and then click '**G**et **S**equence'.

Step 4: Click the **S**ave button in the **G**et **S**equence dialog box, and specify an output folder: **C:\Program Files\GenScalpel\Examples\Output**, and press **S**ave.

Step 5: A new file **NC_001566_CYTB11004-12155.txt** will be written in the specified output folder (C:\ Program Files\GenScalpel\Examples\Output\).

2.3.3 Multiple .TXT Files Batch-processing

In this example five .TXT files (NC_002084.txt, NC_001566.txt, NC_001709.txt, NC_001322.txt and NC_003081.txt) were stored in the same folder (C:\Program Files\GenScalpel\Examples). Here we extract respective protein-coding amino acid sequences from the data files.

Step 1: Start **GenScalpel** by Double-clicking on the **GenScalpel.exe** icon;

Step 2: Select the **T**ool|**S**witch to **B**atch menu command to open **B**atch **P**rocessing dialog box;

Step 3: Click the **A**dd **F**older button , and then select the folder **C:\Program Files\GenScalpel\ Examples**;

***Or* Step 3:** Click **A**dd **F**ile to add five files in folder: **C:\Program Files\GenScalpel\Examples** one by one;

Note: In case of a loading mistake the **D**elete button and the **C**lear button can be used to delete or clear items in **F**ile(s) **L**ist.

Step 4: Specify the output folder: **C:\Program Files\GenScalpel\Examples\Output** by clicking the **B**rowser button.

Step 5: Click the **G**et **P**rotein **S**equence button in the **F**unction frame to complete the task.

Step 6: Five new files **NC_002084_protein_seq.txt**, **NC_001566_protein_seq.txt**, **NC_001709_ protein_seq.txt**, **NC_001322_protein_seq.txt** and **NC_003081_protein_seq.txt** will be written in the output folder (C:\Program Files\GenScalpel\Examples\Output).

3 PART II: INPUT DATA TYPES AND FILE FORMAT

3.1 GenScalpel Format

Four input data file formats can be read and automatically recognized by **GenScalpel**:

(i) Accession Number of the NCBI Reference Sequence;

(ii) the GenBank sequence format (.gb);

(iii) standard **TXT** format (easily produced from a WINDOWS Notepad file)

and **(iv)** 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** (Tamura *et al.*, 2007).

3.2 Single Standard TXT Format

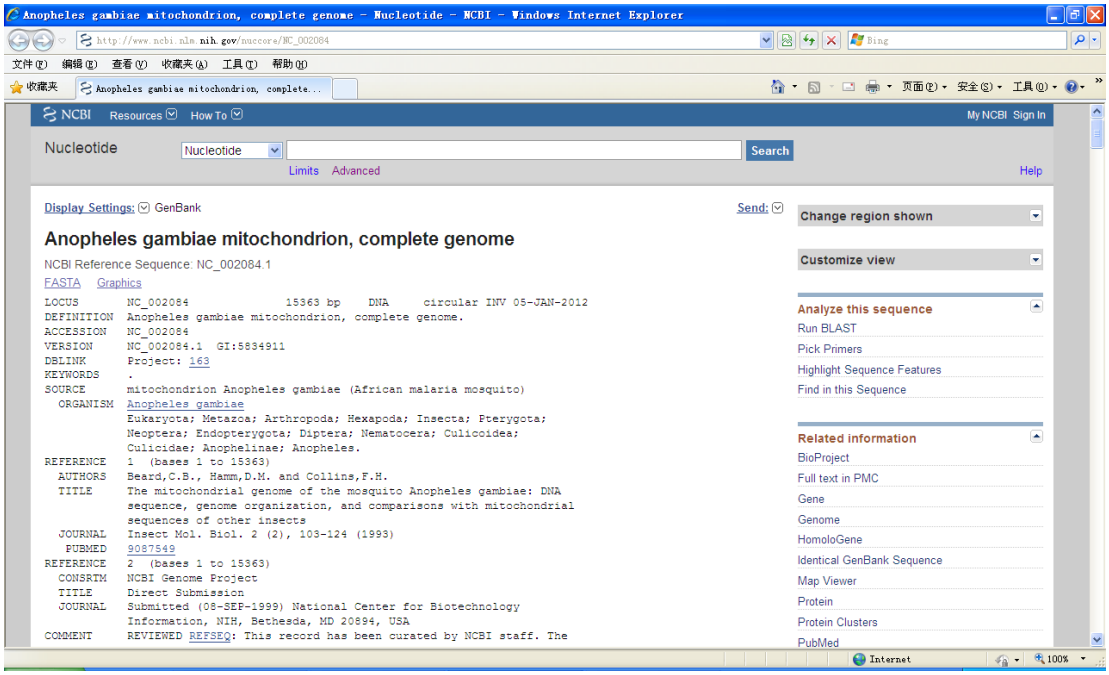
This section provides a guide for creating a text input file from a WINDOWS Notepad file.

Step 1: Create a new .TXT file.

Right click the blank area of the **D**esktop, **N**ew > **T**ext Document, a blank document will be created. Double click the file you just created to open it with Notepad.

Step 2: Open a Webbrowser like Mozilla **F**irefox or WINDOWS **I**nternet Explorer, then visit the homepage of the NCBI (http://www.ncbi.nlm.nih.gov/). Search for the sequence you want to process. Change the format to GenBank Flatfile format.

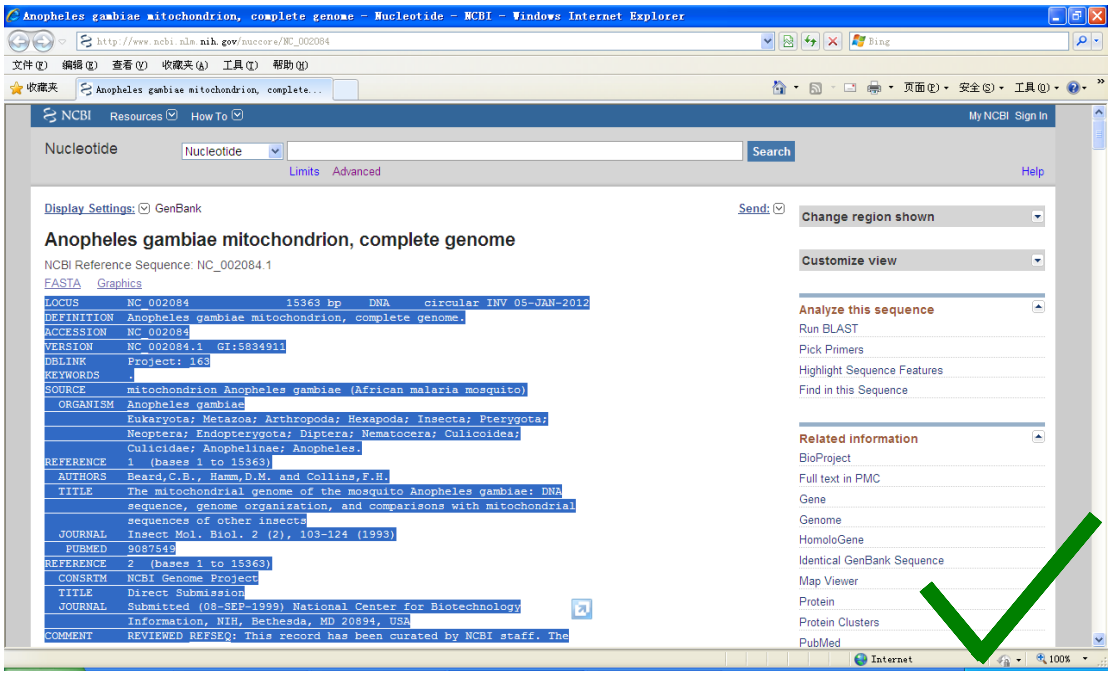
***or* Step 2:** If you already know the URL address, all you need to do is open a Webbrowser and paste the link into the address bar. (e.g. url: **http://www.ncbi.nlm.nih.gov/nuccore/NC_002084**)

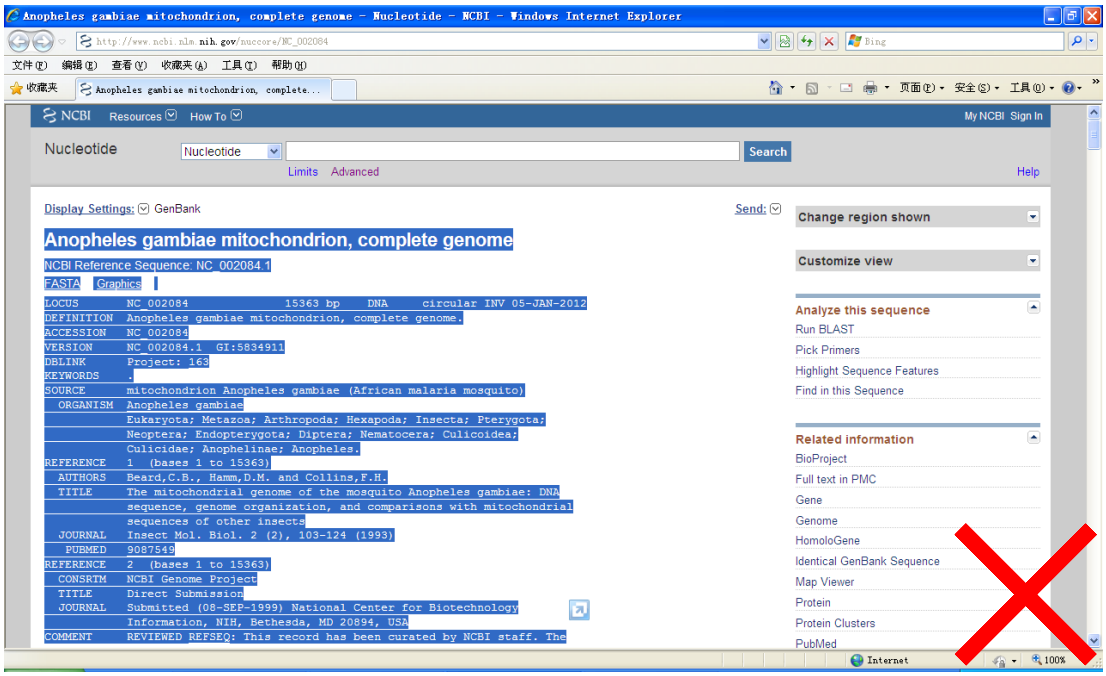


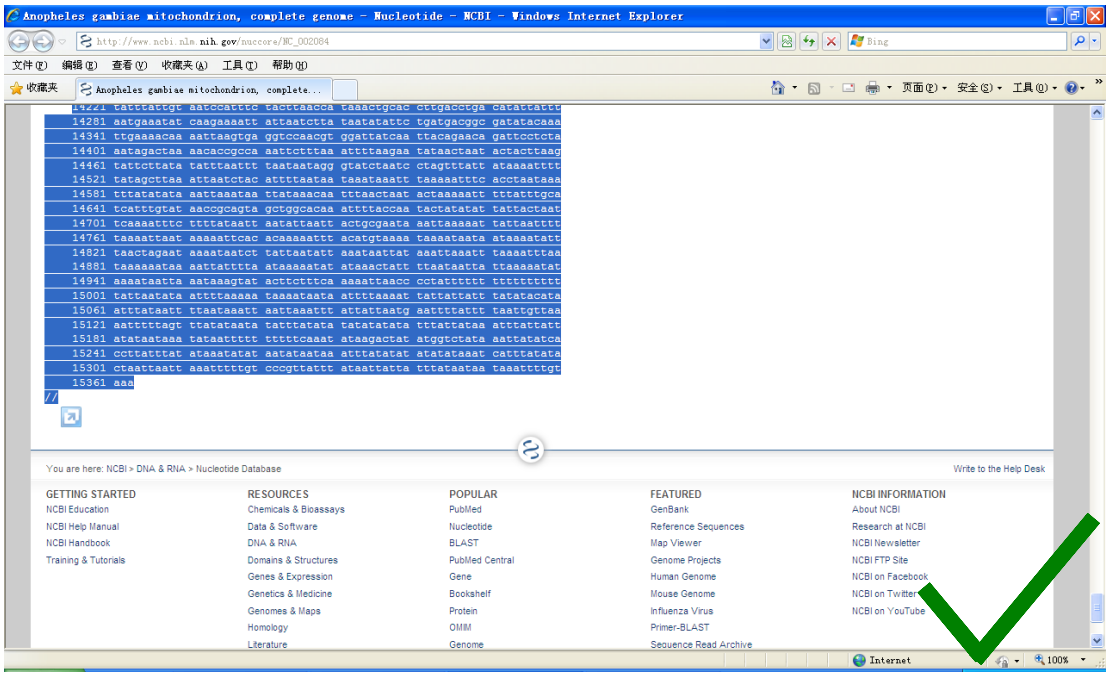
Step 3: Copy the information from a Webpage into a document.

Select the information starting at the word **LOCUS**, and click **C**opy in the **E**dit menu. Open the Texteditor and click **P**aste in the **E**dit menu.

Note: It is strongly recommended that you avoid extracting the whole page of records presented in the webpage. As a result of a wrong GenBank flatfile format, **GenScalpel** can fail to read.







Step 4: Documentation save.

Click **S**ave As in the **F**ile menu of Texteditor. In the pop-up dialog box (the **S**ave File dialog), you can select a different location or specify a different **N**ame, and then click **S**ave. Once this has been finished, click the **C**lose button in the top right corner.

Note: A context-sensitive file name is advised. (e.g. **NC_002084**.txt in this example)

3.3 Importing Data From A Folder

In this section we briefly describe how **GenScalpel** handles multiply files in batch processing.

Step 1: Create a new folder.

Right click the blank area of the **D**esktop, **N**ew > **F**older, a blank folder will be created. Double click the folder icon to open it.

Step 2: **S**elect all the .TXT files that need to be batch processed, then right click the mouse and select **C**opy from the pop-up menu.

Step 3: Copy the .TXT files to a new folder.

Open the folder just created and click **P**aste in the **E**dit menu. This completes the preparation for importing data from a folder.

4 APPENDIX

References

Beard,C.B., Hamm,D.M. and Collins,F.H. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol. Biol.*, **2**, 103–124.

Clary,D.O. and Wolstenholme,D.R. (1985) The mitochondrial DNA molecular of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.*, **22**, 252–271.

Crozier,R.H. and Crozier,Y.C. (1993) The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics*, **133**, 97–117.

Friedrich,M. and Muqim,N. (2003) Sequence and phylogenetic analysis of the complete mitochondrial genome of the flour beetle *Tribolium castanaeum*. *Mol. Phylogenet. Evol.*, **26**, 502–512.

Kimura,M. (1977) Prepondence of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, **267**, 275–276.

Kumar,S. and Dudley,J. (2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics*, **23**, 1713–1717.

Lalitha,S. (2000) Primer Premier 5. *Biotech Software & Internet Report*, **1**, 270–272.

Lee,T.H., Kim,Y.K. and Nahm,B.H. (2008) GBParsy: A GenBank flatfile parser library with high speed. *BMC Bioinformatics*, **9**, 321.

Lewin,B. (2004) Gene. Pearson Prentice Hall, Upper Saddle River.

Lewis,D.L., Farr,C.L. and Kaguni,L.S. (1995) *Drosophila melanogaster* mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. *Insect Mol. Biol.*, **4**, 263–278.

McEntyre,J. and Ostell,J. (2002) *The NCBI Handbook*. National Center for Biotechnology Information (US), Bethesda.

Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276–277.

Sayers,E. (2010). Entrez programming utilities help. National Center for Biotechnology Information (US), Bethesda.

Schwartz,R.L., Phoenix,T., and Foy,B.D. (2009). Learning perl. O'Reilly Media, Sebastopol

Stajich,J.E. *et al.* (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res.*, **12**, 1611–1618.

Tamura,K. *et al.* (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.

Wishart,D.S. *et al.* (2000) PepTool and GeneTool: platform-independent tools for biological sequence analysis. *Meth. Mol. Biol.*, **132**, 93–113.

Yang,Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.