dcj-sat

Git Repository Setup and Usage Manual

Contents

1 Setting up the Repository **2**

1.1 Cloning the Repository 2

1.2 Setting up the Environment 2

1.3 Building Required Files 2

2 Input File Format **2**

3 Testing **2**

3.1 Running the Test 2

3.2 Test Files 3

3.2.1 Real Data 3

3.2.2 Simulated Data 3

1 Setting up the Repository

1.1 Cloning the Repository

```
# Clone the repository to your local machine
git clone https://github.com/Shao-Group/dcj-sat.git
cd dcj-sat/
```

1.2 Setting up the Environment

```
# Create a Python virtual environment using venv (example)
python3 -m venv venv
source venv/bin/activate

# Install the pySAT library
pip install 'python-sat[aiger,approxmc,cryptosat,pblib]'
```

1.3 Building Required Files

```
# Make the build script executable and run it
chmod +x build.sh
./build.sh
```

2 Input File Format

Each input file represents a genome and should conform to the following format:

```
<GENE_ID> <GENE_FAMILY> <CHROMOSOME_NAME> <CHROMOSOME_TYPE>
```

- GENE_ID: A unique identifier for each gene.
- GENE_FAMILY: The gene family as an integer.
- CHROMOSOME_NAME: Name of the chromosome as an integer.
- CHROMOSOME_TYPE: Type of chromosome (1 for linear, 2 for circular).

3 Testing

3.1 Running the Test

```
# Make the testing script executable and run it
chmod +x run_sat.sh
./run_sat.sh <path_to_g1_file> <path_to_g2_file>
```

3.2 Test Files

3.2.1 Real Data

You can find the test files for real data in the ’test_files/real_data’. Inside it are three directories corresponding to including all genes and including genes with less than 2 and less than 3 gene families. In each each of these folders is a folder corresponding to each pair. For eg, to compare gorilla and human containing genes with less than 3 gene families:

```
<path_to_g1_file>
test_files/real_data/less_than_three/gorilla_human/gorilla.dcj

<path_to_g2_file>
test_files/real_data/less_than_three/gorilla_human/human.dcj
```

3.2.2 Simulated Data

Test files for simulated data can be found in ’test_files/simulations’. Inside are two folders ’variable_dcj_ops’ and ’variable_gene_families’. Each of these folders contain a folder corresponding to an instance. This folder contains three files, the original genome, and the two pairs of genomes to compare. For eg, to run the package on the genome with 500 genes, 340 gene families and 150 DCJ operations:

```
<path_to_g1_file>
test_files/simulations/variable_gene_families/sim_500_340_150/
    sim_500_340_150_1.dcj

<path_to_g2_file>
test_files/simulations/variable_gene_families/sim_500_340_150/
    sim_500_340_150_2.dcj
```