Snakemake workflows guidelines and best-practices

SciLifeLab

NBIS

Structure of a snakemake workflow

Workflows can be organized basically anyway you want

- Rules can be added to a single Snakefile, or split across several files
- Files can be named anything you want (`Snakefile`, `rulefile.rules`, `myrules.smk` etc.)
- You can mix and match snakemake rules with python code

```
def trim_params(config):
  return config["trim_settings"]

rule trim_reads:
  input: "data/reads.fastq.gz"
  output: "results/trimmed.fastq.gz"
  params: settings = trim_params(config)
  shell: "read-trimmer {input} {output}"
```

What are some best-practices and guidelines?

1. Snakemake documentation

Distribution and Reproducibility

> It is recommended to store each workflow in a dedicated git repository of the following structure

```
|-- .gitignore
|-- README.md
|-- LICENSE.md
|-- workflow
|   |-- rules
|   |   |-- module1.smk
|   |   |-- module2.smk
|   |-- envs
|   |   |-- tool1.yaml
|   |   |-- tool2.yaml
|   |-- scripts
|   |   |-- script1.py
|   |   |-- script2.R
|   |-- notebooks
|   |   |-- notebook1.py.ipynb
|   |   |-- notebook2.r.ipynb
|   |-- report
|   |   |-- plot1.rst
|   |   |-- plot2.rst
|   |-- Snakefile
|-- config
|   |-- config.yaml
|   |-- some-sheet.tsv
|-- results
|-- resources
```

What are some best-practices and guidelines?

1. Snakemake documentation

Distribution and Reproducibility

- workflow code goes into a subfolder `workflow/`
- `workflow/Snakefile` marks the entrypoint of the workflow
- the (optional) rules can be stored in subfolder `workflow/rules/` (they should end in `.smk`)
- scripts should be stored in `workflow/scripts/`
- notebooks should be stored in `workflow/notebooks/`
- conda environments should be fine-grained and stored in `workflow/envs/`
- report caption files should be stored in `workflow/report/`
- all output files should be stored under `results/`
- all resource files should be stored under `resources/`
- configuration is stored in a subfolder `config/`

What are some best-practices and guidelines?

2. The Snakemake-Workflows project

> a joint effort to create workflows for common use cases of the Snakemake workflow management system

Guidelines

- A workflow repository shall consist of one Snakemake workflow.
- The workflow should be configurable via a well documented YAML-based configuration file and (when necessary) a sample and a unit sheet.
- Whenever possible, Snakemake wrappers should be used.
- The structure of the workflow should follow our template.

The snakemake workflows template

- A cookiecutter template available on GitHub
- Can be used to initialize new workflows

How to use the template

Step 1. Install `cookiecutter` if you haven't already

```
$ conda create -n cc -c conda-forge cookiecutter
<...>
$ conda activate cc
```

SciLifeLab

NBIS

How to use the template

Step 2. Create a new workflow with cookiecutter

```
$ (cc) cookiecutter gh:snakemake-workflows/cookiecutter-snakemake-workflow
full_name [Johannes Köster]: John Sundh
```

How to use the template

Step 2. Create a new workflow with cookiecutter

```
$ (cc) cookiecutter gh:snakemake-workflows/cookiecutter-snakemake-workflow
full_name [Johannes Köster]: John Sundh
email [johannes.koester@protonmail.com]: john.sundh@scilifelab.se
```

How to use the template

Step 2. Create a new workflow with cookiecutter

```
$ (cc) cookiecutter gh:snakemake-workflows/cookiecutter-snakemake-workflow
full_name [Johannes Köster]: John Sundh
email [johannes.koester@protonmail.com]: john.sundh@scilifelab.se
username [johanneskoester]: johnne
```

How to use the template

Step 2. Create a new workflow with cookiecutter

```
$ (cc) cookiecutter gh:snakemake-workflows/cookiecutter-snakemake-workflow
full_name [Johannes Köster]: John Sundh
email [johannes.koester@protonmail.com]: john.sundh@scilifelab.se
username [johanneskoester]: johnne
project_name [RNA-Seq]: gut-microbiota
repo_name [rna-seq]: gut-microbiome-repo
min_snakemake_version [5.7.0]: 5.11.0
```

How to use the template

Step 3. Inspect your new workflow

```
$ ls gut-microbiome-repo/
-rw-r--r--  1 john  staff   1.0K Oct  1 15:49 LICENSE
-rw-r--r--  1 john  staff   4.9K Oct  1 15:49 README.md
drwxr-xr-x  4 john  staff   128B Oct  1 15:49 config
drwxr-xr-x  3 john  staff    96B Oct  1 15:49 resources
drwxr-xr-x  6 john  staff   192B Oct  1 15:49 results
drwxr-xr-x  8 john  staff   256B Oct  1 15:49 workflow
```

How to use the template

Step 3. Inspect your new workflow

Your info has been inserted into the `README.md` file

```
# Snakemake workflow: gut-microbiota

[![Snakemake](https://img.shields.io/badge/snakemake->5.11.0-brightgreen.svg)](https://snakemake.bitbucket.io)
[![Build Status](https://travis-ci.org/snakemake-workflows/gut-microbiome-repo.svg?branch=master)](https://travis-ci.org/snakemake-workflows/gut-microbiome-repo)

This is the template for a new Snakemake workflow. Replace this text with a comprehensive description covering the purpose and domain.
Insert your code into the respective folders, i.e. `scripts`, `rules`, and `envs`. Define the entry point of the workflow in the `Snakefile` and the main configuration in the `config.yaml` file.

## Authors

* John Sundh (@johnne)
```

How to use the template

Step 3. Inspect your new workflow

Your info has been inserted into the `README.md` file

Snakemake workflow: gut-microbiota

This is the template for a new Snakemake workflow. Replace this text with a comprehensive description covering the purpose and domain. Insert your code into the respective folders, i.e. `scripts`, `rules`, and `envs`. Define the entry point of the workflow in the `Snakefile` and the main configuration in the `config.yaml` file.

Authors

- John Sundh (@johnne)

What are some best-practices and guidelines?

3. Snakemake linting

Since `v5.11` Snakemake comes with code quality checker that can be used on your workflow.

To get a list of warnings for parts of the workflow that does not conform to best-practices, run `snakemake --lint`

What are some best-practices and guidelines?

3. Snakemake linting

Examples of warnings:

- Mixed rules and functions in same snakefile.
- No log directive defined
- Migrate long run directives into scripts or notebooks

```
Lints for rule compose_sample_sheet (line 6, rna-seq-kallisto-sleuth/workflow/rules/diffexp.smk):
    * No log directive defined:
      Without a log directive, all output will be printed to the terminal. In distributed environments, this means that errors are harder to discover. In local environments,
      output of concurrent jobs will be mixed and become unreadable.
      Also see:
      https://snakemake.readthedocs.io/en/stable/snakefiles/rules.html#log-files
```

Questions?

SciLifeLab

NBIS