CloudForest Documentation

Release 2.0.1

Nicholas Crawford and Brant Faircloth

December 02, 2015

Contents

Contents:

CHAPTER 1

Quick Start:

The first thing you need to do to get CloudForest running to get it installed. We've tried to make this as painless as possible by using package installers and , but you'll still need to type some commands at the command line.

1. Open terminal.

   Prepare to cut-n-paste!

2. Install Python.

   Python should already be installed on OS X or Linux. CloudForest requires version Python2.7 so enter `python --version` at the commandline and make sure you're uptodate. If you're not, I recommend installing **'Enthought Python'_** 2.7 which includes almost all the dependancies necessary to get CloudForest running.

3. Install R.

   R is *the* open source statistical software package. You should be able to download and install a graphical package installer from the R website. We recommend using a recent version such as R 2.15 or greater.

4. Install Git.

   Git is a distributed version control systerm. It's open source, easy to use, and integrates with github for easy collaboration and distribution.

   - We recommend first installing HomeBrew and then running `brew install git` at the command line to install git.

5. Install Pip.

   Pip is a package installer for python that makes adding and managing packages and modules a breeze.

6. Install CloudForest by running `pip install -U cloudforest` at the commandline. Alternately you can install the most cutting edge development version by running `pip install git+git://github.com/ngcrawford/CloudForest.git` at the commandline.

- CloudForest's `setup.py` script should install all the dependancies you need. However, Numpy can be troublesome. If you get errors when numpy is building you may need to install numpy manually. Running `sudo pip install numpy` should do the trick.

7. Install Phybase.

- First install ape Phybase's only dependancy.
  - To do this open R. At the R command line type: `install.packages('ape')`. Follow the instructions.
- Then download the gziped source code for phybase.

- `cd` to that directory and type `R CMD INSTALL phybase_1.3.tar.gz`. If the version of phybase is newer, you may have to adjust the gzip filename to refect this.

That should do it for installation.

1.1 Configure Software:

CloudForest itsself doesn't require any configuration, but one of it's dependancies does. MrJob has a pretty simple config file you'll need to setup if you want to run analyses on EMR. A full explaination is available here.

You'll need to make a mrjob.conf file. Mine looks something like this:

```
runners:
  emr:
    aws_access_key_id: YOURIDYOURIDYOURIDYOURIDYOURID
    aws_region: us-east-1
    aws_secret_access_key: YOURSECRETKEYYOURSECRETKEYYOURSECRETKEY
    ec2_key_pair: mr-keypair
    ec2_key_pair_file: /Users/YourUserName/.ssh/mr-keypair.pem
    ec2_instance_type: m1.small
    num_ec2_instances: 1
    setup_cmds: &setup_cmds
    ssh_tunnel_is_open: true
    ssh_tunnel_to_job_tracker: true
```

1.2 Test the Installation:

1. Run the UnitTests with nose.

I'm not yet sure how to do this with an 'installed package.'

CHAPTER 2

FAQ:

Why should you use CloudForest to infer a Species-Tree?

> Incomplete Lineage Sorting (ILS) is becoming increasing recognized as influential process in genomic evolution. ILS describes what happens when by random chance different versions of the same allele end up in the 'wrong' species during speciation and thus produce conflicting phylogenies along the genome. If you are interested in inferring the phylogenetic relationships between species from genome sized data (= hundreds of loci) it's important to account for ILS. Unfortuantely traditional 'concatenation style analyses' do not account for ILS. Cloudforest provides a pipeline that incorporates ILS into phylogenetic inference from thousands of independant loci. For inferring many thousands of trees, you'll need to do this if you want to bootstrap your dataset, as well as a wrapper for the R package Phybase which provides a number of methods for generating species-trees from gene-trees.

Why should you use CloudForest

CHAPTER 3

Indices and tables

- genindex
- modindex
- search