



Développement de tests automatiques pour des outils dans l'environnement Galaxy



Etudiant : Thomas EYMARD

Tuteur de stage : Bérénice BATUT

Stage de 2^e année de DUT Bio-informatique

Année Universitaire 2015-2016

Galaxy

- Plate-forme ouverte et open source
- Utilisation de Galaxy :
 - Galaxy dispose de serveurs publics
 - Il existe de nombreuses instances Galaxy « fermées »
 - Possibilité d'installer galaxy sur n'importe quel ordinateur
- Permet de lancer des outils bioinformatiques en ligne de commande afin d'effectuer des analyses complètes et reproductibles

Galaxy et wrapper

- Plupart des outils bioinformatiques utilisables en ligne de commande

```
graphlan.py input.xml image.png --format pdf --dpi 100 --size 7 --pad 2
```

- Wrapper sert d'interface entre l'utilisateur et l'outil
 - Utilise une version fixée de l'outil
 - Gère les fichiers et les paramètres d'entrée
 - Gère les dépendances ainsi que les packages

Galaxy et wrapper

The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the center. The top right corner shows 'Using 0 bytes'.

The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: 'COMMON TOOLS' (Get Data, Manipulate files, Manipulate sequence files, Manipulate BAM/SAM files), 'SEQUENCE PREPARATION TOOLS' (Assemble paired-end sequences, Control quality, Chimera detection, dereplication, clustering, ... with VSEARCH, Cluster sequences, Manipulate RNA, Search similarity, Map against reference genomes), 'METAGENOMIC TOOLS' (Assign taxonomy on all sequence type, Analyze metabolism, Combine functional and taxonomic results), and 'STATISTICS AND VISUALIZATION TOOLS' (Visualize data, Export to GraPhlAn, GraPhlAn to produce graphical output of an input tree, Generation, personalization and annotation of tree for GraPhlAn, Krona pie chart from taxonomic profile, Plot barplot with R, Plot grouped barplot with R, Plot generic X-Y plot with R, Histogram of a numeric column).

The main content area shows the 'GraPhlAn' tool configuration page. The title is 'GraPhlAn to produce graphical output of an input tree (Galaxy Version 0.9.7)'. The 'Input tree' section has a file upload button and a text input field with the placeholder 'No txt dataset available.' and a dropdown menu. Below this is a note: 'The tree must be in PhyloXML, Newick or text format.' The 'Output format' section has a dropdown menu set to 'PNG' with the label '(-format)'. The 'Dpi of the output image (Optional)' section has a text input field. Below this is a note: 'For non vectorial formats (--dpi)'. The 'Size of the output image (in inches)' section has a text input field set to '7' with the label '(-size)'. The 'Distance between the most external graphical element and the border of the image (Optional)' section has a text input field with the label '(-pad)'. At the bottom of the configuration area is a blue 'Execute' button.

Below the configuration area, the 'What it does' section explains that GraPhlAn is a software tool for producing high-quality circular representations of taxonomic and phylogenetic trees, focusing on concise, integrative, informative, and publication-ready representations. It also provides a link to the 'user manual'.

The right sidebar shows the 'History' section with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source'.

Test logiciel

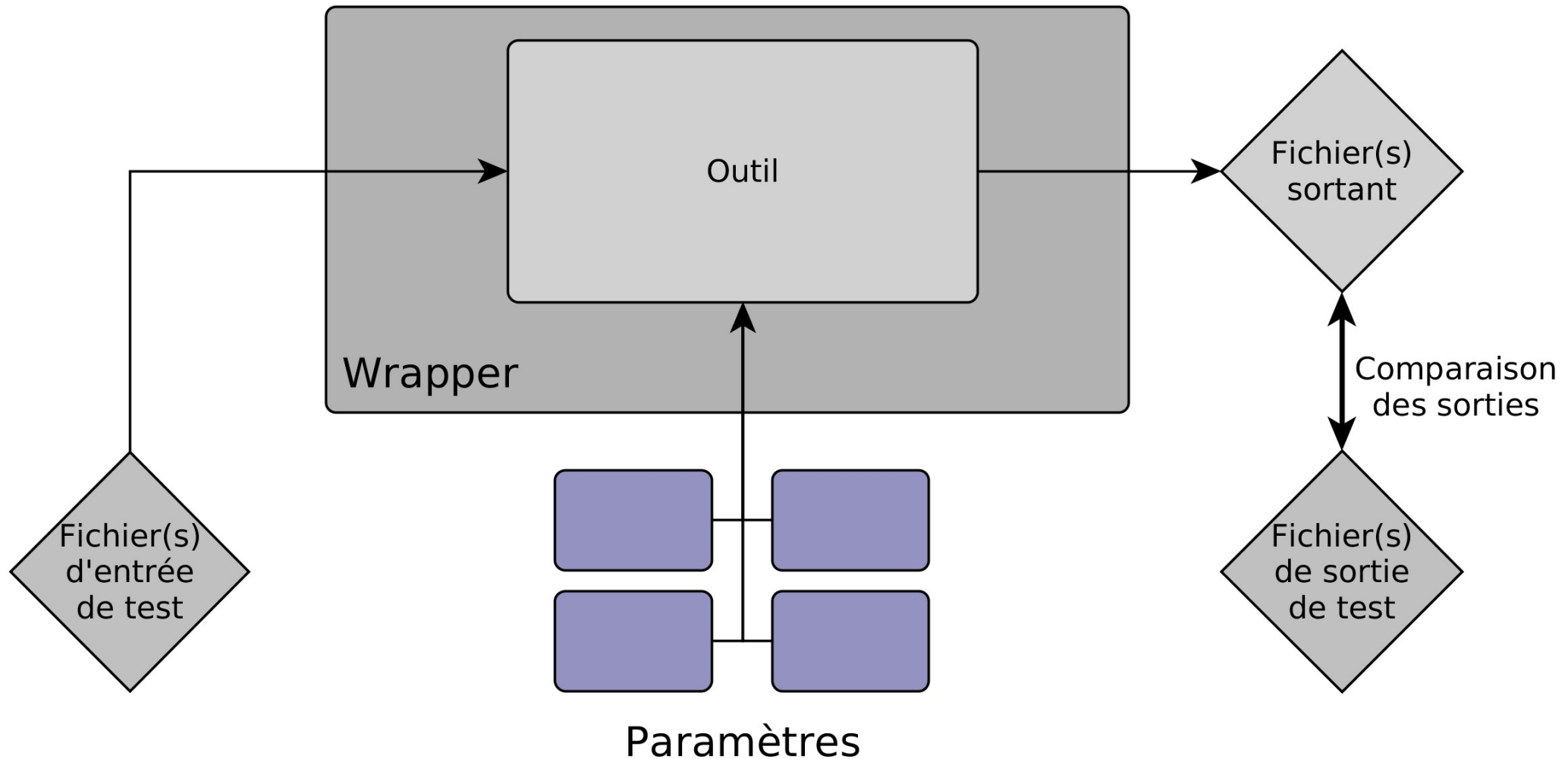
« Technique de contrôle consistant à s'assurer que le comportement d'un programme est conforme à des données préétablies »

- Intérêts du test :
 - Garantir le fonctionnement du programme
 - Garantir que des possibles changements n'affectent pas le comportement attendu du programme

Test logiciel

- Tests Unitaires: test d'un seul élément ou un seul module du logiciel
- Tests fonctionnels : test du comportement d'un programme entier en « conditions réelles »

Le test de wrappers

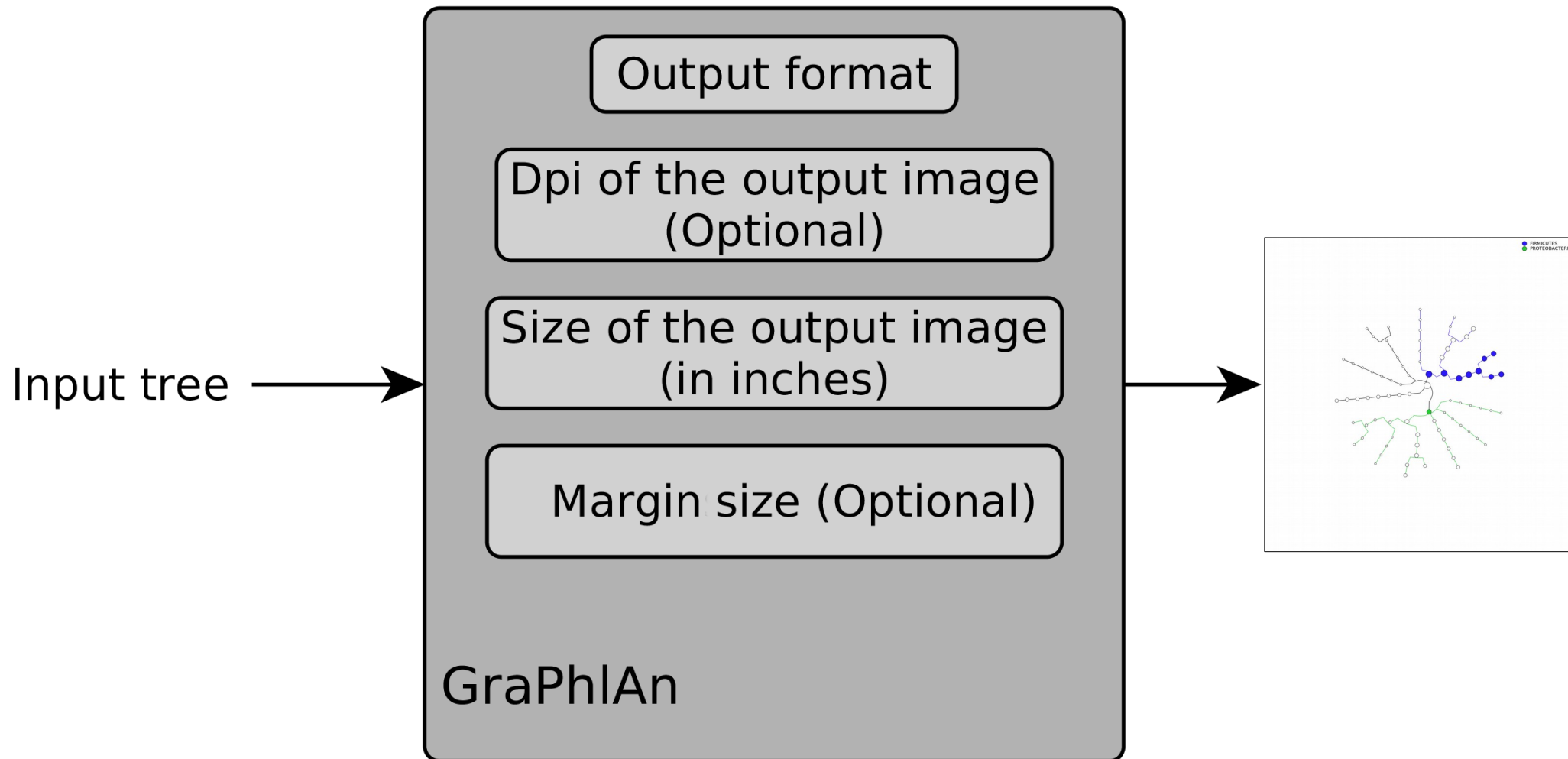


Développement de tests pour des wrappers dans Galaxy

Familiarisation avec l'outil bioinformatique appelé par le wrapper

- Identification des entrées et sorties de l'outil (nombre, format, contenu)
- Identification des paramètres de l'outil
- Génération des entrées/sorties nécessaires pour le test

Familiarisation avec l'outil bioinformatique



Modification du wrapper

```
<tool id="normalize_dataset" name="Normalize a dataset by" version="0.1.0">
  <description>row or column sum to obtain proportion or percentage</description>

  <requirements>
  </requirements>

  <stdio>
    <exit_code range="1:" />
    <exit_code range=":-1" />
  </stdio>

  <version_command></version_command>

  <command><![CDATA[
    python $_tool_directory_/normalize_dataset.py
      --input_file $input_file
      --output_file $output_file
      --normalization $normalization
      --format $format
  ]]></command>

  <inputs>
    <param name="input_file" type="data" format="tabular,tsv,csv" label="Input file" help="File
    in tabular format with tab-separated columns and header in first line (--input_file)"/>

    <param name="normalization" label="Normalization on" type="select" help="--normalization">
      <option value="column" selected="True">Column</option>
      <option value="row">Row</option>
    </param>

    <param name="format" label="Output format" type="select" help="--format">
      <option value="proportion" selected="True">Proportion</option>
      <option value="percentage">Percentage</option>
    </param>
  </inputs>

  <outputs>
    <data name="output_file" format="tabular"
      label="${tool.name} on ${on_string}: Normalized dataset" />
  </outputs>

  <tests>
    <test>
      <param name="input_file" value="input_file.tabular"/>
      <param name="normalization" value="column"/>
      <param name="format" value="proportion"/>
      <output name="output_file" file="output_column_proportion.tabular"/>
    </test>
  </tests>
```

Modification du wrapper

```
<tests>  
  <test>  
  </test>  
</tests>
```



```
<tests>  
  <test>  
    <param name="input_tree" value="intermediary_tree.txt"/>  
    <param name="format" value="png"/>  
    <param name="dpi" value="100"/>  
    <param name="size" value="7"/>  
    <param name="pad" value="2"/>  
    <output name="png_output_image" file="png_image.png" />  
  </test>  
</tests>
```

Tests locaux avec Planemo

- Planemo : outil pour faciliter la création et le développement de wrappers pour Galaxy
- Commandes de test :
 - 'planemo lint' : Test la forme du .xml
 - 'planemo shed_lint' : Test la forme du fichier dans sa globalité
 - 'planemo test' : Test fonctionnel du wrapper dans une instance Galaxy nue

```
planemo test --conda_dependency_resolution --conda_prefix  
$HOME/conda --galaxy_branch release_16.04 --galaxy_source  
$GALAXY_REPO --skip_venv tools/graphlan/
```

Intégration des tests dans le dépôt GitHub

- Outils à tester présents sur le dépôt Github :

<https://github.com/ASaiM/galaxytools>

- Démarche :

- Fork : copie du répertoire
- Création de branche : permet d'éditer des fichiers sans modifier le répertoire d'origine
 - édition et tests
- Pull Request
- Review du code

Intégration dans le dépôt GitHub

Addtest for `combine_methaph_lan2_human2` #16

Edit

Merged bebatut merged 6 commits into ASaiM:master from themard1:br_combine_methaphLan2_human2 on 15 Apr

Conversation 4

Commits 6

Files changed 7

+33,293 -29,435



themard1 commented on 13 Apr

+ 😊

to resolve #1

Labels

None yet

Milestone

No milestone

Assignees

No one assigned

Notifications

Unsubscribe

You're receiving notifications because you authored the thread.

2 participants



themard1 added some commits on 12 Apr



Add missing test

b44e727



Remove some lines

✗ a5222f1



bebatut commented on an outdated diff on 14 Apr

Show 1 comment



bebatut commented on 14 Apr

+ 😊

I will add this test for CI



themard1 added some commits on 14 Apr



Merge branch 'master' of https://github.com/ASaiM/galaxytools into br...

a34c544



Add good outputs and edit tt_blacklist

✓ e79736b



update .tt_blacklist

✓ 3f792eb



bebatut commented on 15 Apr

+ 😊

The test files are not good ones. I can send you some if you want



Change inputs and outputs files

✓ a34b11a



bebatut commented on 15 Apr

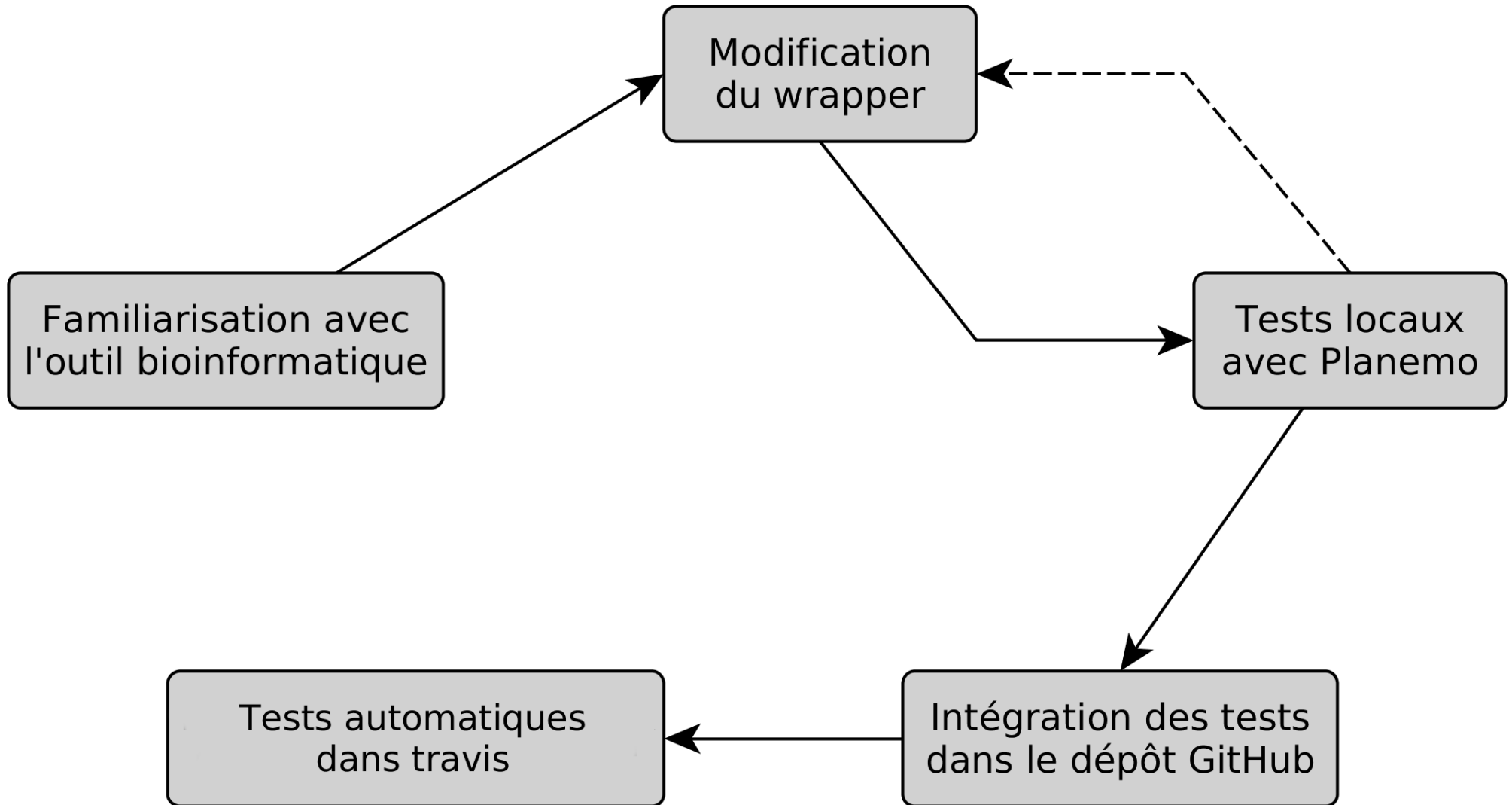
+ 😊



Intégration continue et automatisation des tests

- Intégration continue : Consiste à vérifier à chaque mise à jour que les modifications n'entraînent pas d'erreurs dans l'application développée.
- Tests automatiques : gain de temps
- Travis CI :
 - Création d'environnement totalement vierge et installation des outils nécessaires
 - Pas de pollution dues à d'anciennes versions ou des logiciels tiers
 - Interface avec Github

En résumé



Travail effectué

Wrappers modifiés pour l'ajout de tests

Outils humann2 :

- humann2_join_tables
- humann2_reduce_table
- humann2_regroup_table
- humann2_rename_table
- humann2_renorm_table
- humann2_split_table
- compare_humann2_output

Outil metaphlan2 :

- format_metaphlan2_output

Outil de combinaison :

- combine_metaphlan2_humann2

Wrappers modifiés pour l'ajout de tests

Outils graphlan :

- graphlan
- graphlan_annotate
- export2graphlan

Outils cdhit :

- cd_hit_est
- cd_hit_protein
- format_cd_hit_output

Outils plot :

- plot_barplot
- plot_generic_x_y_plot
- plot_grouped_barplot

Autres outils :

- extract_min_max_lines
- fasta_add_barcode
- normalize_dataset
- compute_wilcoxon_test
- extract_sequence_file

Quelques problèmes rencontrés

humann2_split_table :

Problème : L'outil renvoie un nombre de fichiers dépendant du fichier d'entrée, on ne peut déterminer ce nombre à l'avance

```
<tests>
  <test>
    <param name="input_file" value="joined_pathway_coverage_abundance.tsv"/>
    <output_collection name="split_tables" type="list" >
      <element name="humann2_Abundance" file="split_joined_table_abundances.tsv" />
      <element name="humann2_Coverage" file="split_joined_table_coverage.tsv" />
    </output_collection>
  </test>
</tests>
```

Solution : Rechercher des exemples sur le sujet

Ex : <https://github.com/galaxyproject/tools-iuc/tree/master/tools>

Quelques problèmes rencontrés

group_humann2_uniref_abundances_to_GO :

Problème : L'outil n'accepte pas un paramètre pourtant correct

```
-p ${HUMANN2_DIR}
```

```
| Dataset Info:  
| discarding /home/cidam/conda/bin from PATH  
| prepending /tmp/tmpxjziTr/job_working_directory/000/5/conda-env/bin to PATH  
| Fatal error: Exit code 1 ()  
| Option -p requires an argument.
```

Solution : Se renseigner sur le paramètre créant l'erreur
(documentation de l'outil)

Quelques problèmes rencontrés

Problème : Des dépendances n'étant pas installées

Fréquemment, un outil a besoin de nombreuses dépendances pour fonctionner correctement et ces dépendances ne sont pas toujours correctement gérées.

Solution : Se renseigner sur les packages et leurs dépendances, documentation, puis installer ou mettre à jour ce qui est nécessaire

Quelques problèmes rencontrés

humann2_merge_abundance_tables :

Problème : Des outils mal documentés :


9. Merge abundance tables

```
$ humann2_merge_abundance_tables --input-genes $INPUT_GENES.tsv --input-pathways $INPUT_PATHWAYS.tsv --output $OUTPUT.tsv
```

- \$INPUT_GENES.tsv = a file containing the gene families (or EC) abundance table (tsv format)
- \$INPUT_PATHWAYS.tsv = a file containing the pathways abundance table (tsv format)
- \$OUTPUT.tsv = the file to write the new merged abundance table (tsv format)
- Optional: `--remove-taxonomy` remove the taxonomy from the output file

Solution : effectuer des recherches sur internet, documentation, Google Groups, etc. Et tester l'outil localement en «aveugle»

Après les tests?

 **Galaxy Tool Shed**

Repositories Groups Help User

3904 valid tools on May 15, 2016

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	74
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	40
Combinatorial Selections	Tools for combinatorial selection	6
Computational chemistry	Tools for use in computational chemistry	21
Convert Formats	Tools for converting data formats	64
Data Managers	Utilities for Managing Galaxy's built-in data cache	32
Data Source	Tools for retrieving data from external data sources	35
Epigenetics	Tools for analyzing Epigenetic/Epigenomic datasets	3
Fasta Manipulation	Tools for manipulating fasta data	76
Fastq Manipulation	Tools for manipulating fastq data	54
Genome-Wide Association Study	Utilities to support Genome-wide association studies	20
Genomic Interval Operations	Tools for operating on genomic intervals	42
Graphics	Tools producing images	42
Imaging	Utilities to support imaging	1
Metabolomics	Tools for use in the study of Metabolomics	25
Metagenomics	Tools enabling the study of metagenomes	58
Micro-array Analysis	Tools for performing micro-array analysis	8
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	105
Ontology Manipulation	Tools for manipulating ontologies	10
Phylogenetics	Tools for performing phylogenetic analysis	19
Proteomics	Tools enabling the study of proteins	78
RNA	Utilities for RNA	84
SAM	Tools for manipulating alignments in the SAM format	80

Conclusion

- Nombreux tests ajoutés permettant de finaliser de tester automatiquement et de publier ces outils sur le ToolShed.
- Apris :
 - Git et l'utilité d'un gestionnaire de versions
 - Github et le travail collaboratif
 - Test logiciel
 - Intégration continue et Travis CI
 - Prendre exemple sur les outils finalisés