University of Zurich[UZH]

S3IT

Running tasks in sequence: `SequentialTaskCollection` and `StagedTaskCollection`

Riccardo Murri <riccardo.murri@uzh.ch>

S3IT: Services and Support for Science IT

University of Zurich

November 14–17, 2016

Basic use of SequentialTaskCollection

```
from gc3libs.workflow \
  import SequentialTaskCollection

class MySequence (SequentialTaskCollection) :
  # ...
  def __init__(self, ...):
    app1 = FirstApp(...)
    app2 = SecondApp(...)
    SequentialTaskCollection.__init__(
      self, [app1, app2])
```

A `SequentialTaskCollection` runs a list of tasks one at a time, in the order given.

Basic use of SequentialTaskCollection

```
from gc3libs.workflow \
  import SequentialTaskCollection

class MySequence(SequentialTaskCollection):
  # ...
  def __init__(self, ...):
    app1 = FirstApp(...)
    app2 = SecondApp(...)
    SequentialTaskCollection.__init__(
      self, [app1, app2])
```

Initialize a `SequentialTaskCollection` with a list of tasks to run.

Running tasks in sequence

```
class MyScript(SessionBasedScript):
  # ...
  def new_tasks(self, extra):
    tasks_to_run = [
      MySequence(...)
    ]
    return tasks_to_run
```

You can then run the entire sequence by returning it from `new_tasks()`.

Detour: BLAST, again

Another use of the BLAST tool is to search for given "query" proteins in a data base. Large curated DBs are available, but one may want to build a custom DB.

Building a DB from a set of FASTA-format files `p1.faa` `p2.faa` and `p3.faa`, and querying it is a 3-step process:

```
cat p1.faa p2.faa p3.faa > db.faa
formatdb -i db.faa
blastpgp -i q.faa -d db.faa -e ...
```

The `formatdb` step produces output files `db.faa.phr`, `db.faa.pin`, and `db.faa.psq`; all these files are *inputs* to the `blastpgp` program.

Exercise 8.A: Write a `blastdb.py` script to build a BLAST DB and query it.

The `blastdb.py` script shall be invoked like this:

```
$ python topblast.py query.faa p1.faa [p2.faa ...]
```

where arguments `new.faa`, `p1.faa`, etc. are FASTA-format files.

The script should build a BLAST DB out of the files p*N*.faa. Then, it should query this database for occurrences of the proteins in `query.faa` using `blastpgp`.

Exercise 8.B: Find out by running the `blastdb.py` script of Ex. 8.A:

1. What happens if an intermediate step fails and does not produce complete output?

2. After the whole sequence turns to TERMINATED state, what is the value of its signal and exitcode?

3. How could you implement a "cleanup" feature that removes intermediate results (e.g., the "`.phr`" files) and only keeps the output from `blastpgp` **if the whole sequence was successfully executed?**

Running jobs in sequence

`StagedTaskCollection` provides a simple interface for constructing sequences of tasks, but only when the number and content of steps is *known and fixed* at programming time.

(By contrast, the most general `SequentialTaskCollection` can alter the sequence on the fly, insert new stages while running and loop back. But the code is also harder to write.)

```
class Pipeline (StagedTaskCollection) :
  def __init__(self, image):
    StagedTaskCollection.__init__(self)
    self.source = image

  def stage0(self):
    # run 1st step
    return Application(...)

  def stage1(self):
    if self.tasks[0].execution.exitcode != 0:
      # set collection signal
and exit code,
      # and state to
TERMINATED
      return (0, 1)
    else:
      # run 2nd step
      return Application(...)
  # ...
```

Example of a `StagedTaskCollection` subclass.

```
class Pipeline(StagedTaskCollection):
  def __init__(self, image):
    StagedTaskCollection.__init__(self)
    self.source = image

  def stage0(self):
    # ...

  def stage1(self):
    # ...

  # ...
  def stageN(self):
    # ...
```

Stages are numbered starting from 0.

You can have as many stages as you want.

```
class Pipeline(StagedTaskCollection):
  # ...

  def stage0(self):
    # run 1st step
    return Application(
      ['convert', self.source,
       '-colorspace', 'gray',
       'grayscale_' + self.source],
      inputs = [self.source],
      ...)

  # ...
```

Each `stage`*N* method can return a `Task` instance, that will run as step *N* in the sequence.

```
class Pipeline(StagedTaskCollection):
  # ...

  def stage1(self):
    if self.tasks[0].execution.exitcode != 0:
      # set collection signal
and exit code,
      # and state to
TERMINATED
      return (0, 1)
    else:
      # run 2nd step
      return Application(...)

  # ...
  def stageN(self):
    # ...
```

In later stages you can check the exit code of earlier ones, and decide whether to continue the sequence or abort.

```
class Pipeline(StagedTaskCollection):
  # ...

  def stage1(self):
    if self.tasks[0].execution.exitcode != 0:
      # set collection signal
and exit code,
      # and state to
TERMINATED
      return (0, 1)
    else:
      # run 2nd step
      return Application(...)

  # ...
  def stageN(self):
    # ...
```

To abort the sequence, return an integer (termination status) or a pair *(signal, exit code)*, instead of a `Task` instance.

Exercise 8.C: Rewrite the `blastdb.py` script from Ex. 8.A to use a `StagedTaskCollection` and be sure to check that a step is successful before proceeding to the next one.

Upon successful completion of the pipeline, move the `blastpgp` output into directory `/home/ubuntu/results` and then delete all intermediate files and directories.