BioCompass Documentation

Release dev

November 16, 2016

Contents

Contents:

CHAPTER 1

BioCompass

Python package for gene clustering

- Free software: BSD license
- Documentation: https://BioCompass.readthedocs.io.

1.1 What is BioCompass?

An emergent need in the field of natural products is to dereplicate biosynthetic pathways at the genomic level. This dereplication consists of grouping the biosynthetic gene clusters (BGCs) into families according to their nucleotide sequence homology, and a procedure known as ‘gene cluster networking’. Since the source code from published networking approaches are still not publicly available, we adapted our own strategy for the discovery of gene cluster families, named Biosynthetic Gene Cluster Comparative Synteny Software (BioCompass).Please note that this is a beta version of BioCompass which is still undergoing final testing before its official release. The website, its software and all content found on it are provided on an “as is” and “as available” basis.

1.2 How BioCompass Works?

BioCompass groups BGCs into gene cluster families based on synteny and homology. These clusters need to be identified by antiSMASH, with the ClusterFinder option preferentially turned off. A similarity matrix is used to divide each given BGC into subclusters based on synteny with the best MultiGeneBLAST hits (obtained using antiSMASH 3.0) and the functional annotation of each gene in the queried cluster. This information is then incorporated into a query-specific database to search for the best matches for each subcluster. This newly created database includes microbial BGCs identified by antiSMASH (downloaded from NCBI database, Genbank NR) and the latest version of the well annotated MIBiG database of known gene clusters. Additional gene clusters, for example missing from NCBI and MIBiG, can also be added in by the user. Final similarity scores are calculated by MultiGeneBLAST for each subcluster and then stored as tables. The outputs can be displayed as a network diagram using Cytoscape v3.2.1.

1.3 Future Implementations

One of the issues of using networking approaches in the natural products research area concerns the concept of networking itself. For accurate dereplication of families (both molecular families, as used in GnPS (link), and gene cluster families), it’s required to define a threshold for which once trespassed, two gene clusters are not part of the same family anymore. Analogously to the species definition when using the 16S rRNA gene, this threshold is empirical and can be imprecise in some cases. Hence, BioCompass envisions to implement a cutoff calibration feature to minimize this issue. The new feature consists on the user evaluating the network diagram for both gene homology (scored via multigene blast) and domain homology (score via Jaccard index, another feature to be implemented soon), visually deciding which cutoff would better represent those scores for the particular query. The user will use an internal standard to aid in the decision making process.

CHAPTER 2

Installation

2.1 Stable release

To install BioCompass, run this command in your terminal:

```
$ pip install BioCompass
```

This is the preferred method to install BioCompass, as it will always install the most recent stable release.

If you don’t have pip installed, this Python installation guide can guide you through the process.

2.2 From sources

The sources for BioCompass can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/NP-Omix/BioCompass
```

Or download the tarball:

```
$ curl  -OL https://github.com/NP-Omix/BioCompass/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ cd BioCompass/

$ python setup.py install
```

CHAPTER 3

Usage

The current version of BioCompass can network the gene clusters from ONE strain/genome against the gene clusters from NCBI/GenBank and MIBiG databases. Future Implementations will provide the option to analyze multiple strains.

3.1 Using BioCompass in your queried genome at your machine

If your installation of BioCompass succeeded, follow the instructions below. If not, please report an issue on our GitHub repo and include the error message you obtained during your installation.

1. Submit your query genome to antiSMASH and "Download all results" (top right corner, icon shaped as a download arrow). If you don't have a sample for testing, feel free to use the strain Moorea producens PAL 15AUG08-1 and its antiSMASH result;
2. Once you downloaded the files, descompress and save them into a folder named for example "antiSMASH_input". We advise renaming the subfolder inside antiSMASH_input to a shorter name representing your strain. As an example for this tutorial, since we're using the genome of Moorea producens PAL 15AUG08-1, we'll rename the subfolder to "PAL";
3. Download multigeneBLAST 1.1.14 for command line;
4. Now, execute BioCompass using the following comand:

```
    cd path/to/BioCompass/BioCompass
    make INPUTDIR='path/to/antiSMASH_input' REFNAME='NAME' MULTIGENEBLASTDIR='path/to/multigeneblast
```

For our example, using Moorea producens PAL 15AUG08-1 and assuming that both antiSMASH_input and multigeneBLAST are in the main BioCompass folder, the command should look like:

```
make INPUTDIR='/home/Desktop/BioCompass/antiSMASH_input' REFNAME='PAL' MULTIGENEBLASTDIR='/home/Deskt
```

PS: the folder name inside antiSMASH_input and the REFNAME must be the same!

5. Select the cutoff you would like to use for filtering your results. We advise using a low cutoff first, checking the network diagram and then revisiting this step using the code below (that only reruns this step, not the whole pipeline) to find the best cutoff for your data:

```
    make ???
```

6. The table outputs (REFNAME_edges.txt and REFNAME_nodes.txt) can be vizualized as a network diagram using Cytoscape 3.2.1, according to the tutorial here (tutorial under construction!)

PS2: The running time of BioCompass at an average laptop (e.g. MacBook Pro 2.6 GHz Intel Core i5 8 GB 1600 MHz DDR3) for Moorea producens PAL 15AUG08-1 (which contains 44 gene clusters) was approximately 7 hours. We advise the use of the command screen.

3.2 Using BioCompass in your queried genome at Amazon Web Service

We expect that the tutorial above would work at your machine, but in case it doesn't, we can garantee that it will work on AWS if followed the instructions below. The instructions for running BioCompass at AWS are very similar than explained in the tutorial above, although requires the following setup first.

(Tutorial under construction)

3.3 Using BioCompass in your project

To use BioCompass in a project:

```
import BioCompass
```

CHAPTER 4

Contributing

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given.

You can contribute in many ways:

4.1 Types of Contributions

4.1.1 Report Bugs

Report bugs at https://github.com/castelao/gene_cluster_network/issues.

If you are reporting a bug, please include:

- Your operating system name and version.
- Any details about your local setup that might be helpful in troubleshooting.
- Detailed steps to reproduce the bug.

4.1.2 Fix Bugs

Look through the GitHub issues for bugs. Anything tagged with “bug” and “help wanted” is open to whoever wants to implement it.

4.1.3 Implement Features

Look through the GitHub issues for features. Anything tagged with “enhancement” and “help wanted” is open to whoever wants to implement it.

4.1.4 Write Documentation

Gene Cluster Network could always use more documentation, whether as part of the official Gene Cluster Network docs, in docstrings, or even on the web in blog posts, articles, and such.

4.1.5 Submit Feedback

The best way to send feedback is to file an issue at https://github.com/castelao/gene_cluster_network/issues.

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.
- Remember that this is a volunteer-driven project, and that contributions are welcome :)

4.2 Get Started!

Ready to contribute? Here's how to set up *gene_cluster_network* for local development.

1. Fork the *gene_cluster_network* repo on GitHub.
2. Clone your fork locally:

```
$ git clone git@github.com:your_name_here/gene_cluster_network.git
```

3. Install your local copy into a virtualenv. Assuming you have virtualenvwrapper installed, this is how you set up your fork for local development:

```
$ mkvirtualenv gene_cluster_network
$ cd gene_cluster_network/
$ python setup.py develop
```

4. Create a branch for local development:

```
$ git checkout -b name-of-your-bugfix-or-feature
```

   Now you can make your changes locally.

5. When you're done making changes, check that your changes pass flake8 and the tests, including testing other Python versions with tox:

```
$ flake8 gene_cluster_network tests
$ python setup.py test or py.test
$ tox
```

   To get flake8 and tox, just pip install them into your virtualenv.

6. Commit your changes and push your branch to GitHub:

```
$ git add .
$ git commit -m "Your detailed description of your changes."
$ git push origin name-of-your-bugfix-or-feature
```

7. Submit a pull request through the GitHub website.

4.3 Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. The pull request should include tests.

2. If the pull request adds functionality, the docs should be updated. Put your new functionality into a function with a docstring, and add the feature to the list in README.rst.
3. The pull request should work for Python 2.6, 2.7, 3.3, 3.4 and 3.5, and for PyPy. Check https://travis-ci.org/castelao/gene_cluster_network/pull_requests and make sure that the tests pass for all supported Python versions.

4.4 Tips

To run a subset of tests:

```
$ py.test tests.test_gene_cluster_network
```

CHAPTER 5

Credits

5.1 Development Lead

- Tiago Leao <tferreir@ucsd.edu>
- Gui Castelão <guilherme@castelao.net>

5.2 Contributors

None yet. Why not be the first?

CHAPTER 6

History

6.1 0.0.1 (2016-06-24)

- First python package prototype.

CHAPTER 7

Indices and tables

- genindex
- modindex
- search