FAIR_bioinfo for bioinformaticians

Introduction to the tools of reproducibility in bioinformatics

C. Hernandez[1] T. Denecker[1] J. Sellier[2] G. Le Corguillé[2]
C. Toffano-Nioche[1]

[1]Institute for Integrative Biology of the Cell (I2BC)
UMR 9198, Université Paris-Sud, CNRS, CEA
91190 - Gif-sur-Yvette, France

[2]IFB Core Cluster taskforce

Sept. 2020

Conda

CONDA: an environment manager

Conda concepts, objetcs

- Environment: a set of packages/tools in a directory (added to our PATH)
- Conda: an open source package + a general-purpose environment management system (installation, execution, upgrade). For any programming language, multi-platform (Windows, MacOS, Linux).
- Conda package: a compressed tarball of a tool

Why using an environment manager?

- avoid compilation and dependencies problems: an environment manager will take care of everything!
- have several environments in parallel each with their own set of tools
- useful when cross-tools dependencies are incompatible with each other

CONDA: Access

Conda distribution

- Anaconda: a data science platform, comes with a lot of packages
- Miniconda: come without installed packages

Anconda cloud, the "conda hub"

- Anaconda cloud (private company) relies on the community of developers, concerns many domains (Machine Learning, Data Visualization, Dashboarding-web, Image Processing, Natural Language Processing, etc)
- Anaconda cloud: made up of channels/owners. Each channels contains one or more conda packages
- be careful when downloading any packages from an untrusted source, always inspect before installation

About channels

Some conda channels

- defaults
- bioconda: bioinformaticians contributions
- conda-forge: many popular python packages (analogous to PyPI but with a unified, automated build infrastructure and more peer review of recipes)
- r: for packages in R language

Channels list order

- when different channels have the same package $\Rightarrow$ collisions
- collisions resolved following the order of your channels list $\Rightarrow$ put supplemental channels at the bottom of your channel list

CONDA command

simple commands

```
conda create env -n myenv # creation of a conda environment
conda info --envs # list environments (* for the active one)
conda activate myenv # active the myenv environment
conda deactivate # inactivate the environment
conda list # list packages (only in an active environment)
conda install package # installation of a tool/package
conda remove package # suppress the tool from the
    environment
conda env remove -n myenv # suppress the myenv environment
```

miniconda3

With the miniconda3 distribution and by default, environments are installed in a miniconda3/envs/ repository

CONDA 2 modes

interactive

- create an environment
- activate the environment
- install some conda packages

configuration file

- list all conda packages in a configuration file (yaml or json format)
- create the environment based on the configuration file (option `-f`)
- activate the environment

reproducibility

- good practice: use a configuration file
- specify a precise version of a package:
  `<channel>::<package>=<version>`

Conda Exercise

Conda setup

How to access conda?

- Conda is so used that it could even be installed by default to your machine. To test this: `conda --version`
- if not, may install it or got it by a docker image:

```
docker run -i -t -v ${PWD}:/data continuumio/miniconda3
```

- on the IFB cluster, with modules: `module load conda`

Conda environment

We have already (blindly) use a conda configuration file in the workflow session:

```
conda env create -n envfair -f envfair.yml
conda activate envfair
```

We will next detail the content of the configuration file, `envfair.yml`

Example of a conda configuration file

envfair.yml

```
channels:
  - conda-forge
  - bioconda
  - main
  - default
dependencies:
  - python=3.7.6 # specify python version (not required but
    can help with downstream conflicts)
  - snakemake-minimal=5.10.0 # workflow manager
  - graphviz=2.42.3 # for visualisation
  - xorg-libxrender
  - xorg-libxpm
  - wget=1.20.1 # for downloading files
  - fastqc=0.11.9 # for the RNAseq analysis
  - bowtie2=2.4.1
  - samtools=1.10
  - subread=2.0.1
```

How to access tools?

Manage Conda environment

1. create the working environment:

```
conda create env -n myenv
```

2. activate it:

```
conda activate myenv
```

3. if not yet done, install packages (specify the channel):

```
conda install -c bioconda bowtie2
```

4. work with the tools
5. quite the environment:

```
conda deactivate
```

Install Snakemake

Objective

Create a conda configuration file to install the snakemake tool.

Hint

- Search its channel in the Anaconda cloud web pages
- the "minimal" environment is sufficient

Install Snakemake

condaEnvSnakemake.yml

```
channels:
  - conda-forge
  - bioconda
  - main
dependencies:
  - snakemake-minimal=5.10.0
```

run

```
conda create env -n condaEnvSnakemake -f condaEnvSnakemake.
    yml
conda activate condaEnvSnakemake
snakemake ...
```