Python in Bioinformatics

Gang Chen
chengang@bgitechsolutions.com

November 8, 2014

华大科技 BGI·tech
香港中文大學 The Chinese University of Hong Kong

Outline

Next

Python Package

- python setup
- easy_install
- pip

Python Package Development

- Tutorial:https://packaging.python.org/en/latest/distributing.html
- Example:https://github.com/pypa/sampleproject

Example

see hello directory

Next

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

scipy project

Installation and Example

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

BioPython

Biopython is a set of freely available tools for biological computation written in Python by an international team of developers.

Installation and Example

- Download
- python setup.py
- see bio.py as an example

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

Python for Machine Learning

- scikit-learn
- pyml

PyML

PyML is an interactive object oriented framework for machine learning written in Python. PyML focuses on SVMs and other kernel methods.

scikit-learn

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

igraph for network visualization

igraph

igraph is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. igraph is open source and free. igraph can be programmed in GNU R, Python and C/C++.

igraph is implemented in C++, but can be programmed in R, Python and C/C++.

Install and Example

- pip install python-igraph
- Tutorial: http://igraph.org/python/doc/tutorial/tutorial.html

Next

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

Clouding Computing

- Amazon Web Service: aws.amazon.com
- Aliyun: aliyun.com
- Google Compute Engine: cloud.google.com
- Microsoft Azure: azure.microsoft.com
- ...

NCI Cancer Genomics Cloud Pilots

Current Needs in Cancer Research

The challenges posed by the need to disseminate, manage, and interpret large, multi-scale data pervade efforts to advance understanding of cancer biology and apply that knowledge in the clinic.

Bioinformatics in the Cloud

- DNANexus: DNANexus.com
- SBGenomics: SBGenomics.com
  rabix: rabix.org
- GeneDock: GeneDock.com
- L3-Bioinformatics: l3-bioinfo.com
- tute genomics, Variant Analysis from Qiagen, ...

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

Overview

https://wiki.dnanexus.com/Developer-Portal

Installation

```
pip install dxpy
```

Next

1. Python Package
2. Python based Bioinformatics Projects
   - Scipy
   - BioPython
   - Machine Learning
   - Network Visualization and Analysis
3. Bioinformatics in the Cloud using Python
   - Clouding Computing and Bioinformatics
   - Python SDK of DNANexus
   - Rabix from SBGenomics

Reproducible Research

Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better understood and verified.

References

- https://www.coursera.org/course/repdata
- http://cran.r-project.org/web/views/ReproducibleResearch.html

Rabix Project

see rabix_ismb.pdf