

Use of Semantically Annotated Resources in the MobyLe2 Web Framework

Hervé Ménager, Bertrand Néron, Olivia Doppelt-Azeroual,
Olivier Sallou
MobyLe development team

BOSC 2014

Introduction - Mobylye

- A web-based workbench for bioinformatics (v1.5.3)

The screenshot displays the Mobylye web interface. At the top left, it says "Mobylye @Pasteur". On the right, there are user options: "(guest) set email | sign-in | activate | sign-out" and "refresh workspace". Below this is a navigation bar with tabs: "Welcome", "Forms", "Data Bookmarks", "Jobs", and "Tutorials". A search bar is located at the top left of the main content area. On the left side, there is a "Programs" sidebar with a tree view of various bioinformatics tools, including alignment, consensus, differences, formatter, multiple, display, hmm, information, clustalO-multialign, clustalO-profile, clustalO-sequence, clustalw-multialign, clustalw-profile, clustalw-sequence, comalign, dca, dialign, edialign, extend_align, mafft, msa, msaprops, muscle, pima, plotcon, T-Coffee@rpbs, tranalign, pairwise, structure, assembly, database, genetics, and hmm. The main content area shows the "mafft" program selected. It has a title "mafft 6.849" and buttons for "Run" and "Reset". Below the title is a description: "Multiple alignment program for amino acid or nucleotide sequences." and a button for "only simple options". The "Input Options" section includes a "Sequences File" dropdown with "paste", "db", and "upload" options, and an "EDIT" button. Below this is a text area labeled "Enter your data below:". The "Sequences type" is set to "Automatic". The "Use structural alignment(s)" section has a "Structural alignment 1" dropdown with "paste" and "upload" options, and an "EDIT" button. Below this is another text area labeled "Enter your data below:".

- adapted to the needs of biologists, **sometimes very punctual.**
- also provides a **persistent workspace if needed.**

Mobyle 2

- Major rewrite of Mobyle
- Currently in development
- New features:
 - groupware functionalities
 - secure data sharing
 - REST api
 - *ontology based service descriptions*

Resource description

Classification

Parameters typing

```
<program>
  <head>
    <name>mafft</name>
    <version>6.849</version>
    <doc>
      <title>mafft</title>
      <description>
        <text lang="en">Multiple alignment program for amino acid or nu
      </description>
      <doclink>http://mafft.cbrc.jp/alignment/software/about.html</doclink>
    </doc>
    <category>alignment:multiple</category>
  </head>
  <parameters>
    <paragraph>
      <name>input_opt</name>
      <prompt lang="en">Input Options</prompt>
    </parameters>

    <parameter ismandatory="1" issimple="1">
      <name>sequences</name>
      <prompt lang="en">Sequences File ( a file containing several
      <type>
        <datatype>
          <class>Sequence</class>
        </datatype>
        <dataFormat>FASTA</dataFormat>
      </type>
      <format>
        <code proglang="perl">" $sequences"</code>
        <code proqlang="python">" " + str( sequences )</code>
      </format>
    </parameter>
  </parameters>
</program>
```

Discovery and integration

Services classification

[\[more\]](#)

Classify: by category by package
Workflows and programs: separate mixed

Programs

- alignment
 - consensus
 - differences
 - formatter
 - squizz_checker
 - squizz_convert
 - multiple
 - pairwise
 - dot_plots
 - global
 - local
 - bl2seq
 - combat
 - treealign
 - wise2
 - structure
- assembly
- database
- display
- genetics
- hmm
- nucleic
- phylogeny
- protein
- sequence
- structure

Workflows

- align-family
- hmm_build_search
- protein_distance_phylogeny

Guided chainings

results

Alignment
muscle.out

```
>tr|Q1KLC8|Q1KLC8_THEFU Alpha-amylase - Thermomonospora fusca.  
-MGVRRSLAALLAALLGCATSLVALTVA-----  
-A---SPAHAAPSGNRDVI VHLFQWR-----WKSIADECRITLGPHGFGAVQVSPPEH  
VVL-----PAEDYP-----WNQDYQPVS YKLDQTRRGSRAD  
FID--MVNTCREAGVKIYDAVINHMT-----GTGS-----AGAGPGSAGSSYSKY  
DYPGIYQSQDFNDCRRDI---TNMNDKWEVQHCELVGLADLKTSSPYVQDRIAA-YLNEL  
IDL-GVAGFRIDAANKH----IPEGDLQA----ILSRLKNVHPAWGGKPYIFQEVIADS  
TISTGSYTHLGSVTEFQYHR---DISHAFANGIAHLTGLGSLTP-----SDKAVV--F  
VVAL--LPTDQVDFDLTDFDAP--VPLAOKFMLALDYCTDQVMSQVDM
```

as muscle.out or

- parameters
- job execution

- showalign (sequence)
- jalview-print (alignment)
- mview (alig)
- BMGE (infile)
- muscle (profile2)
- muscle (profile1)
- quicktree (aligfile)
- protpars (infile)
- distmat (sequence)

tr|Protein Sequence Parcimony Method - Alianmer

Format detection/conversion

parameters

Blast program (-p) ? blast

- Database
 - Protein db (-d) ? uniprot_sprot
- Query Sequence
 - Query (-i) (Sequence) ?
 - abcd3_human.sp (SWISSPROT)
 - reformatted file produced by squizz: abcd3_human.fasta (FASTA)
- Selectivity options

Typing system

- datatype: the “kind” of information stored in the data
 - a datatype is either a reference to a python class
 - e.g.: **Sequence** -> “sequence.py”
 - or a “dynamic” subclass of one of these python classes
 - e.g.: **BlastReport**, subclass of **Report** -> “report.py”
- data format: the format used to store the data
 - e.g.: **FASTA**, **SWISSPROT**, etc.

Shortcomings of Moby1 classification and typing system

- freely extendable
 - highly sensitive to annotation errors (PDB or Pdb?),
 - potentially inconsistent between multiple MobyNet sites,
- confusions between datatypes / data formats
- Moby-specific mechanism
 - hard to share/reuse these annotations

Solution: use of an ontology

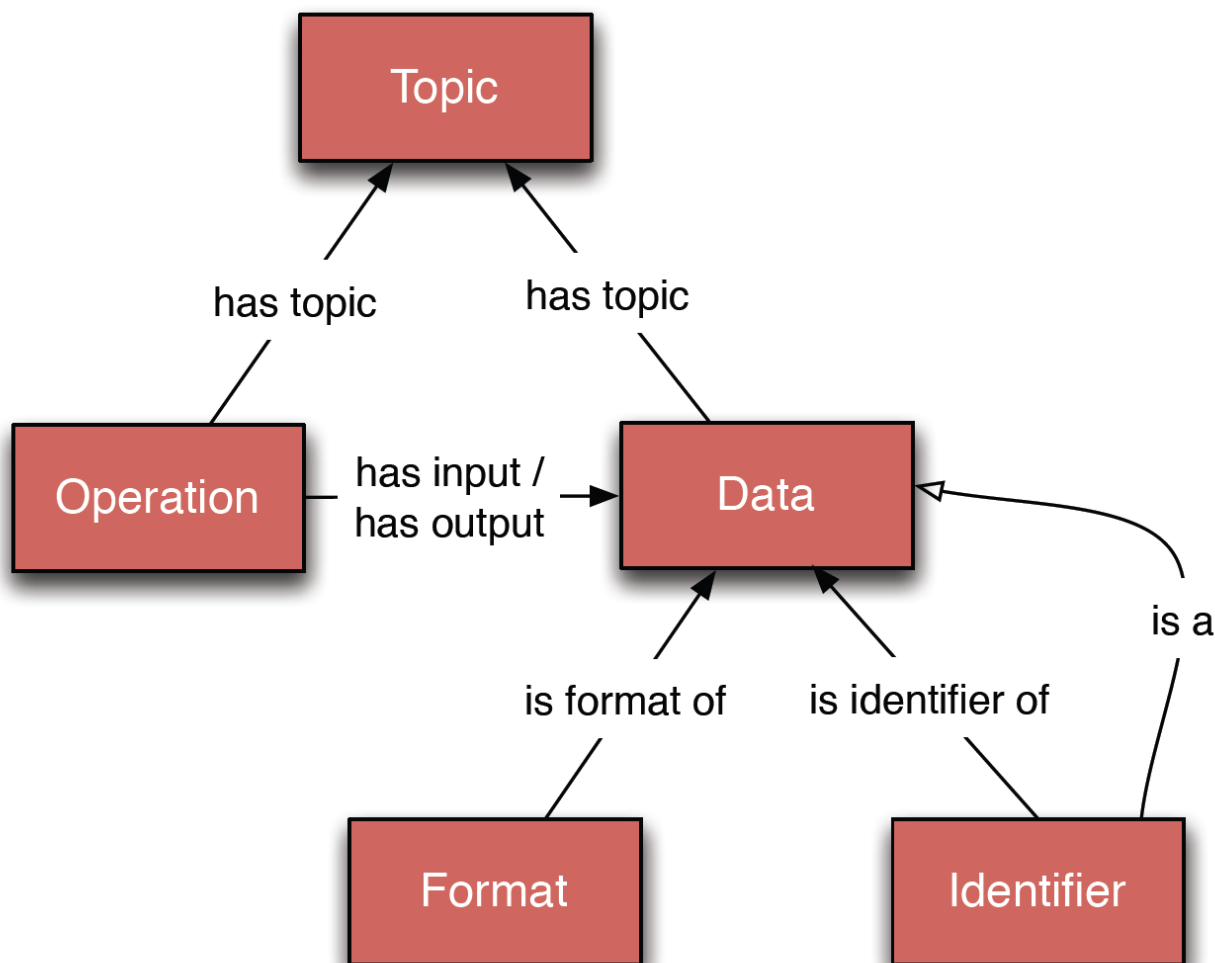
General Properties

- controlled vocabulary,
- describing the relationships between the different terms.

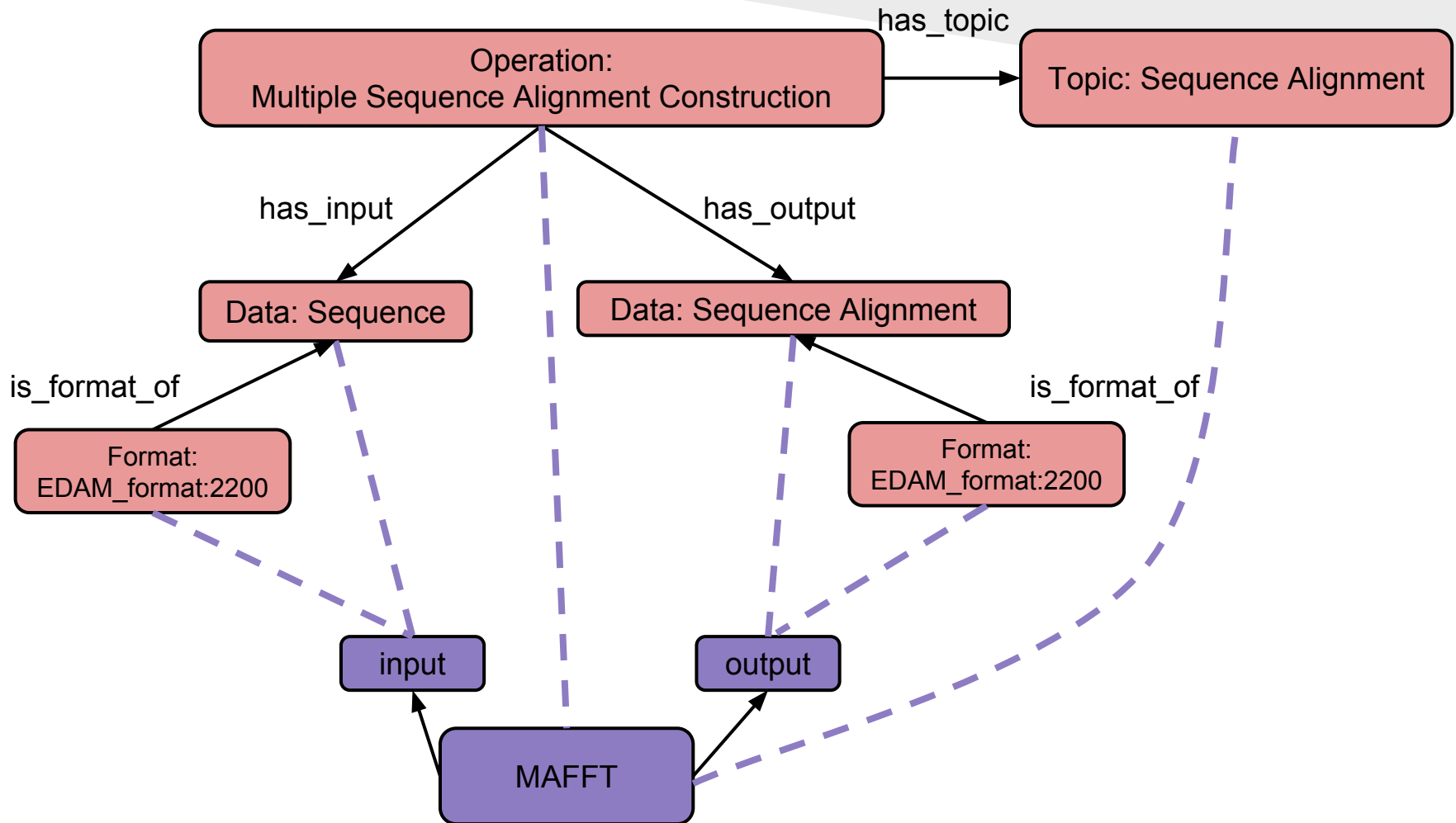
Scope

- for the data and parameters:
 - the type of information described,
 - the format used to store it.
- for the services classification:
 - the hierarchy of categories to store the services.

EMBRACE Data And Methods ontology



Service annotation: a simple example



Services classification

- Structure analysis
 - Structure prediction
 - Protein secondary structure prediction
 - garnier
 - helixturnhelix
 - pepcoil
 - + Nucleic acid structure prediction
 - Nucleic acid structure analysis
 - + Nucleic acid structure prediction
 - banana
 - btwisted
 - RNA structure and alignment
 - rnaifold
 - rnaplot 1.8.4
 - + Protein structure analysis
 - + Structure databases
- Phylogenetics
 - kitsch
 - neighbor
 - clique
 - distmat
 - weighbor 1.2.1
- Data handling

- **Usage of EDAM topics hierarchy**

Guided interactive chainings

```
>FR775944|FR775944.1-Lactobacillus-crispatus-CIP-102990T  
cttacggaagtttggtgaaccactgagcaggcttggtact  
acttaggaattcttaagaagagtattactacttctattccaggactggactggt  
tatgtgcaaattgatggcgtttacacgagttctcaacagttcctggcgtcagagaagat  
gtaaccaagatcatctgaactgaagaaactgaattaaagtctcttcagatgaacaaa  
aggtattgaattggacattgaaggccagccacagtaactgctgatgatcttaaag >>
```

Sequence Profile Generation

>> hmmbuild

Sequence Alignment Rendering

>> boxshade

Sequence Alignment Reformatting

>> mview

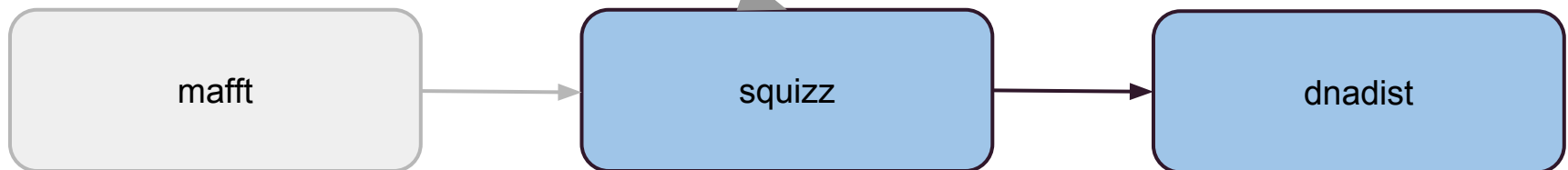
>> squizz

Automatic conversions

```
>FR775944|FR775944.1-Lactobacillus-crispatus-CIP-102990T  
-----cttacggaagttgttgaaccactgagcaggcttggact  
acttaggaattcttaagaagagtattactacttctattccaggactggactgtt  
tatgtgcaaattgatggcgtttacacgagttctcaacagttctggcgtcagagaagat  
gtaaccaagatcatctgaactgaagaaactgaattaaagtctcttcagatgaacaa  
aaggttattgaattggacattgaaggccagccacagtaactgctgatgatcttaaa >>
```

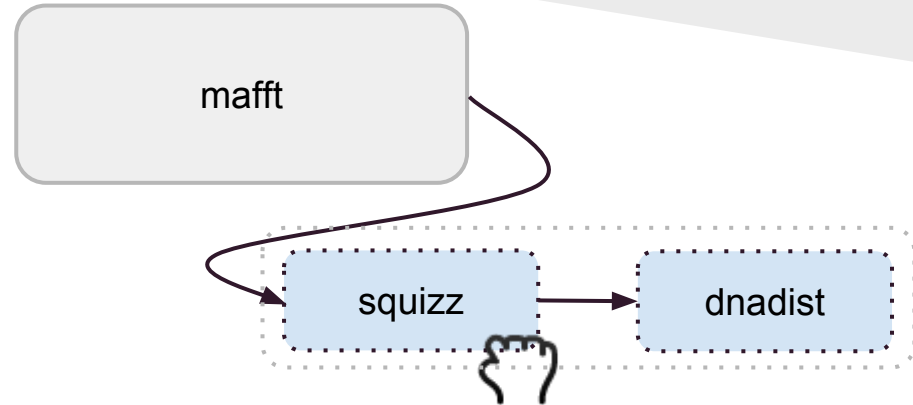
```
Sequence Profile Generation  
>> hmmbuild  
Sequence Alignment Rendering  
>> boxshade  
Sequence Alignment Reformatting  
>> mview  
>> squizz  
Sequence Distance Matrix  
Generation  
>>dnadist
```

EDAM_operation:0335 (File reformatting)

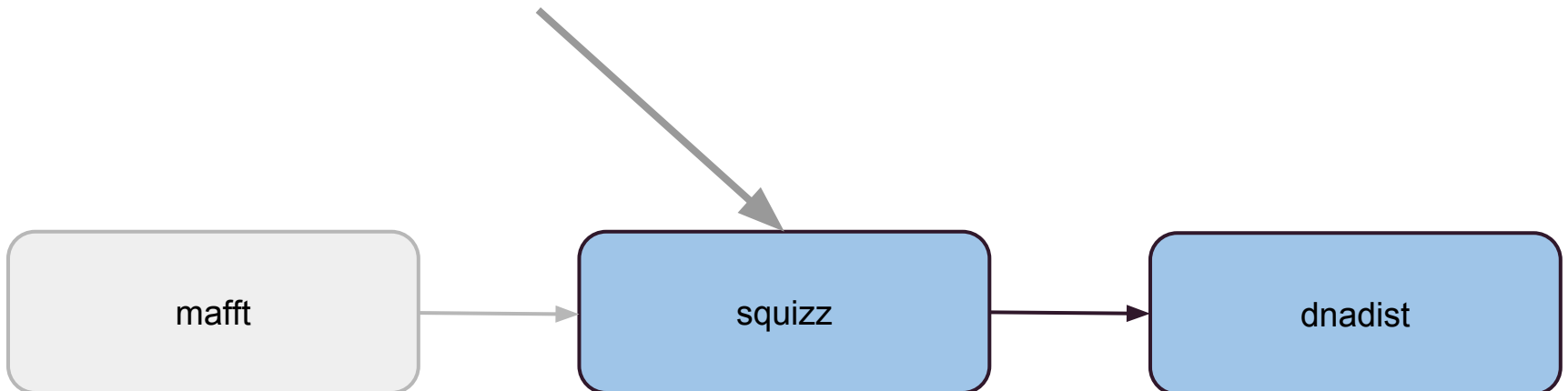


Automatic conversions in workflows

```
Sequence Profile Generation
>> hmmbuild
Sequence Alignment Rendering
>> boxshade
Sequence Alignment Reformatting
>> mview
>> squizz
Sequence Distance Matrix
Generation
>> dnadist
```



EDAM_operation:0335 (File reformatting)

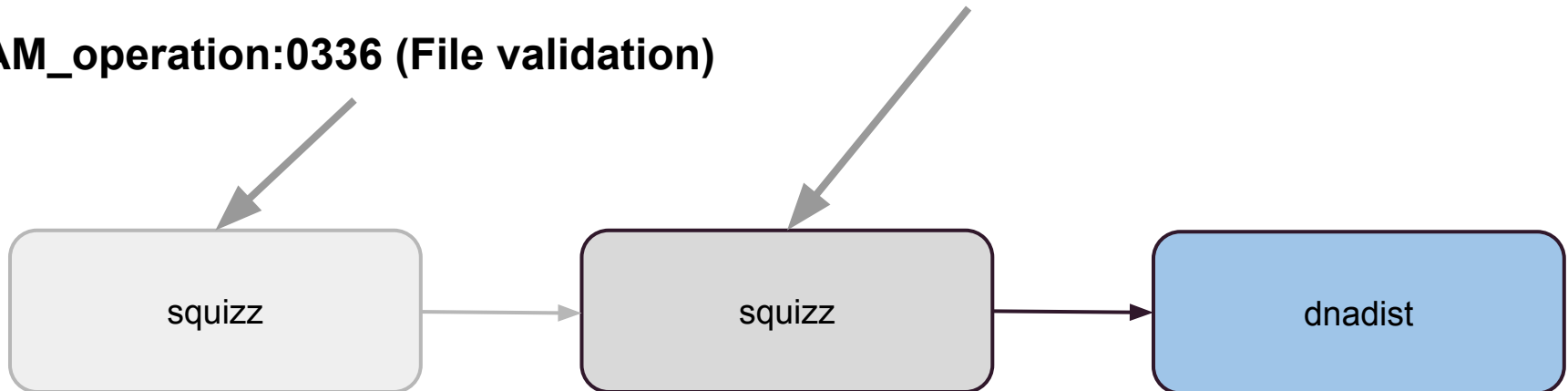


Automatic format detection

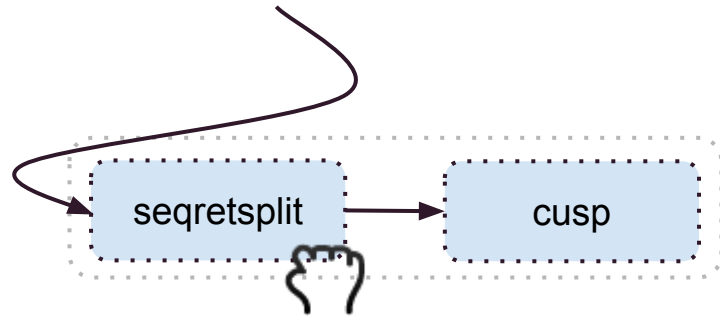
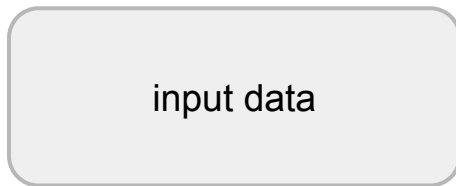
- **if the format of a data item is unknown (e.g., directly uploaded by the user)**
- **detection of the data format, conversion if necessary**

EDAM_operation:0335 (File reformatting)

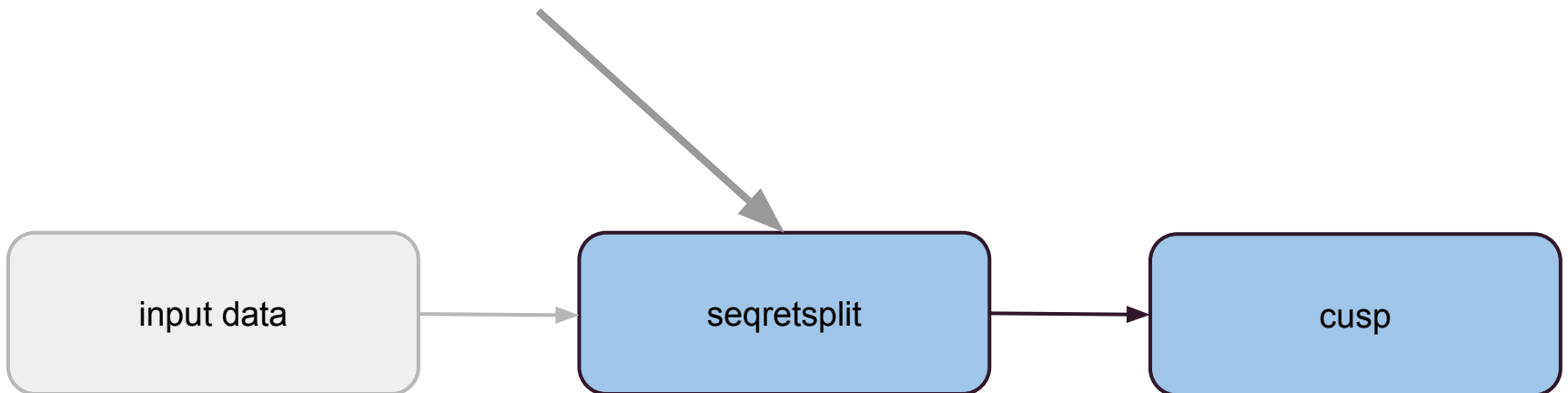
EDAM_operation:0336 (File validation)



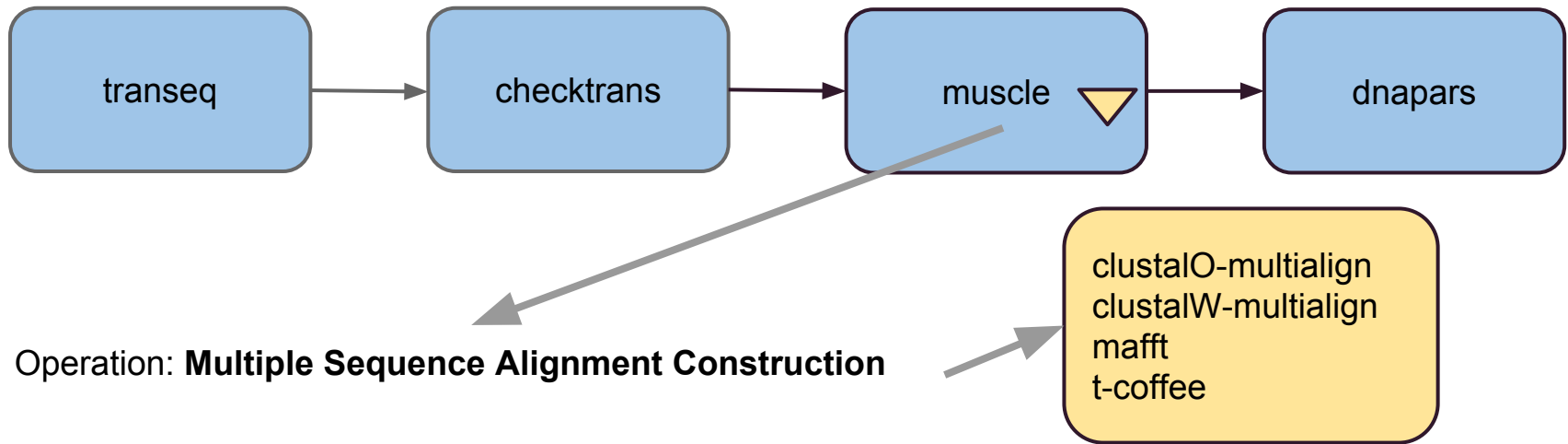
Automatic split and iteration



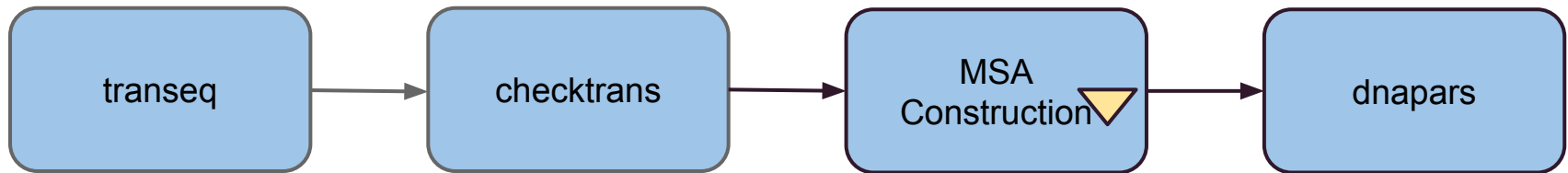
(Sequence splitting)



Equivalent methods



Abstract methods



Operation: **Multiple Sequence Alignment Construction**

MSA Construction Method:

clustalw-multialign



clustalO-multialign

mafft

muscle

t-coffee

Conclusion

- Moby 2 in development stage
- Ontologies are a useful resource for the integration of bioinformatics resources:
 - “service discovery”
 - implementation of high-level services
 - sharing of resource descriptions with other projects
- Remaining challenges:
 - integration in “user-friendly” interfaces
 - quality of the “annotations” in the service descriptions

Acknowledgements

- Mobylye and Mobylye2 development teams,
- our users,
- EDAM and Tools Registry teams: Jon Ison, Matus Kalas, Kristoffer Rapacki, etc.
- **MASTODONS** funding