

Learning GenSeed with the tutorials

1. Introduction

We provide a series of tutorials for the users to test GenSeed and learn how to use it for different applications. All tutorials use real-life data and, as such, the final results can be evaluated by comparison with public databases.

To install the tutorials, please download the file `tutorial_data.tgz` (http://www.coccidia.icb.usp.br/genseed/tutorials/tutorial_data.tgz) to a directory of your choice. Decompress the file using the following command:

```
tar xzvf tutorial_data.tgz
```

or, alternatively...

```
gzip -d tutorial_data.tgz
```

and then... `tar xvf tutorial_data.tar`

This command will create the `tutorial_data` directory, which contains four subdirectories: `tutorial_1`, `tutorial_2`, `tutorial_3`, `tutorial_4` and `tutorial_5`. Each of these directories contains a set of subdirectories, as exemplified below:

- `seed` - this directory contains the seed(s) files.
- `db` - this directory should contain your database multiple sequence FASTA file. Some databases are provided, but some others, especially those that are relatively larger, must be downloaded by the user and stored in this directory. The appropriate links for download are informed.
- `test` - this directory will be your working directory for the tutorial and is initially empty.
- `results` - this directory contains output files previously generated by GenSeed for the corresponding tutorial. We are providing these files just in case you have problems in running GenSeed, and want to check how the output files should look for the respective data set.
- `vector` - this directory, when provided, contains UniVec, NCBI's database of cloning vector sequences, plus sequences of adapters, linkers, and primers commonly used in the process of cloning cDNA or genomic DNA. You can also download the respective file from <ftp://ftp.ncbi.nih.gov/pub/UniVec/>.

2. Third-party program requirements

To run this tutorial, you need first to have the programs listed below installed in your server. All the executables must be in directories listed in the system `PATH`, otherwise GenSeed will not be able to invoke them.

- **BLAST package.** BLAST can be downloaded from <ftp://ftp.ncbi.nih.gov/blast/executables/release/>. GenSeed uses the following executables from the package: `formatdb`, `blastall`, `fastacmd` and `bl2seq`.

- GenSeed can use **Cross_match** if vector masking is desired. GenSeed generates *.ace assembly files that can be easily inspected using **Consed**. Cross_match and Consed are components of the Phred/Phrap/Consed package. Instructions for obtaining this package are provided at <http://www.phrap.org/phredphrapconsed.html>.
- **CAP3**. CAP3 assembly program can be downloaded from <http://seq.cs.iastate.edu/>.
- **HTML browser**. GenSeed produces text and graphical output files in HTML format. These files can be inspected in any HTML browser such as Mozilla Firefox or Internet Explorer.

3. The tutorials

The tutorials describe how to run GenSeed in several applications of a seed-driven progressive assembly. The tutorials follow a logical order and the concepts are gradually introduced. For these reason, we strongly recommend you to practice the tutorials in the suggested order. Four tutorials are provided:

- Tutorial 1 - Reconstruction of a gene or genomic segment using DNA and protein sequence seeds
- Tutorial 2 - cDNA reconstruction using DNA and protein seeds
- Tutorial 3 - cDNA reconstruction using proteomic data
- Tutorial 4 - cDNA reconstruction using SAGE data
- Tutorial 5 - Reconstruction of a complete extrachromosomal genome using DNA and protein seeds

Please let the authors know if you have any suggestion or need any help on using GenSeed. Enjoy your (sequence) germination!