Documentation for Antibiotic Resistance Genes Annotation Tool

Bo Liu

July 3, 2009

Contents

1 Introduction

This program `ardbAnno.pl` is a command-line version of the antibiotic resistance genes annotation tool in Antibiotic Resistance Genes Database at http://ardb.cbcb.umd.edu. Basically, this program takes a set of genes or reads in FASTA format, and can identify and annotate potential antibiotic resistance genes using BLAST similarity search.

2 Installation

This annotation tool requires the installation of Perl, BioPerl toolkit and BLAST programs.

Uncompress the file by typing

```
tar xzvf ardbAnno.tar.gz
```

A directory named `ardbAnno1.0` should result. In this directory, there is a subdirectory named `blastdb`, which contains the reference resistance genes `resisGenes.pfasta`. To compile this data set for BLAST, within this `blastdb` subdirectory type

```
formatdb -i resisGenes.pfasta -p T -o T
```

3 Program Usage

3.1 Input your samples

In the directory `ardbAnno1.0`, there is a file named as `genomeList.tab`, which contains a list of FASTA files that need to be annotated. Open this file using text editor, the content format is

```
>test
test.pfasta
```

`>test` specifies the name of the sample is `test` and the second line is the location of your sample FASTA sequence file. You can include multiple samples into this file, for example

```
>test1
test1.pfasta
>test2
test2.pfasta
```

3.2 Run the annotation program

After you have specified your samples in the previous step, you are ready to run the annotation program. Within the directory `ardbAnno1.0`, type

```
perl ardbAnno.pl
```

3.3 Results

For each sample there are two result files: (1) raw blast file (with blastp or blastx suffix); (2) annotation file (with anno suffix). In addition, for all the samples specified in file `genomeList.tab`, their annotations are summarized into an excel sheet `output.xls`. Format of 'anno' file

```
>NP_878016
BestHit CAL90940
```

```
Type vana
Description VanA type vancomycin resistance operon genes ...
Resistance vancomycin teicoplanin
Require vanra vanha vansa vanya vanxa
Evalue 0.0
QueryLength 343
HitLength 343
HSPlength 343
PercentIdentity 100.00
```

First line >NP_878016 is the ID of the query sequence. BestHit means
the best BLAST hit in the reference antibiotic resistance genes dataset
`resisGenes.pfasta`. Type is the gene type name of the best BLAST hit.
Resistance is the antibiotic resistance profile. Require means vana gene re-
quires other cooperative genes to perform resistance function. HSPlength is
the high scoring pair alignment length from BLAST.

4 Customization (Change Parameters)

For a gene to be annotated as resistance gene, two criteria are used in this
program: (1) E-value; (2) percent identity. The E-value cutoff is specified
in variable `$evalueCutoff` in program `ardbAnno.pl`, which means all resis-
tance genes identification using this same evalue cutoff. The percent identity
for different resistance gene types vary, and this value is specified in the last
column of file `originType.tab` in directory `tabs`.

4.1 Change evalue cutoff

Open program `ardbAnno.pl`, change the value of the variable `$evalueCutoff`
to the value you need.

4.2 Change percent identity

Open file `originType.tab` in directory `tabs`, for each resistance gene (row),
change the value (in percentage) of the last column to your desired percent
identity.

5 Citation

Please cite the following paper as the reference to this program.

Liu, B. and M. Pop (2009). "ARDB–Antibiotic Resistance Genes Database."
Nucleic Acids Res 37(Database issue): D443-7.