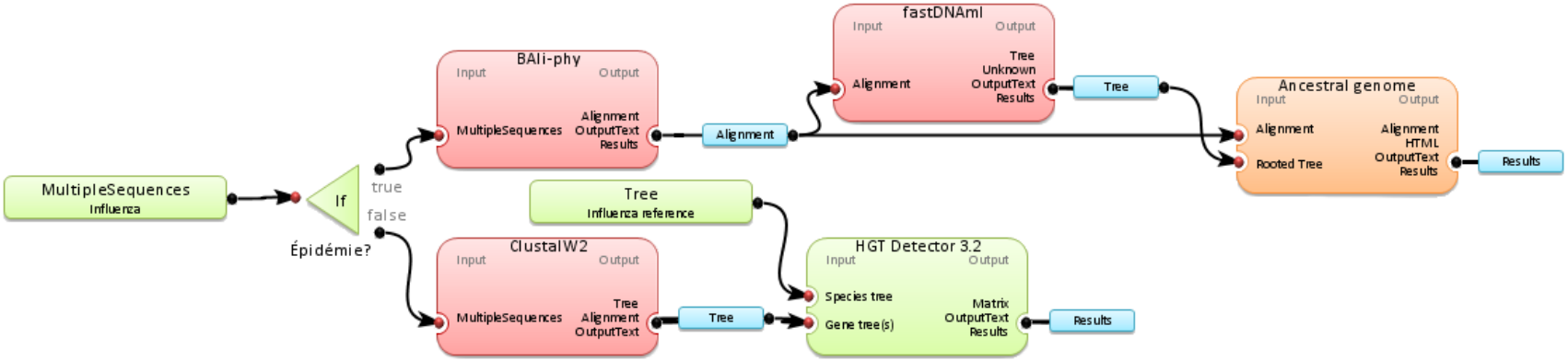
Application des flux de travaux en bioinformatique



Etienne Lord, *Ph.D.*

Chercheur et développeur chez Agriculture Canada

20 février 2018

Laboratoire Bioinformatique UQÀM

UQÀM

Université Laval

Sommaire

- **L'évolution de la bioinformatique**
- **Les flux de travaux / flux de données**
  - Nomenclature des flux de données
  - Systèmes de gestions de flux de travaux
- **Les systèmes de gestions de flux de travaux en bioinformatique**
  - Galaxy
  - Taverna
  - Armadillo

  } Et quelques autres…
- **La plate-forme *Armadillo* orientée vers la phylogénétique**
  - Cas d'utilisations
  - Orientation dans le développement et défis
- **Références**

Évolution des techniques en biologie (problématique 1)

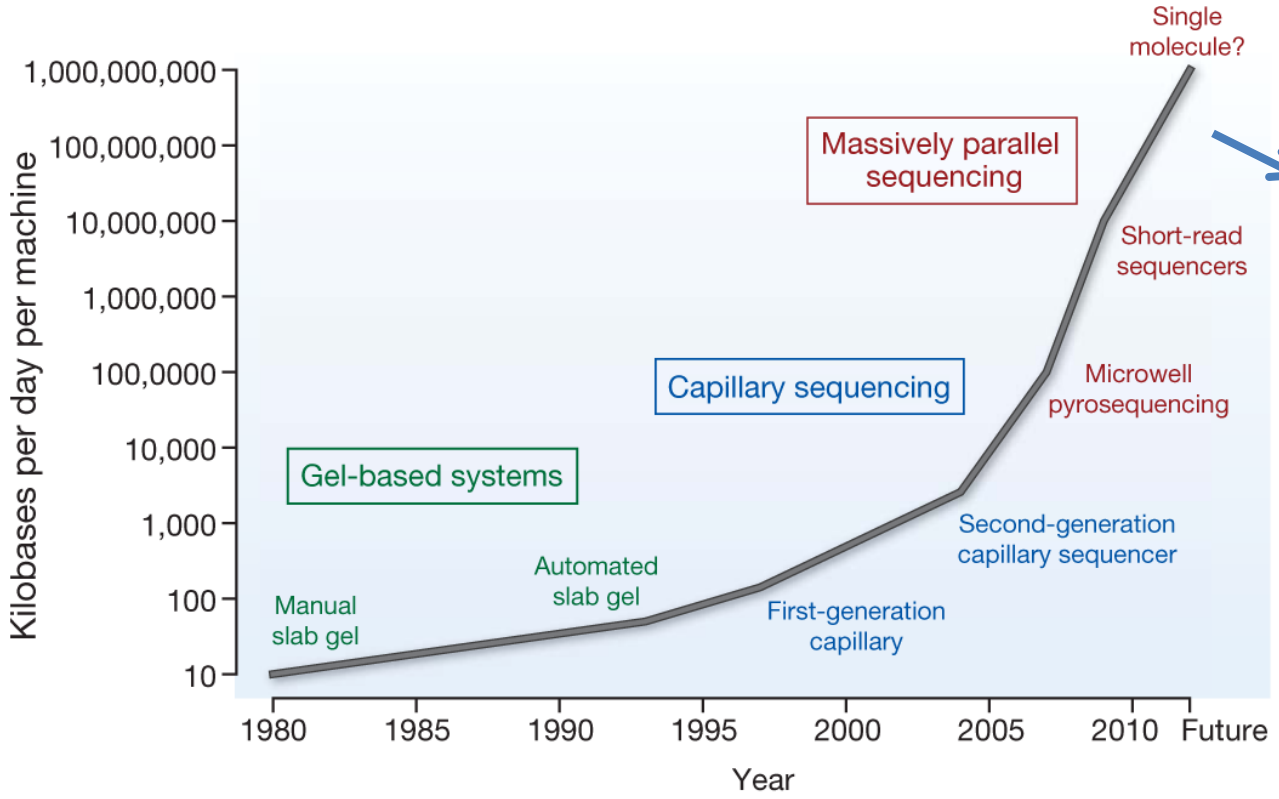
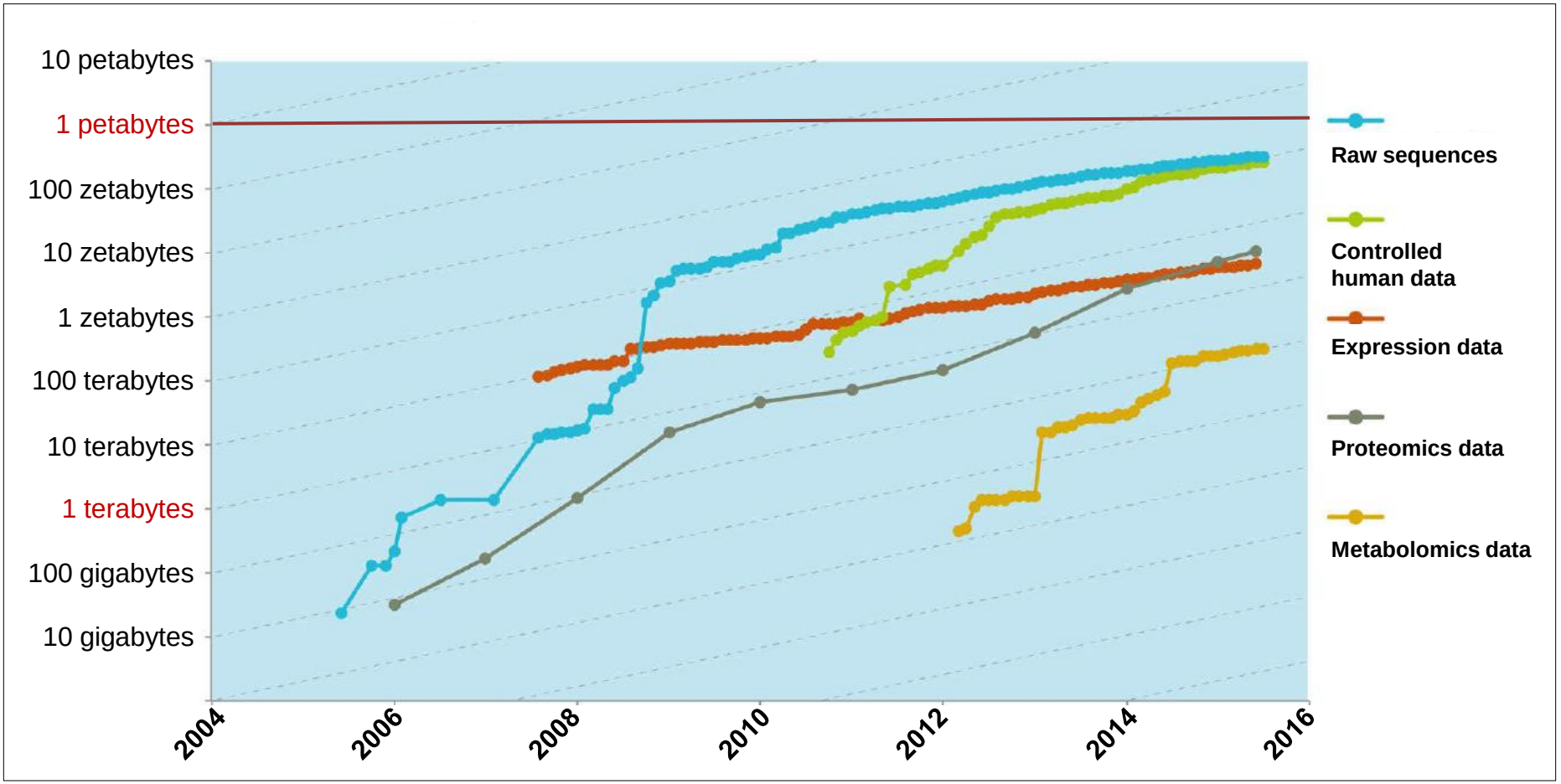


Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

Séquenceur *MinION* de Oxford Nanopore publications.nanoporetech.com/

Stratton MR, Campbell PJ, Futreal PA. (2009) The cancer genome. Nature. 458(7239):719-24.

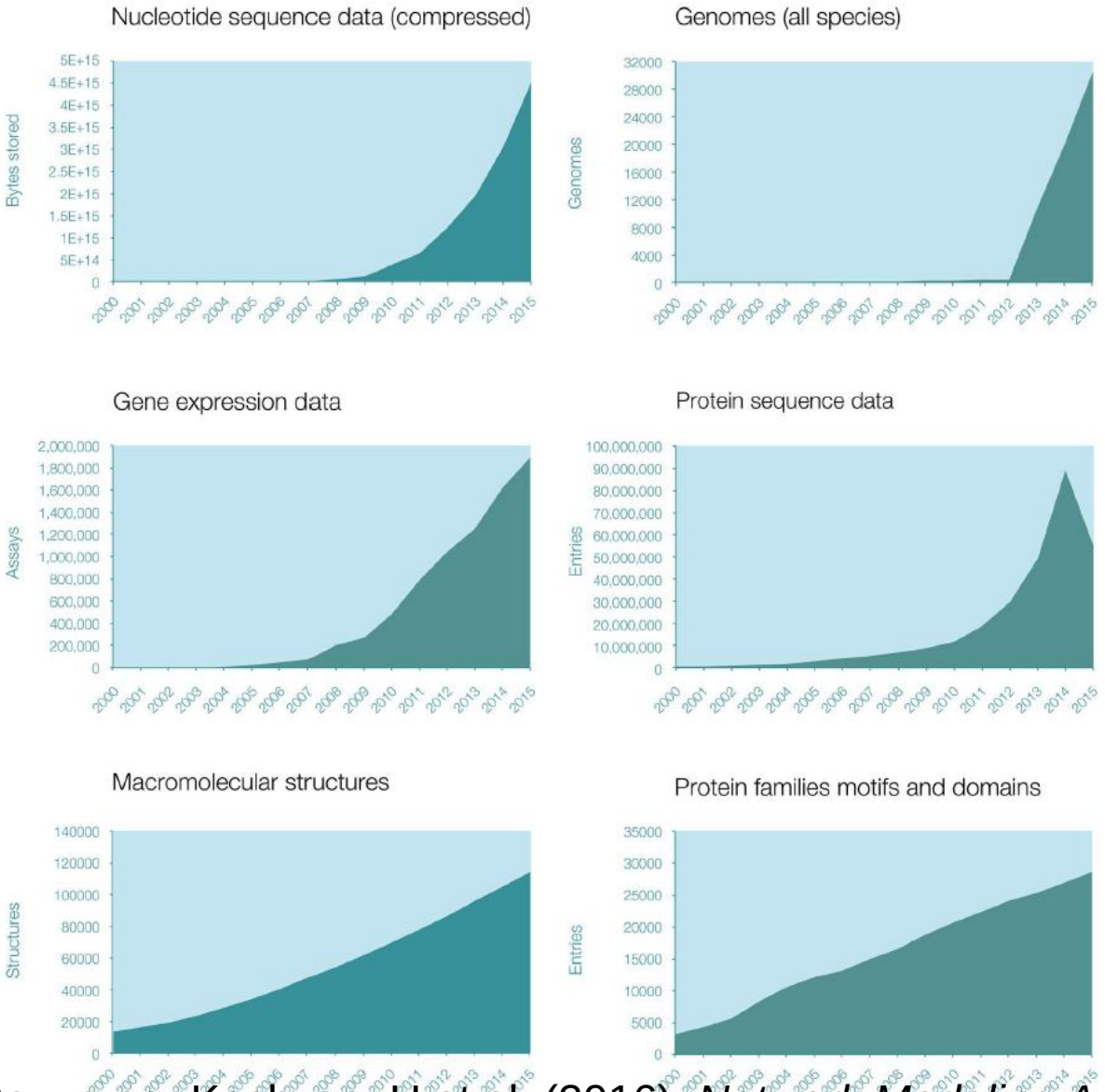
Évolution de la quantité des données (problématique 2)



EMBL-European Bioinformatics Institute (2016) EMBL-EBI annual scientific report 2015

http://www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/external_relations/Documents/EMBL_EBI_ASR_2015_DigitalEdition23Jul2016.pdf

Évolution des types de données (problématique 3)



EMBL-European Bioinformatics Institute (2016) EMBL-EBI annual scientific report 2015

Types de données

1. **Expression** (gènes, protéines,microarray)
2. **Séquences** (ADN, ARN, protéines).
3. **Stuctures** (2D, 3D)
4. **Interactions** (protéines-protein interactions)
5. **Réseaux** (pathways) et **ontologies**
6. **Raw data**? (Geo, etc.)

Revue par Kashyap, H et al. (2016). *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 28.

Une myriade de formats (problématique 4)

BioPortal Browse Search Mappings Recommender Annotator Resource Index Projects

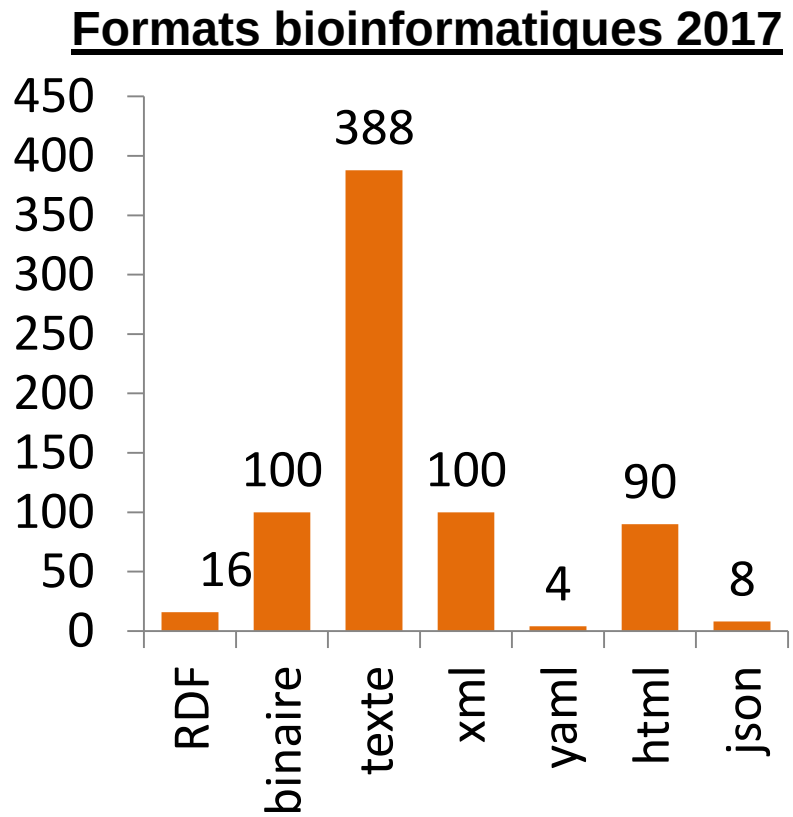
EDAM bioinformatics operations, data types, formats, identifiers and topics

Summary Classes Properties Notes Mappings Widgets

Jump To:

- Data
- DeprecatedClass
- Format
  - Binary format
  - **Format (by type of data)**
    - Alignment format
    - Annotated text format
    - Article format
    - Bibliographic reference format
    - Biodiversity data format
    - Biological pathway or network format
    - Biological pathway or network report format
    - Chemical formula format
    - Data index format
    - Database hits (sequence) format
    - Dirichlet distribution format
    - Document format
    - Enzyme kinetics report format
    - Experiment annotation format
    - Gene annotation format
    - Gene expression report format
    - Graph format
    - HMM emission and transition counts format
    - Image format
    - Phylogenetic tree report (tree distances) format
    - Protein domain classification format

Details Visualization Notes (0) Class Mappings (1)



protégé

Ontology viewer
protege.stanford.edu/

http://bioportal.bioontology.org/ontologies/EDAM?p=classes

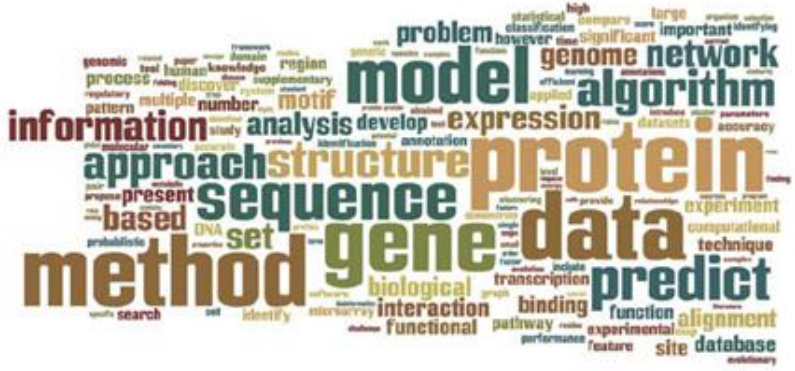
http://edamontology.org/page

L'évolution de la bioinformatique

Les **méthodes**, les **réseaux** et les **données** deviennent de plus en plus importantes dans la recherche.

1993-1997

1998-2002



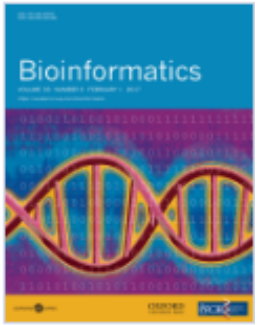
2003-2007

2008-2012

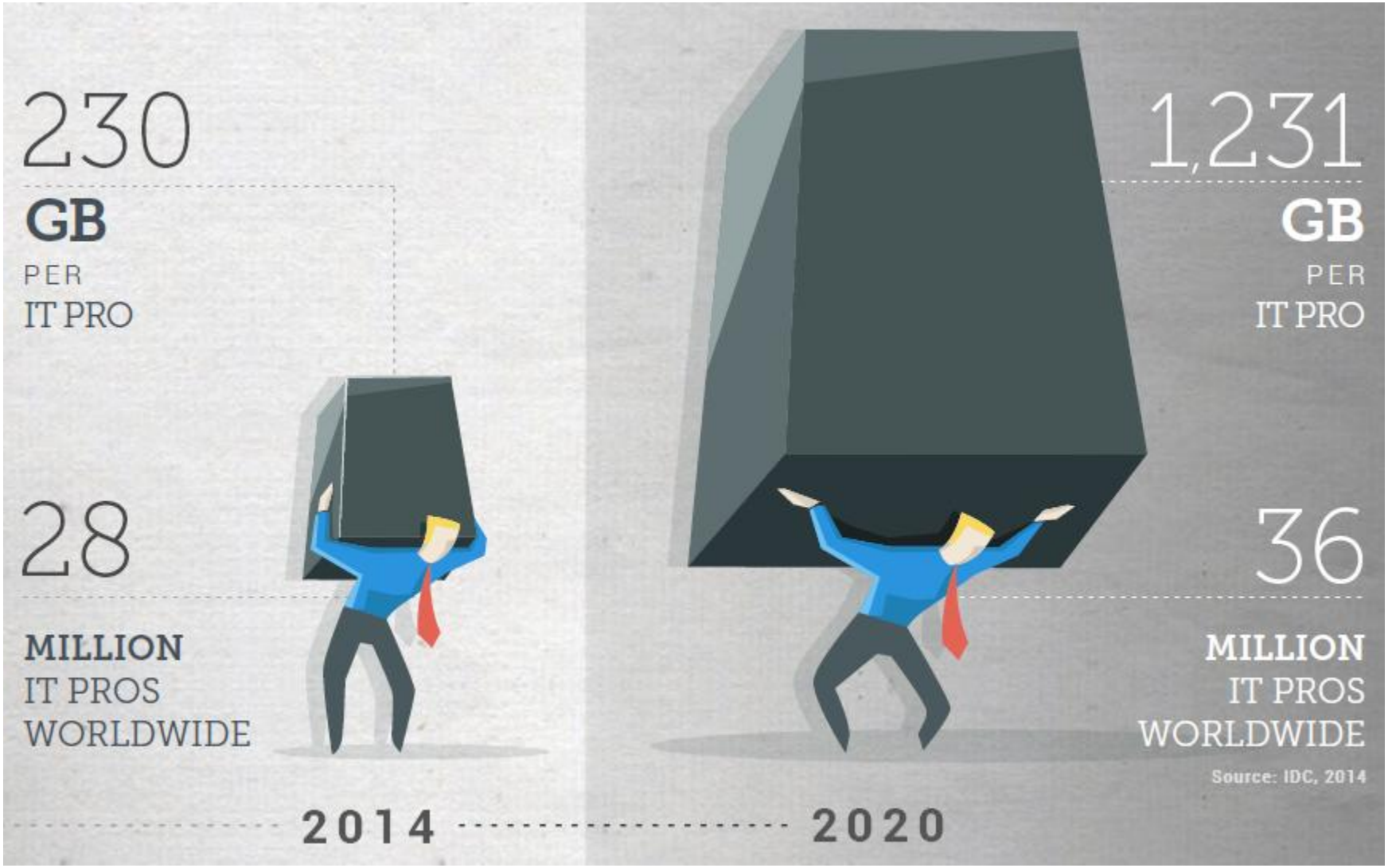
Gibson TA. (2012) The roots of bioinformatics in ISMB. PLoS Comput Biol. 8(8):e1002679.

Bioinformatique 2013-2018

2013-2018

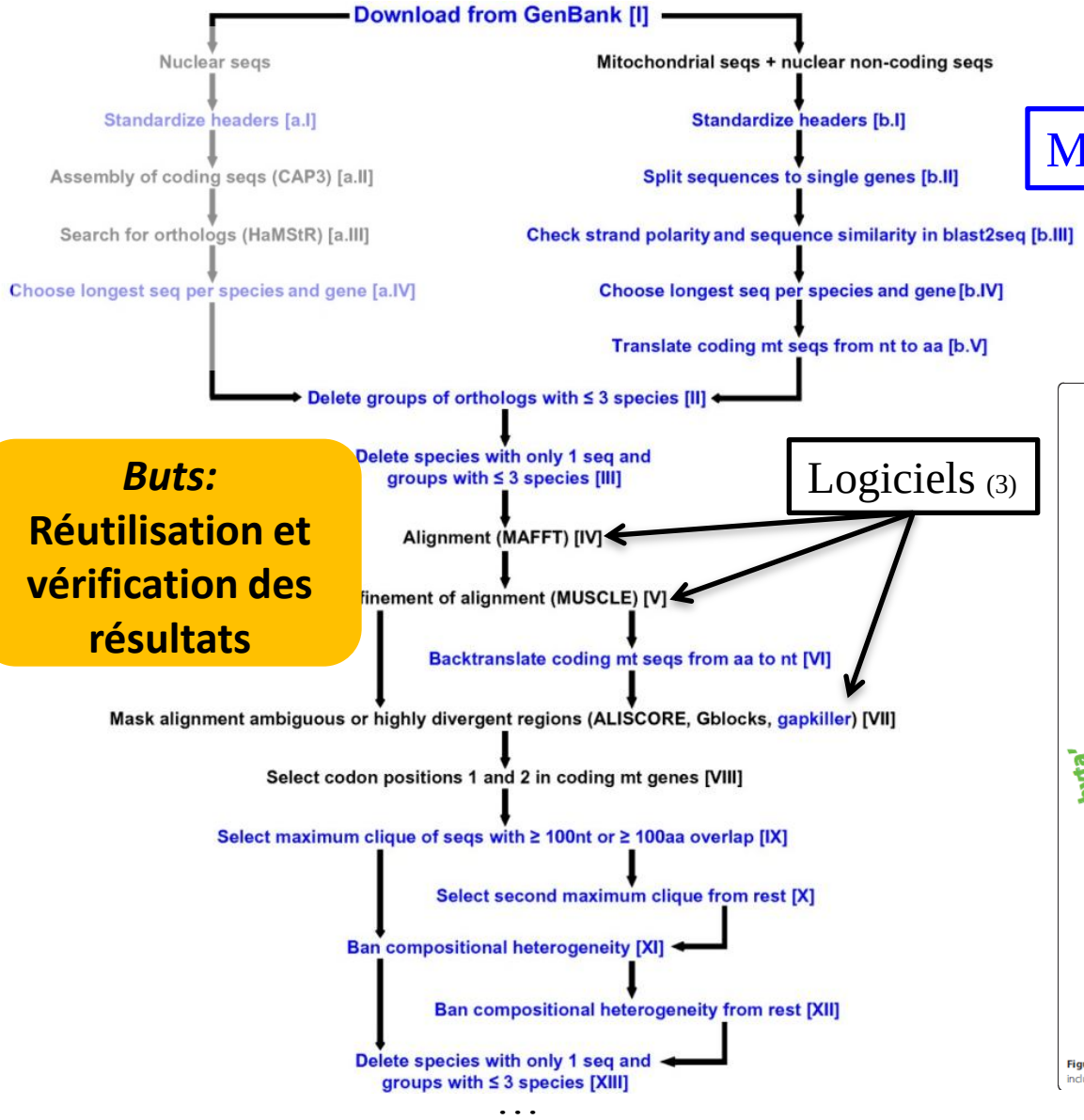


Comment s'en sortir?



Turner V et al. (2014) The digital universe of opportunities:rich data and the increasing value of the internet of things. International Data Corporation, White Paper, IDC_1672

Exemple de flux de travaux en analyse phylogénétique



Méthodes (13)

Logiciels (3)

Buts:
Réutilisation et vérification des résultats

120 000 séquences

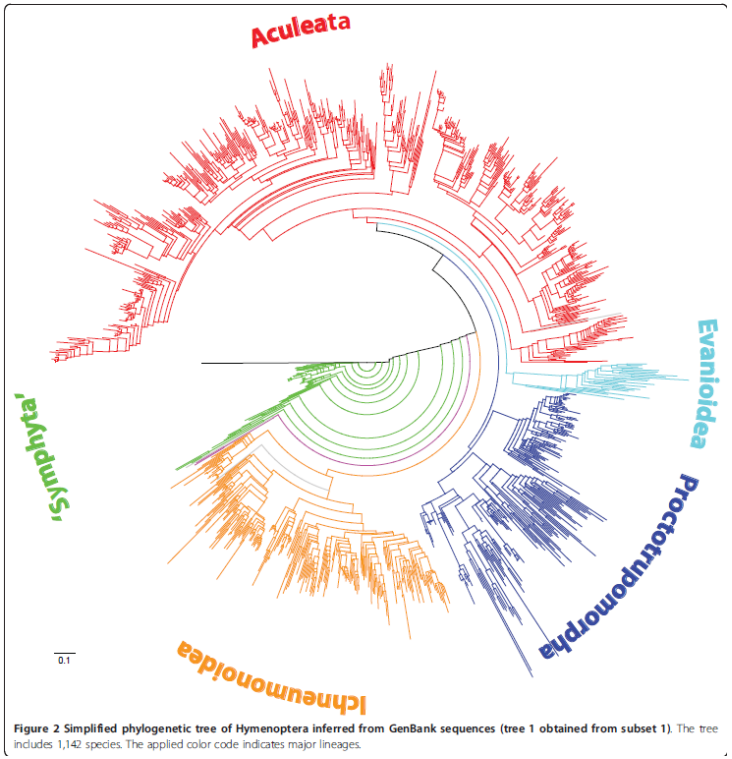


Figure 2 Simplified phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 1 obtained from subset 1). The tree includes 1,142 species. The applied color code indicates major lineages.

Une reconstruction phylogénétique par Peters *et al.* (2011) *BMC Biology* 9, 55.

Exemple d'application en 2015

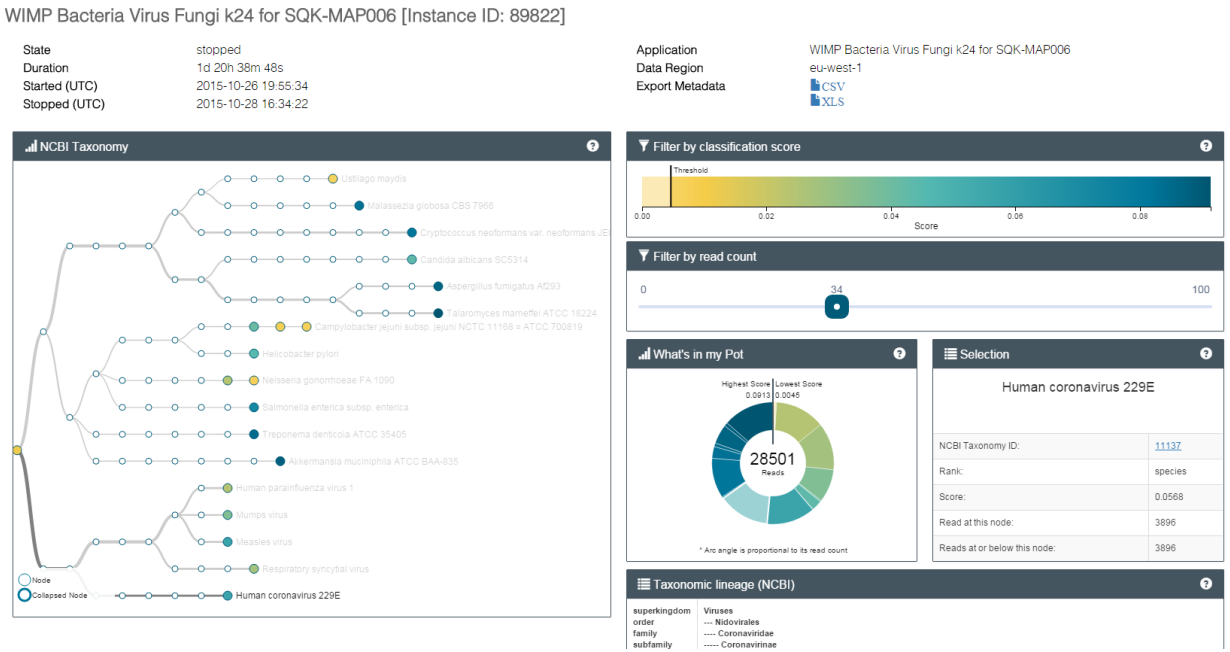
What's in my pot? Real-time species identification on the MinION

BioArxiv doi: http://dx.doi.org/10.1101/030742

Sissel Juul, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett & Daniel J Turner

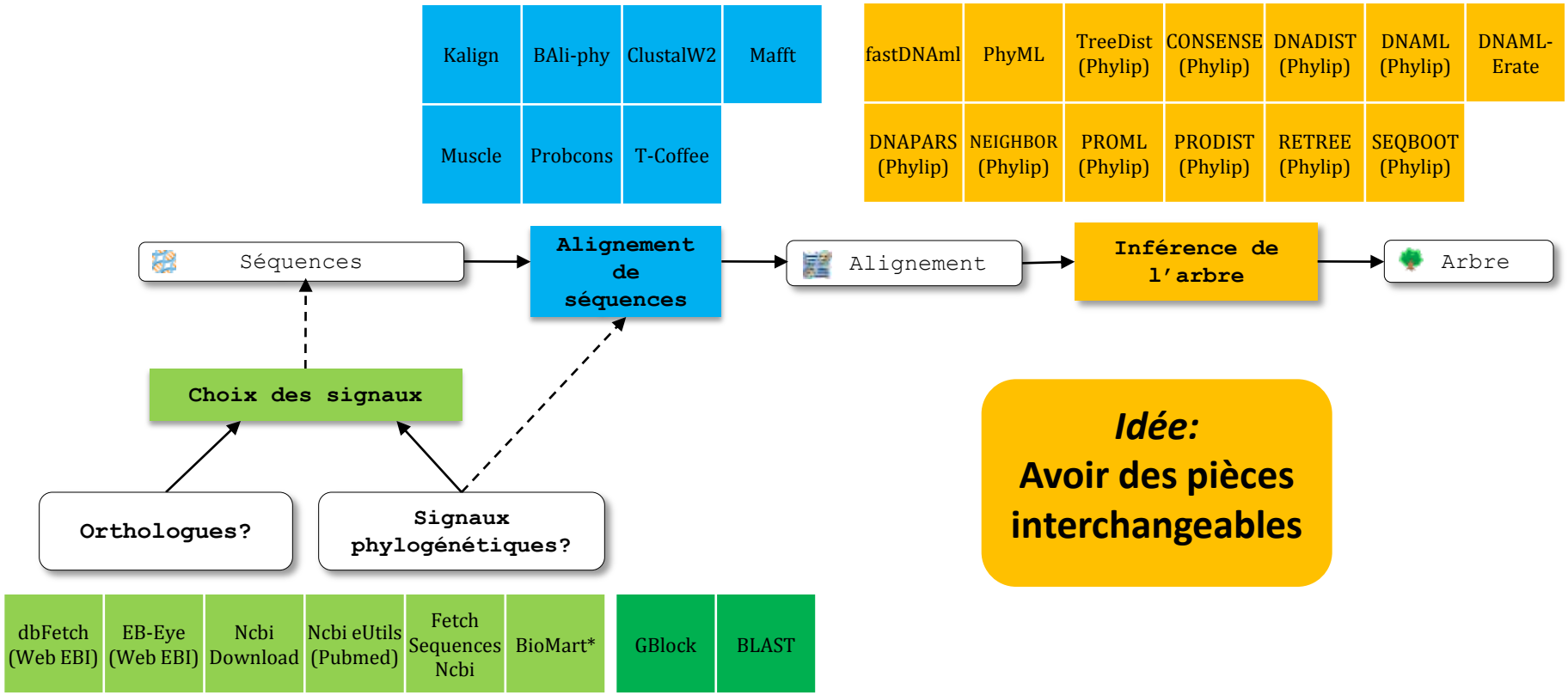
Abstract

Whole genome sequencing on next-generation instruments provides an unbiased way to identify the organisms present in complex metagenomic samples. However, the time-to-result can be protracted because of fixed-time sequencing runs and cumbersome bioinformatics workflows. This limits the utility of the approach in settings where rapid species identification is crucial, such as in the quality control of food-chain components, or in during an outbreak of an infectious disease. Here we present What's in my Pot? (WIMP), a laboratory and analysis workflow in which, starting with an unprocessed sample, sequence data is generated and bacteria, viruses and fungi present in the sample are classified to subspecies and strain level in a quantitative manner, without prior knowledge of the sample composition, in approximately 3.5 hours. This workflow relies on the combination of Oxford Nanopore Technologies' MinION ™ sensing device with a real-time species identification bioinformatics application.



Flux de travaux: définition générale d'un workflow

Un flux de travaux est un **patron de tâches ordonnées** pouvant être **exécutées de manière répétitive**.



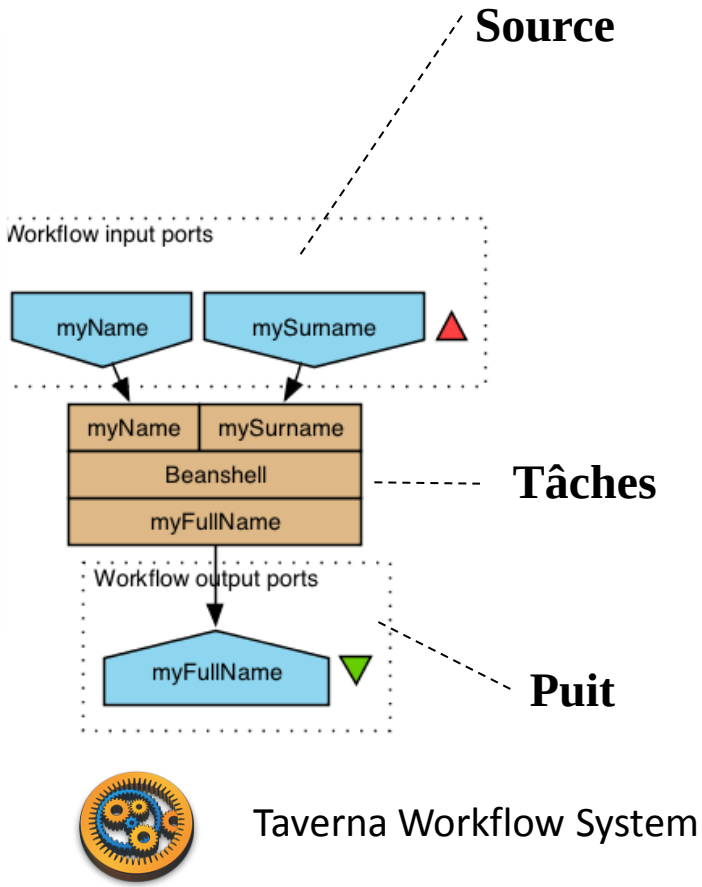
Idée:
Avoir des pièces interchangeables

Schématisation de l'inférence d'un arbre phylogénétique.
Voir pour revue Philippe *et al.* (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602

Composantes des flux de travaux et avantages

```
public MultipleSequences Convert_to_Protein(MultipleSequences multi, boolean remove_stop, int code_number) {
        Config.log("Convert_to_Protein()");
            Translator translator=new Translator();
            GeneticCode code=translator.code.get(code_number);
           MultipleSequences newmulti=new MultipleSequences();
           newmulti.setName("Protein from "+multi.getName());
           newmulti.setNote("Convertion of MultipleSequences to Protein at "+Util.returnCurrentDateAndTime());
           for (biologic.Sequence s:multi.getSequences()) {
               String aa="";
               if (remove_stop)  {
                    aa=translator.translate_without_stop(s, code);
               } else {
                   aa=translator.translate(s, code);
               }
               biologic.Sequence s2=s.clone();
               s2.setId(0);
               s2.setSequence(aa);
               s2.setSequence_type("aa");
               if (s2.getLen()==0) {
                   Config.log("Unable to find start codon for "+s.getName());
                   return multi;
               }
               Config.log(s2.getName()+"\t"+s2.getLen()+" aa ");
               newmulti.add(s2);
           }
        Config.log("Finish...");
       return newmulti;
}
```

Code source



Taverna Workflow System

+ Portabilité (pas de recompilation)
+ Automatisation
+Réutilisation des composants
+ *Scale-up* facilité
+ Simplification de la cognition

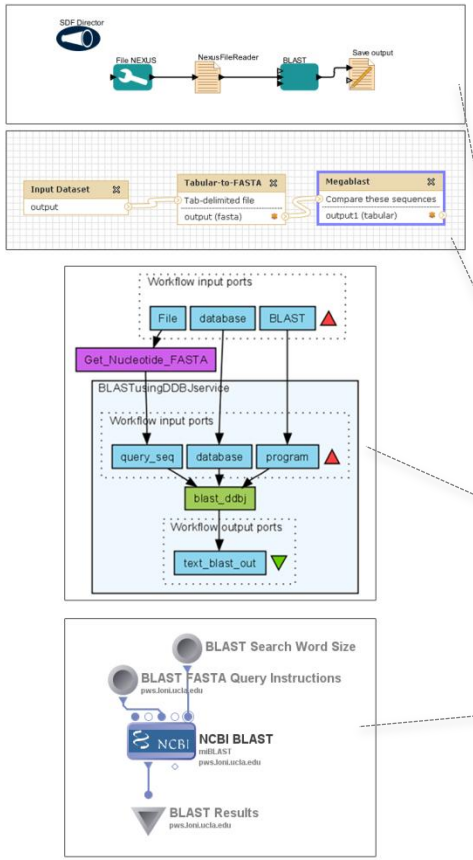
Caractéristiques des systèmes de gestion de workflows (WfMS)

- Permet *l'**interaction avec l'utilisateur*** (suggestion avec ontologie).
- Supporte la ***reproductibilité*** dans les résultats.
- Intègre des services (*e.g.* services Web) et de ***logiciels hétérogènes***.
- Supporte la ***distribution et la gestion de données*** hétérogènes.
- Support pour les ***plate-formes hautes performances***.
- Permet la ***surveillance*** pendant le fonctionnement et permet la ***reprise*** même dans le cas d'erreurs.
- Permet une ***interconnexion avec d'autres systèmes*** (*e.g.* de flux de travaux, d'ordinateurs en réseau)

Lin, C., Lu, S., Fei, X., Chebotko, A., Pai, D., Lai, Z., Fotouhi, F., et Hua, J. (2009). A reference architecture for scientific workflow management systems and the VIEW SOA solution. *IEEE Transactions on Services Computing*, *2*(1), 79-92.

Woollard, D., Medvidovic, N., Gil, Y., et Mattmann, C. A. (2008). Scientific Software as Workflows: From Discovery to Distribution. IEEE Software, 25(4), 37-43.

Les plates-formes de flux de travaux en bioinformatique

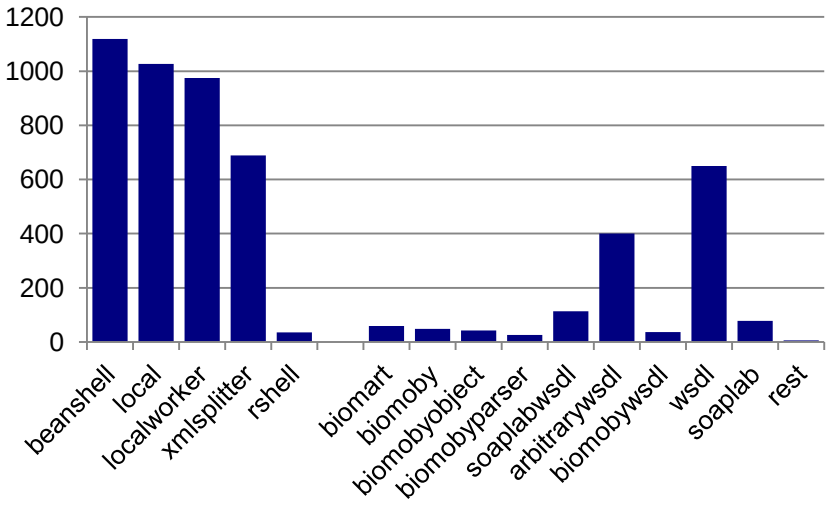


Plates-formes de flux de travaux bioinformatiques

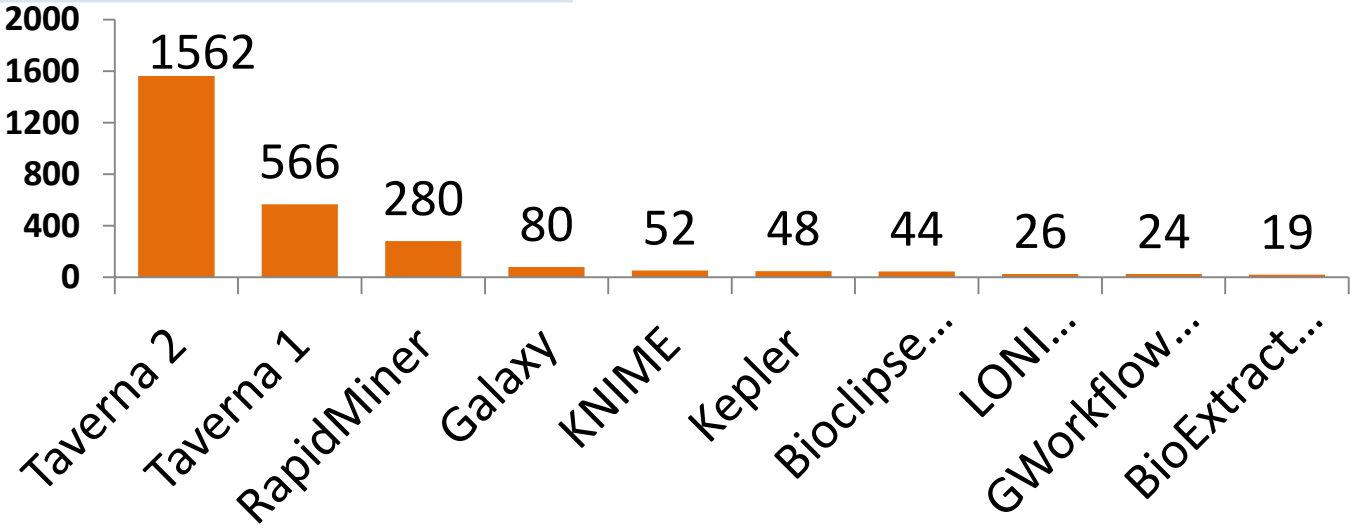
Plate-forme	Classe	Flux de travaux	Accès au « nuage »	Application
Kepler (2004)	**Locale**	*Data-flow*	Oui (Ecogrid)	Général
Galaxy (2005)	**Web/Locale**	*Data-flow*	Oui (Amazon EC2)	Bioinformatique
Taverna (2003)	**Locale, services Web**	*Data-flow*	Oui (myGRID)	Bioinformatique
LONI (2003)	**Locale, services Web**	*Data-flow*	Oui	Bioinformatique
Bio-Jeti (2009)	**Locale, services Web**	*Control-flow*	Oui	Bioinformatique
Triana (1997)	**Client-Serveur**	*Data-flow/ Control-flow*	Oui	Général

Revue récente: Ocaña K, de Oliveira D. Parallel computing in genomic research: advances and applications. *Advances and Applications in Bioinformatics and Chemistry : AABC*. 2015;8:23-35.

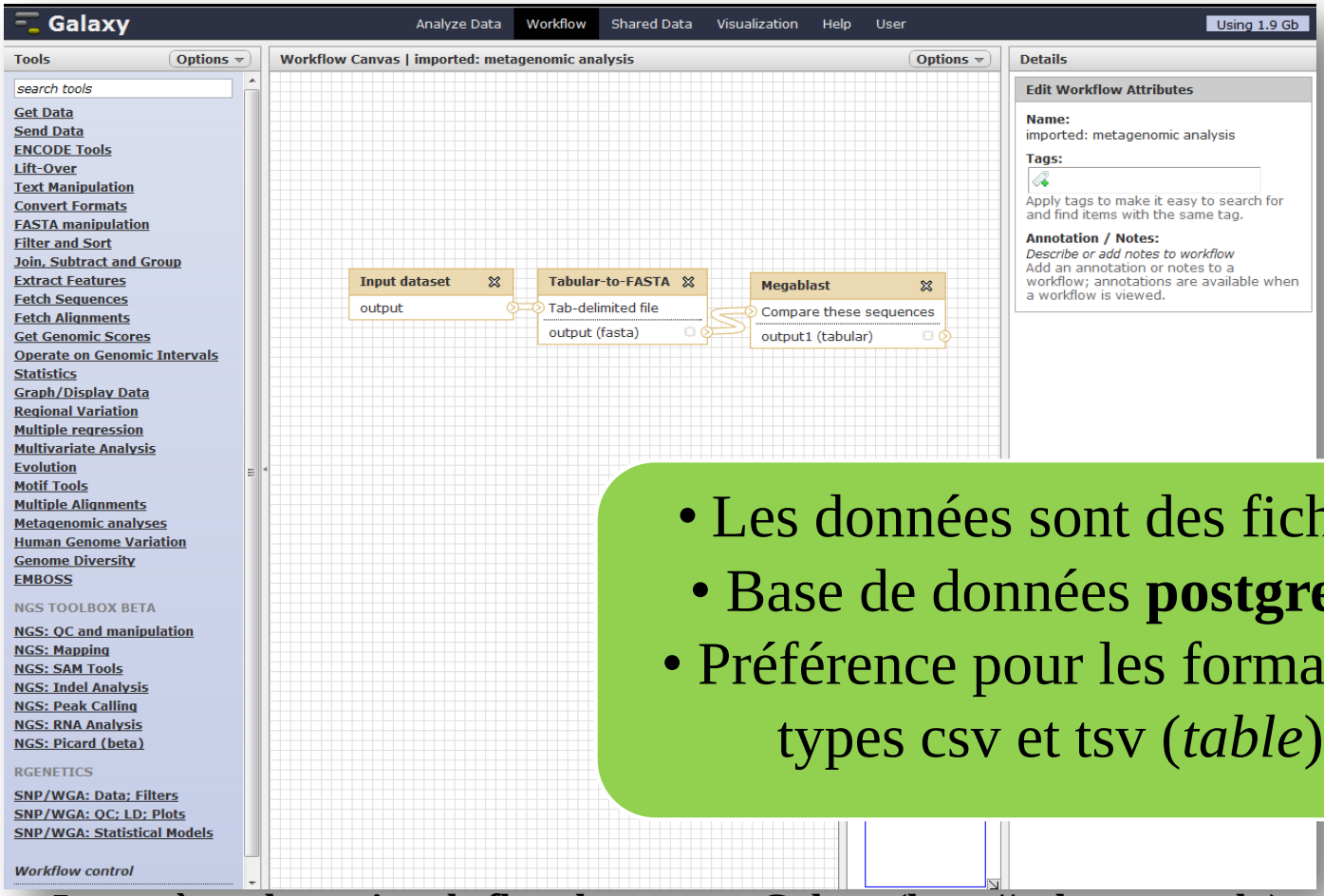
Répertoire de workflows (myexperiment.org)



En 2018:
3892 workflows
~ 10 600 membres



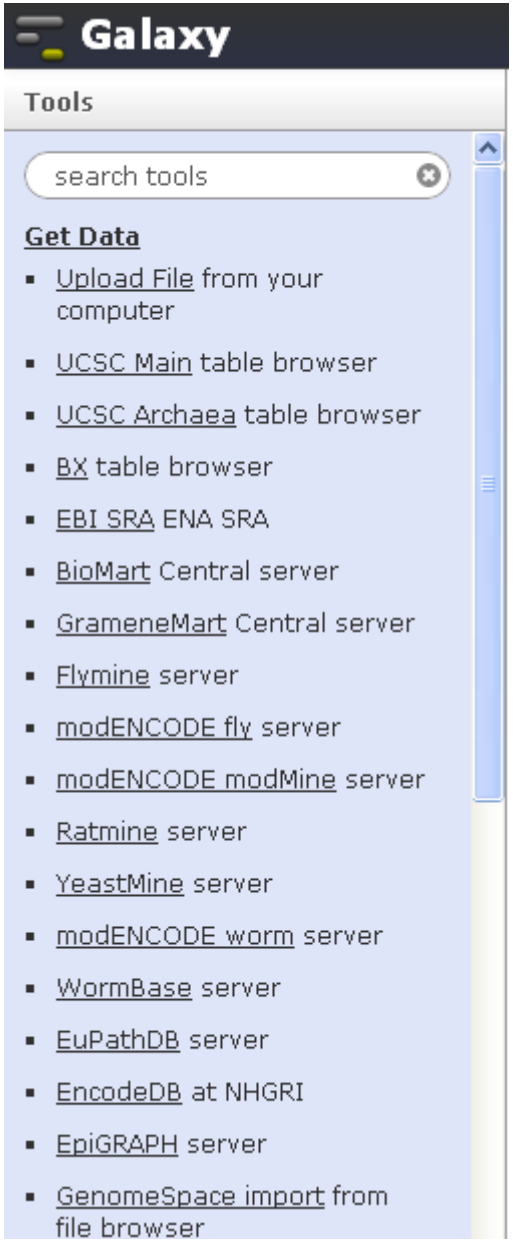
Serveur Web ou Serveur Local | Langage Python et PHP | Data-flow
Données externes | Bioinformatique | Définition des outils en format XML



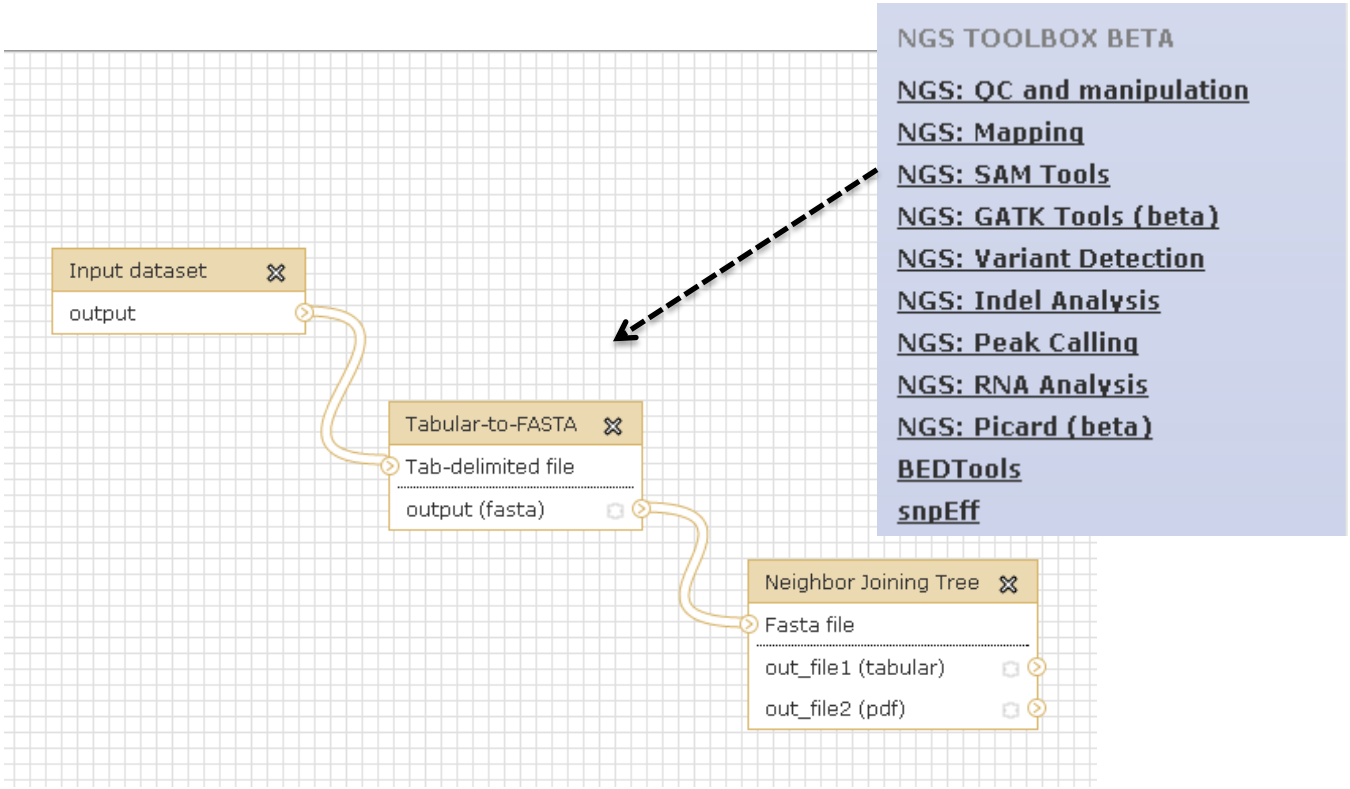
- Les données sont des fichiers
- Base de données **postgresql**
- Préférence pour les formats de types csv et tsv (*table*)

Le système de gestion de flux de travaux Galaxy (http://galaxy.psu.edu)
Giardine, B. et al. (2005). Galaxy: a platform for interactive large-scale genome analysis, Genome Res.15:1451-1455.

Galaxy



Next generation sequencing



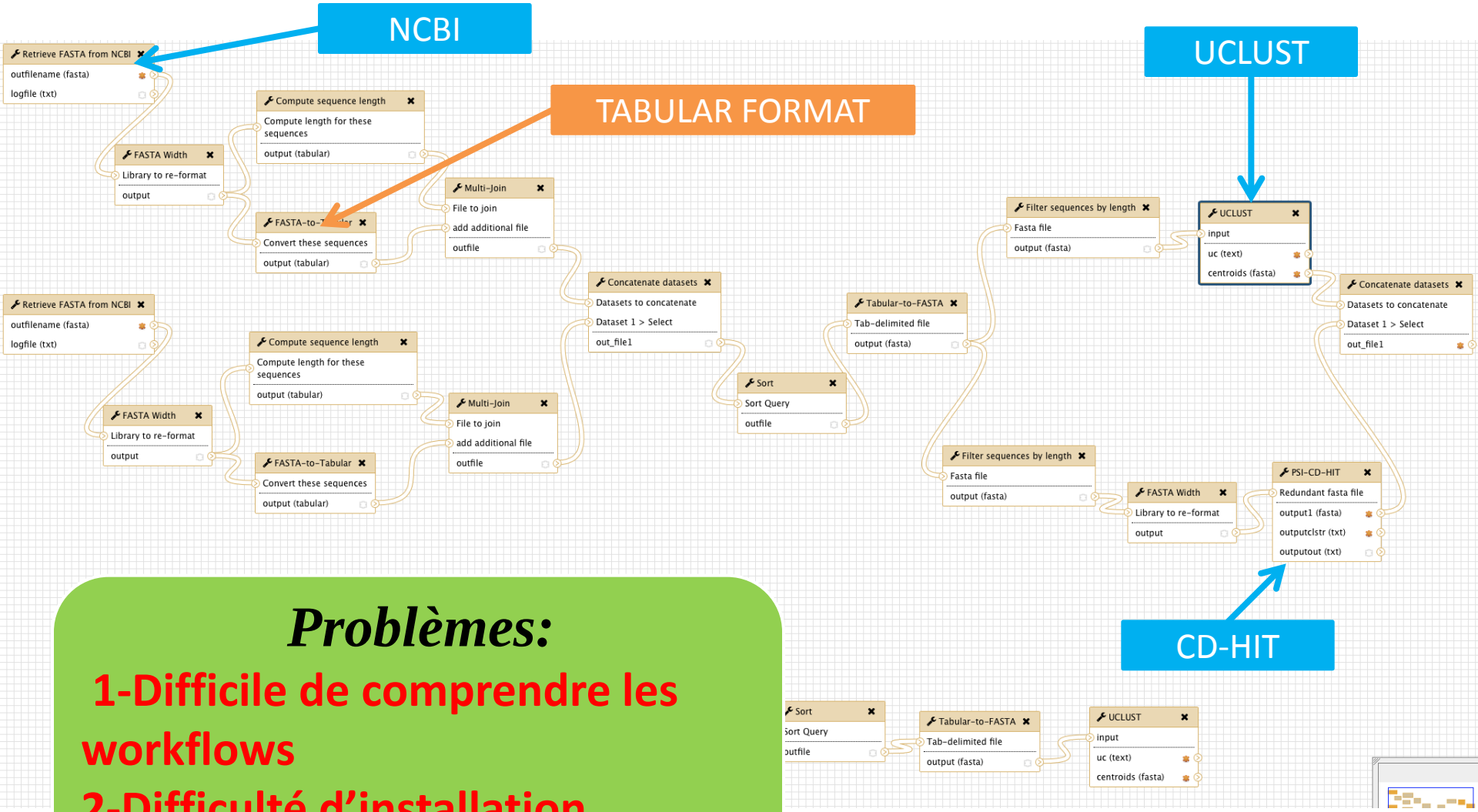
Exemples de flux de travaux

http://main.g2.bx.psu.edu/workflow/list_published

Serveurs publics

https://wiki.galaxyproject.org/PublicGalaxyServers

Galaxy | https://www.myexperiment.org/workflows/4921



Problèmes:

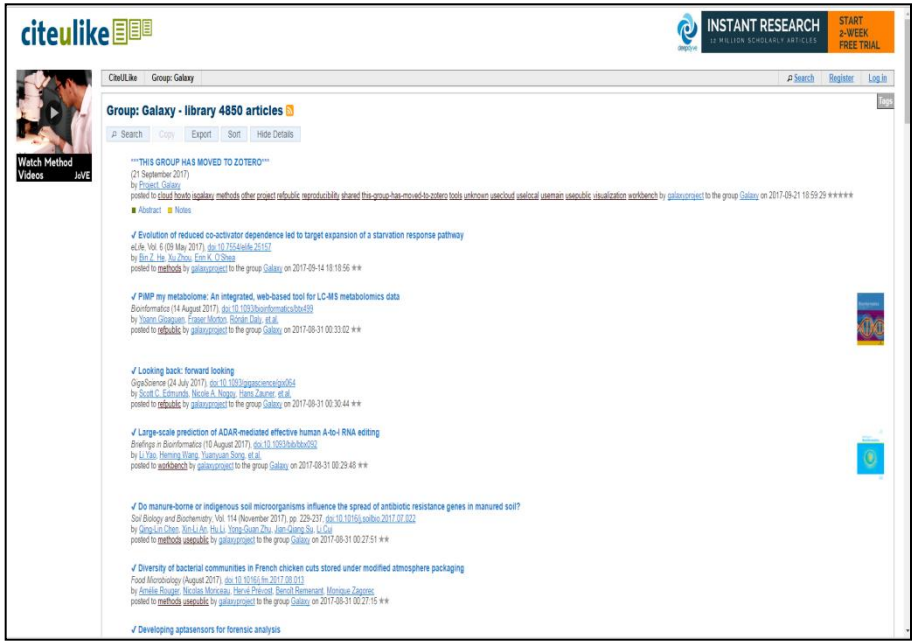
1-Difficile de comprendre les workflows

2-Difficulté d'installation

3-Pas d'outils standards

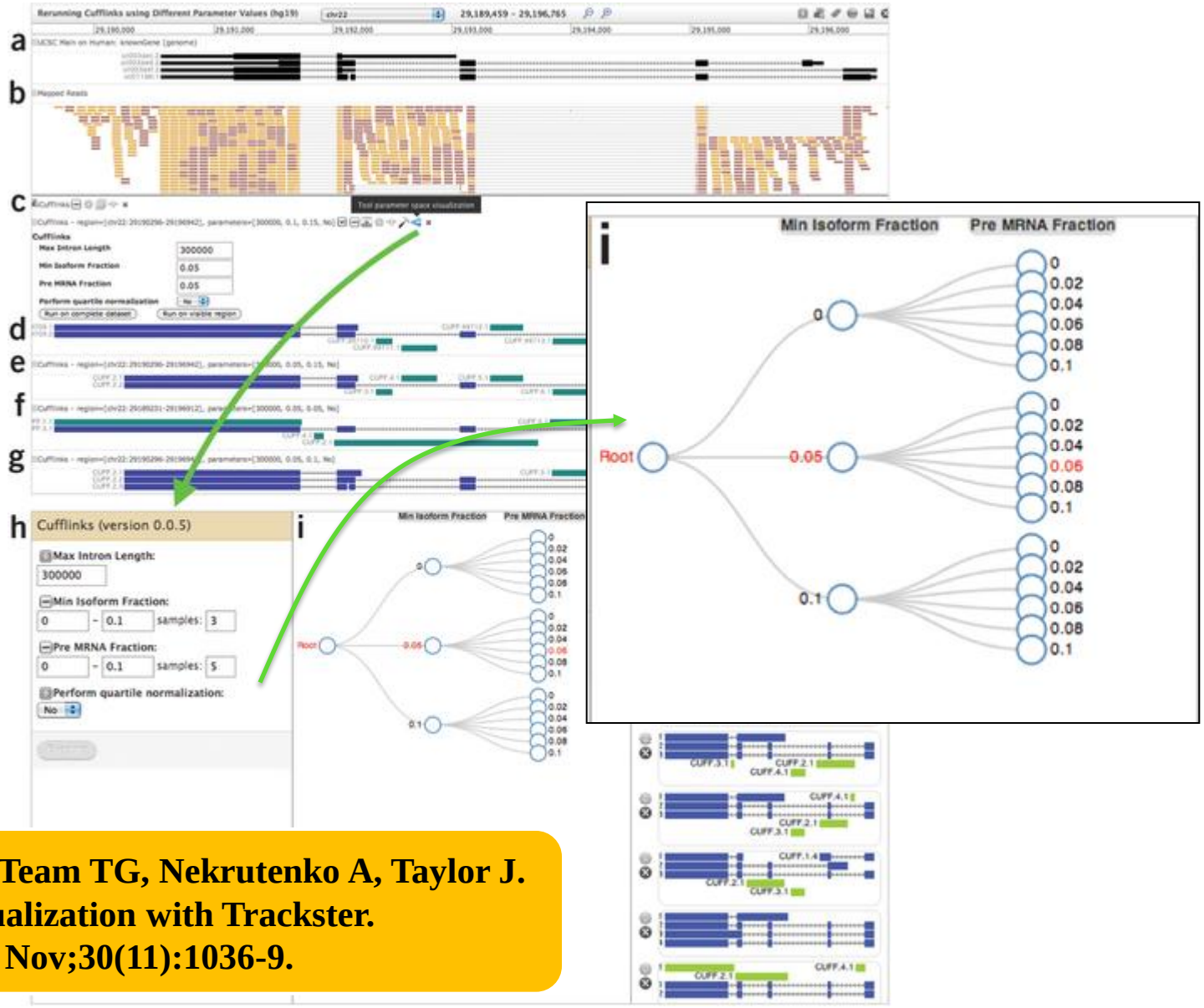
5455 publications

http://www.citeulike.org/group/16008/



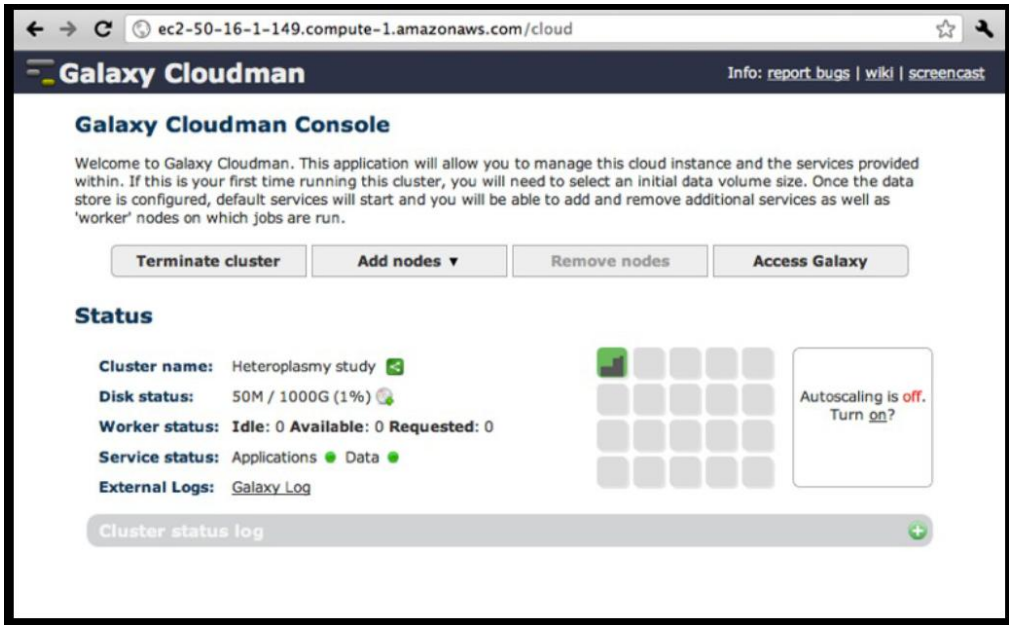
https://www.zotero.org/groups/1732893/galaxy?

Galaxy (workflows imbriqués)



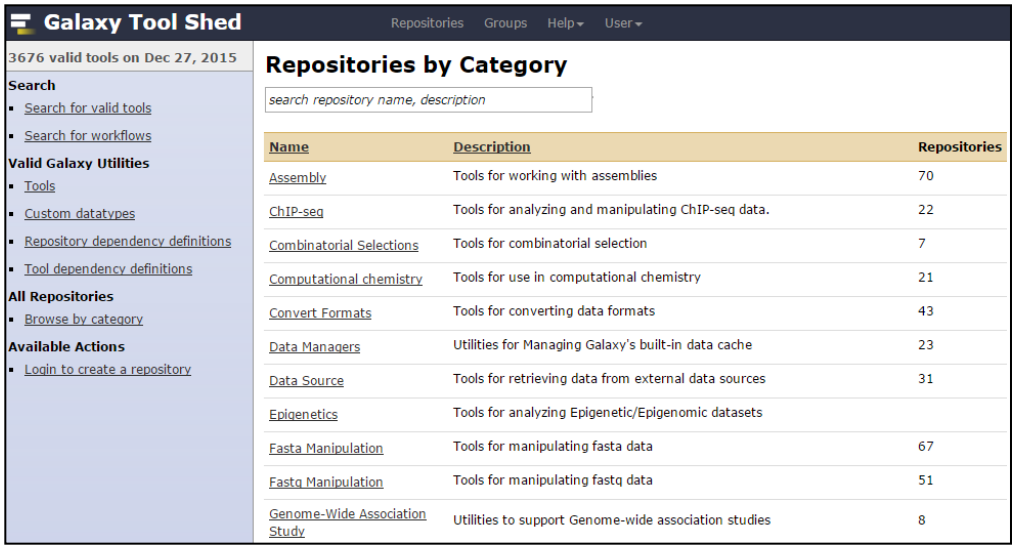
Goecks J, Coraor N, Team TG, Nekrutenko A, Taylor J. NGS analyses by visualization with Trackster. Nat Biotechnol. 2012 Nov;30(11):1036-9.

Galaxy (Amazon Cloud) et Tool Shed



amazon web services™

Afgan, E *et al.* (2012) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics. 11 Suppl 12:S4.

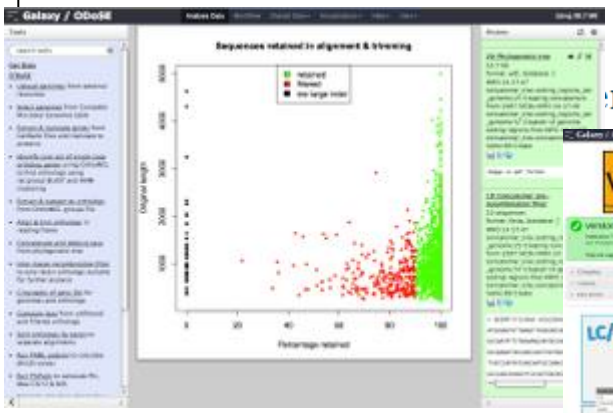


Galaxy Tool Shed

https://toolshed.g2.bx.psu.edu/

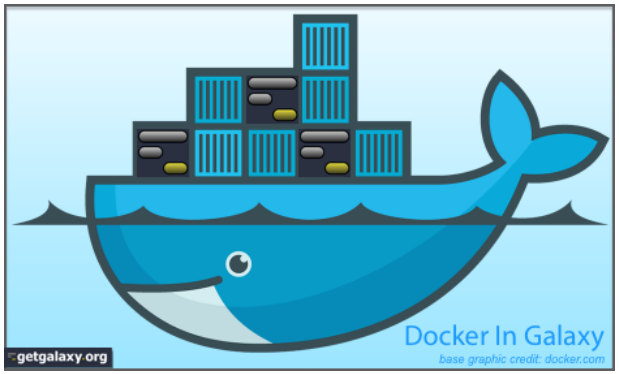
46 outils seulement en phylogénétique!

Galaxy



https://wiki.galaxyproject.org/PublicGalaxyServers

Machines Virtuelles pour Galaxy



Galaxy Virtual Appliance Directory

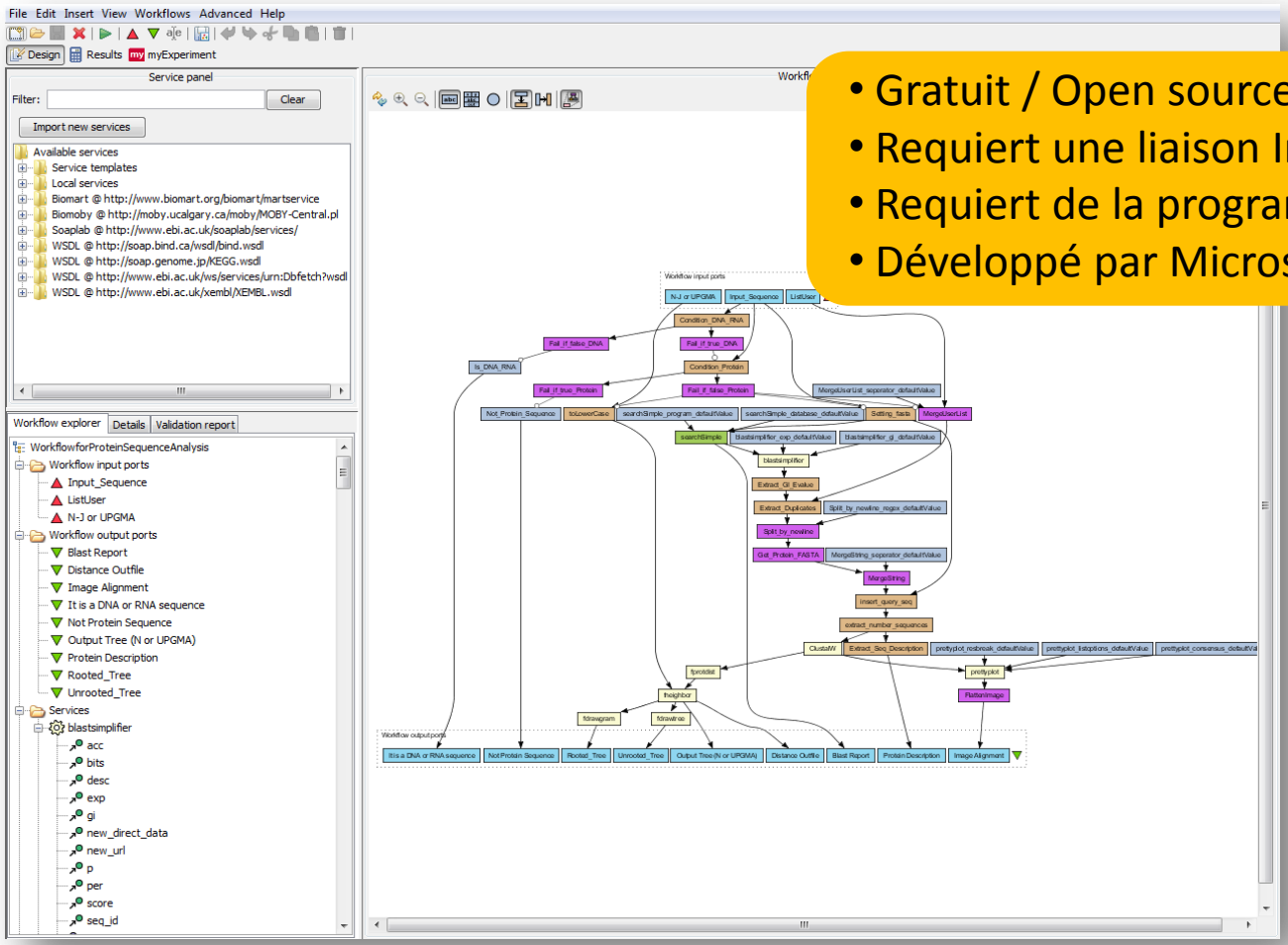
This lists all known available Galaxy virtual appliances. The goal is to make it easy for the Galaxy community to find and deploy them, thus benefiting from the (generous!) work of others.

Appliance	Technology	Domains	Description	Owners	Date Created/Updated
DDBJ Read Annotation Pipeline	VM	Reference genome mapping, *de novo* assembly, and structural and functional annotations	VirtualBox VM created in 2015 so that "Researchers may utilize the virtual machine image for users' sensitive datasets such as human personal genomic sequences."	DDBJ Read Annotation Pipeline, Development team	2015
RNAmapper	VM, AMI	Identify genome region linked to a mutation, and causal candidate mutations	RNAmapper uses RNA-seq data to identify both a region of the genome linked to a mutation as well as candidate mutations that may be causal for the phenotype of interest.	Moens Lab, Fred Hutchinson Cancer Research Center; Megason Lab, Harvard Medical School	2012/08/03
MegaMapper	VM, AMI	positional cloning of mutations	MegaMapper is a computational pipeline for positional cloning of mutations by whole genome sequencing.	Megason Lab, Harvard Medical School	2013/05/02
Automated Selection of Hotspots (ASH)	VM	hotspot detection	automated detection quantitative ranking of hotspots to support histopathologists in selecting the 'hottest' hotspot areas in adrenocortical carcinoma.	David van Zessen <d.vanzessen@erasmusmc.nl>, Erasmus Medical Center	2014/05/16
UC Davis Bioinformatics Core Training Galaxy	AMI	Training	Includes all needed tools and datasets for a weeklong workshop.	UC Davis Bioinformatics Core	2014/06/20
Immunoglobulin Galaxy	VM	immunoglobulin	a set of tools for easy detection and quantitation of immunoglobulin heavy chain alternative transcripts.	David van Zessen <d.vanzessen@erasmusmc.nl>, Erasmus Medical Center	2014/08/19
ballaxy	Docker	computer aided drug design, molecular modelling	Contains the BALL (Biochemical Algorithms Library) Project tools, i.e. computer aided drug design and molecular modelling based on protein and ligand structure data.	Hans-Peter Lenhof Lab, Saarland University; Oliver Kohlbacher Lab, University of Tübingen; Andreas Hildebrandt Lab, University of Mainz	2014/09/02
Pitagora-Galaxy	VM, AMI	General Purpose	Includes the tools and workflows that are described at our wiki.	Genome Science Division, RCAST, University of Tokyo	2014/11/20
SADI-Galaxy Container	Docker	Web services, RDF	Make SADI services available in a Galaxy platform.	Mikel Egaña Aranguren	2015/01/07
RNA Galaxy Workbench	Docker	RNA-Seq	A Docker based RNA-Seq workbench	RNA Bioinformatics Center, part of the German Network for Bioinformatics Infrastructure (de.NBI).	2015/02/17

https://wiki.galaxyproject.org/VirtualAppliances, https://hub.docker.com/r/anhi/ballaxy/

Taverna (v2.5 et 3.0)

Serveur Web ou Serveur Local | Langage Java | Données externes
Data-flow et control-flow avec programmation |Bioinformatique

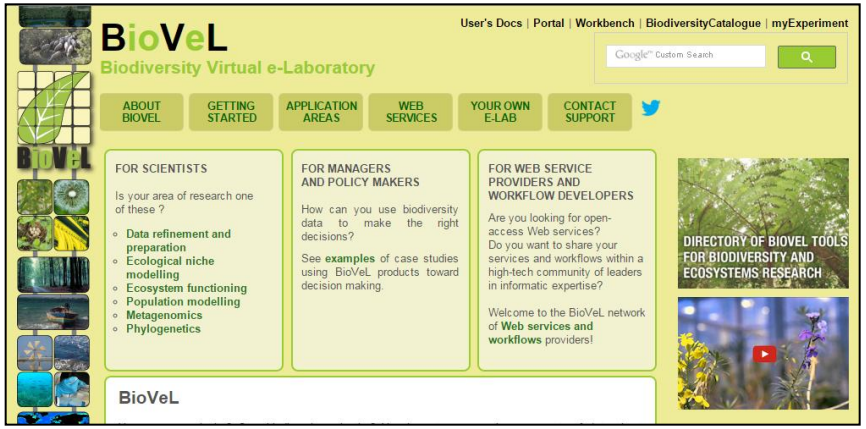


- Gratuit / Open source
- Requiert une liaison Internet
- Requiert de la programmation (java)
- Développé par Microsoft, Apache

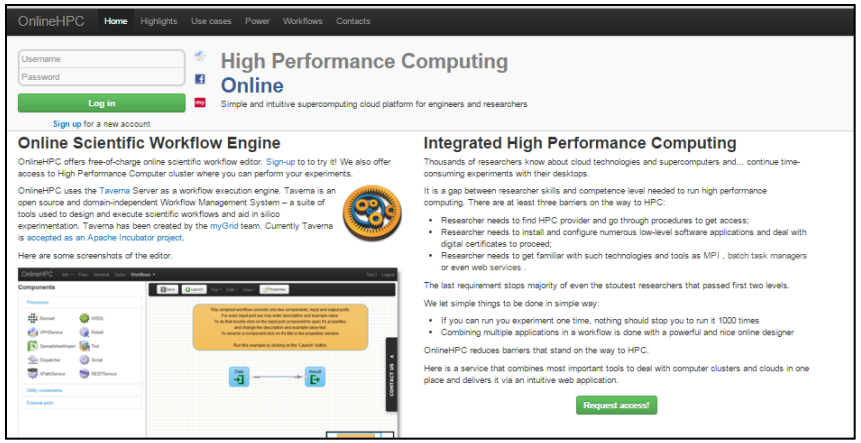
https://taverna.incubator.apache.org/documentation/taverna-galaxy/

Stevens, et al. (2003). myGrid: personalised bioinformatics on the information grid. Bioinformatics, 19:i302-i304.

Répertoire de Services Webs / Intégration



https://www.biodiversitycatalogue.org/
https://www.biovel.eu/



http://onlinehpc.com/site/main

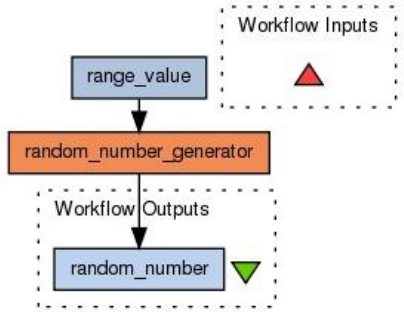
Integrated with other myGrid tools

- myexperiment – workflow sharing environment for scientists
- *Bio*Catalogue – curated catalogue of Web services

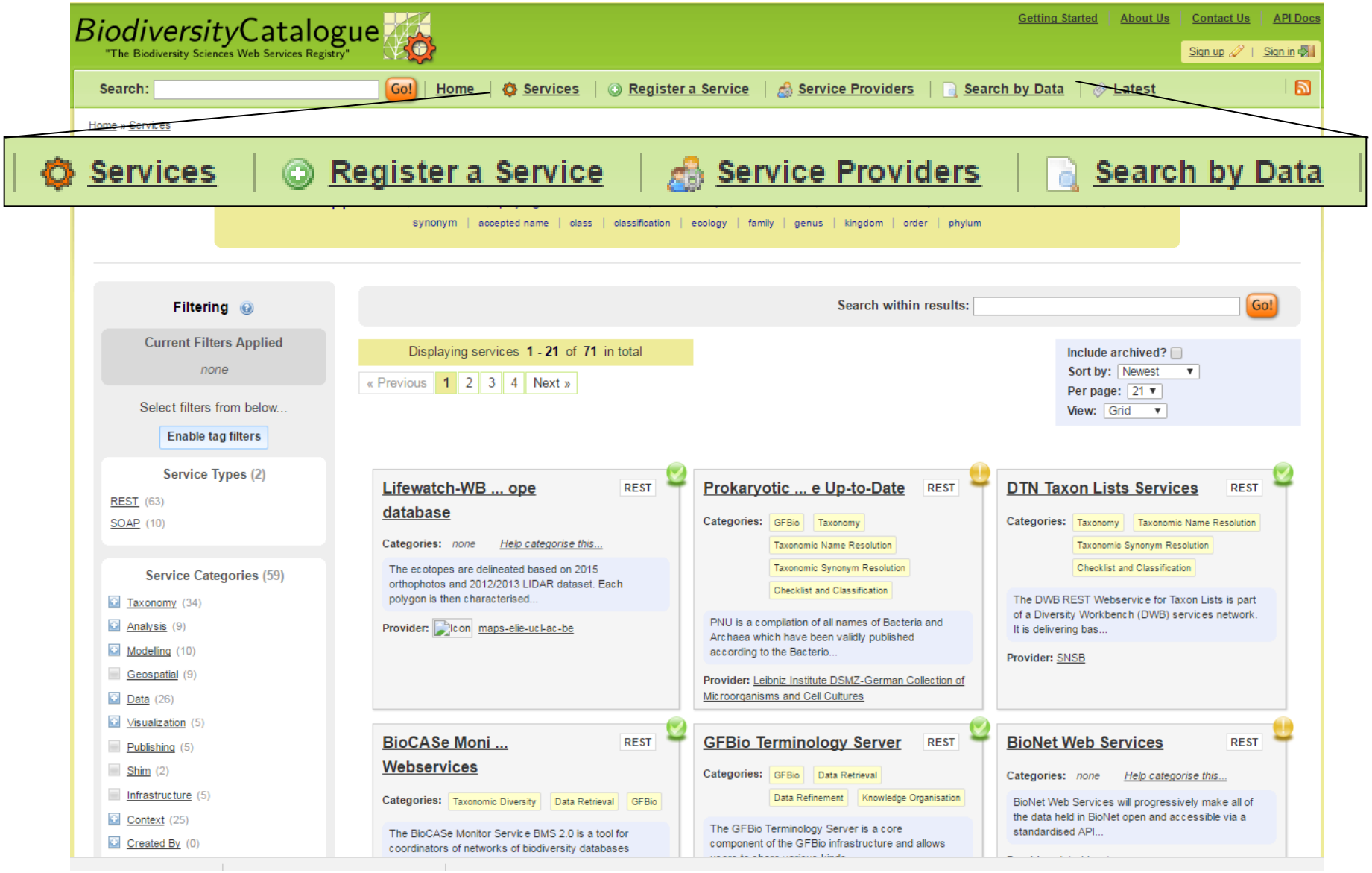
Works with …

- biomart – data integration system for large scale data querying
- SoapLab – Web services wrapping existing command-line analysis programs (such as EMBOSS sequence analysis)
- SADI – semantic Web services described using SADI
- R – software for statistical computing and graphics
- Bioconductor – tools for the analysis and comprehension of high-throughput genomic data
- A number of third party plugins

Intégration des langages de programmation: Python, R, Perl, Java.
Exemple de python:
www.myexperiment.org/workflows/2475



BiodiversityCatalogue



BioVel (https://www.biovel.eu/)

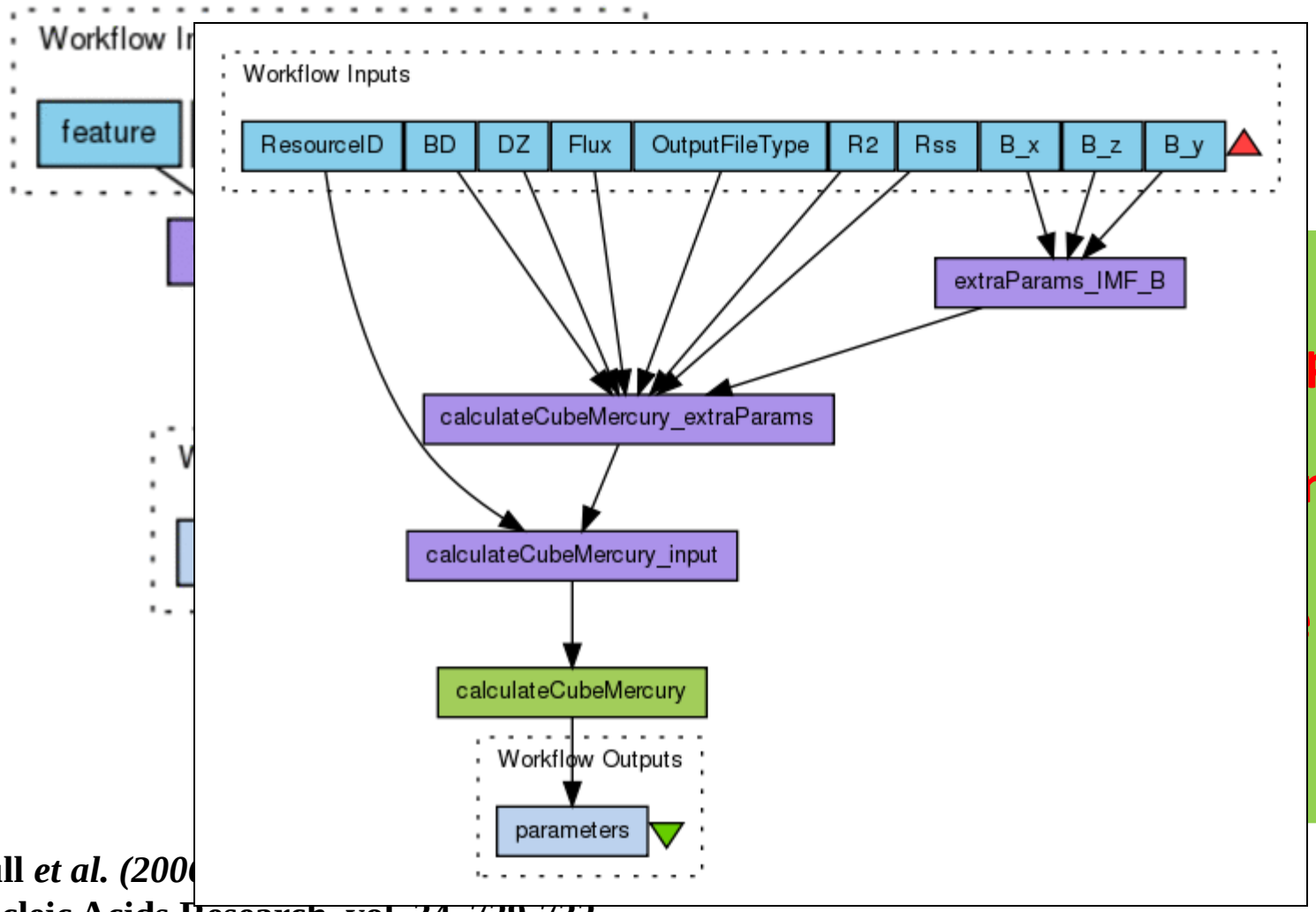
Biodiversity Virtual e-Laboratory

Phylogenetics	Bayesian phylogenetic inference workflows are for performing phylogenetic inference for systematics and diversity research. Bayesian methods guide selection of the evolutionary model and a post hoc validation of the inference is also made. Phylogenetic partitioning of the diversity across samples allows study of mutual information between phylogeny and environmental variables purl: http://www.purl.ox.ac.uk/researchobj/myexp-370 Portal: https://www.portal.biovel.eu/workflows/466 https://www.portal.biovel.eu/workflows/549 https://www.portal.biovel.eu/workflows/550 https://www.portal.biovel.eu/workflows/525 PDP workflow, using PhyloH for partitioning environmental sequencing data using both categorical and phylogenetic information purl: http://www.purl.ox.ac.uk/workflow/myexp-3570.5 Portal: https://www.portal.biovel.eu/workflows/434 https://www.portal.biovel.eu/workflows/71 MSA-PAD workflow performs a multiple DNA sequence alignment coding for multiple/single protein domains invoking two alignment modes: gene and genome Gene mode purl: http://www.purl.ox.ac.uk/workflow/myexp-4549.1 Portal: https://www.portal.biovel.eu/workflows/712 (access on request) Genome mode purl: http://www.purl.ox.ac.uk/workflow/myexp-4551.1 Portal: https://www.portal.biovel.eu/workflows/713 (access on request) SUPERSMART (self-updating platform for estimating rates of speciation and migration, ages and relationships of taxa) is a pipeline analytical environment for large-scale phylogenetic data mining, taxonomic name resolution, tree inference and fossil-based tree calibration url: https://www.biodiversitycatalogue.org/services/78 Scientific studies: [44–47]

Web service

Hardisty et al. (2016) BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecol. 2016 Oct 20;16(1):49. PubMed PMID: 27765035

Approche de Taverna



ples

non une

nom des

Hull *et al.* (200 Nucleic Acids Research, vol. 34, 729-732.

Oinn *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 20(17):3045-54.

Taverna

Ajout d'information sur la provenance des données et ontologie

github.com/taverna/taverna-prov

Command Line Tool 2.5

Taverna Command Line Tool 2.5 is basically the Taverna Workbench stripped of its GUI so you can run workflows from a command prompt headlessly (i.e. without the graphical environment).

It is available in several editions. The Core edition contains services for running scientific workflows in any domain, accessing general services such as REST or SOAP Web services and external command line tools.

Domain-specific editions (see below) add support for specific service types.

Enterprise Taverna is an 'all inclusive' edition which contains all available service types, equivalent to the Core edition with all official plugins included.

Available editions

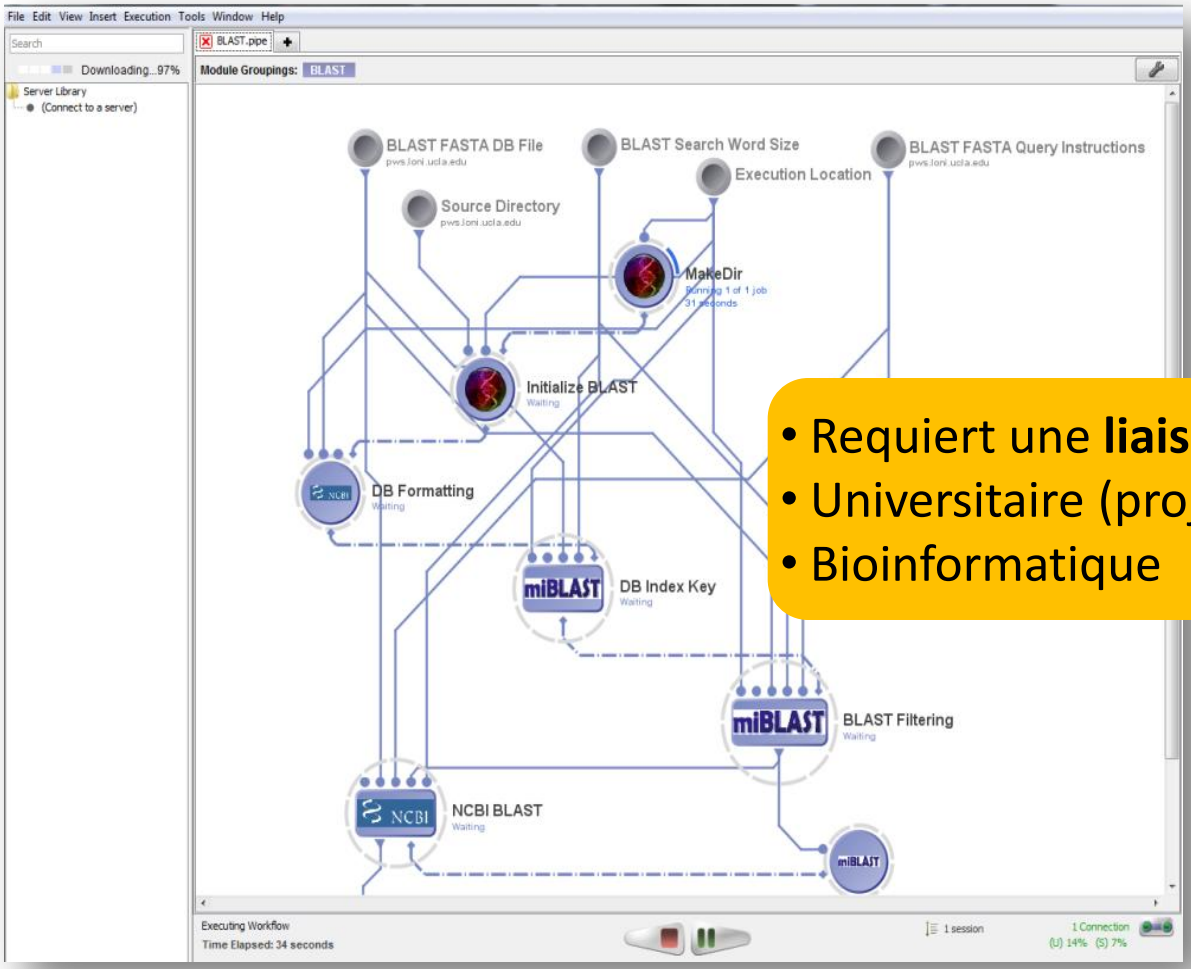
Check system requirements before you download Taverna.

- Taverna Command Line Tool Core 2.5
- Taverna Command Line Tool Astronomy 2.5
- Taverna Command Line Tool Bioinformatics 2.5
- Taverna Command Line Tool Biodiversity 2.5
- Taverna Command Line Tool Digital Preservation 2.5
- Taverna Command Line Tool Enterprise 2.5

Quelques outils bioinformatiques disponibles

Loni

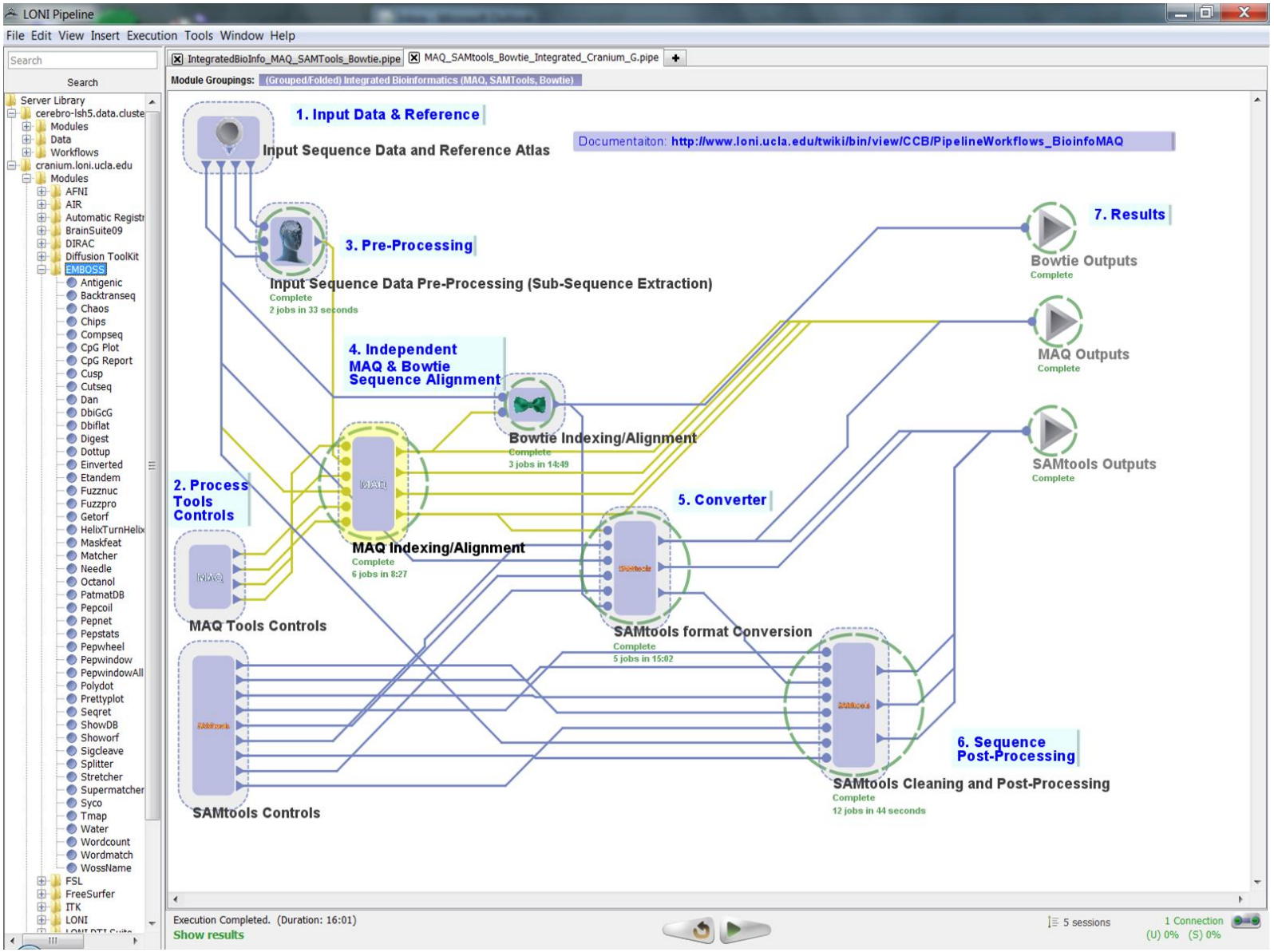
Service webs et Local| Langage Java| Données externes
Data-flow sans programmation |Bioinformatique



- Requiert une **liaison Internet**
- Universitaire (projet de recherche)
- Bioinformatique

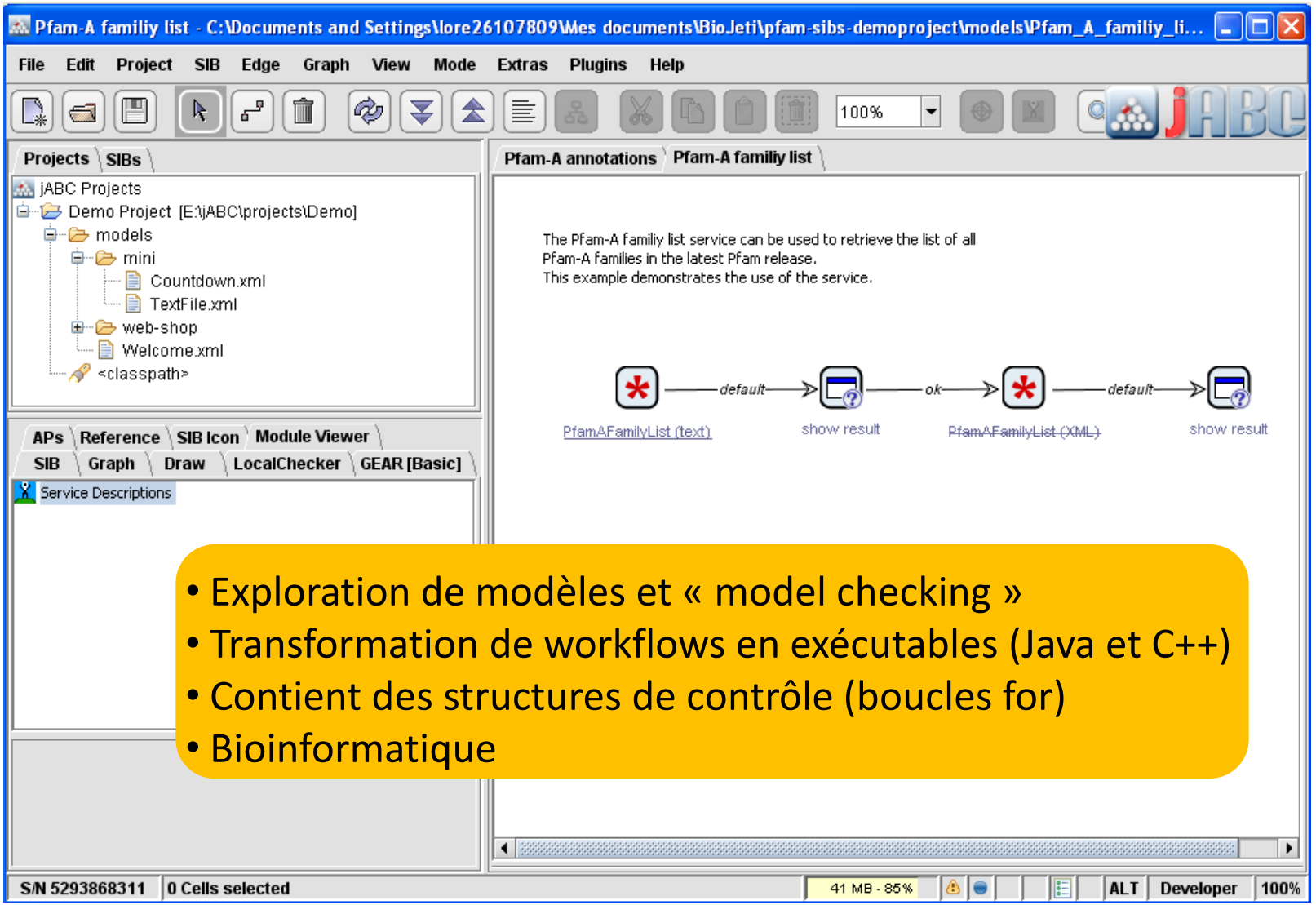
pipeline.loni.ucla.edu

LONI



http://www.ccb.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMAQ

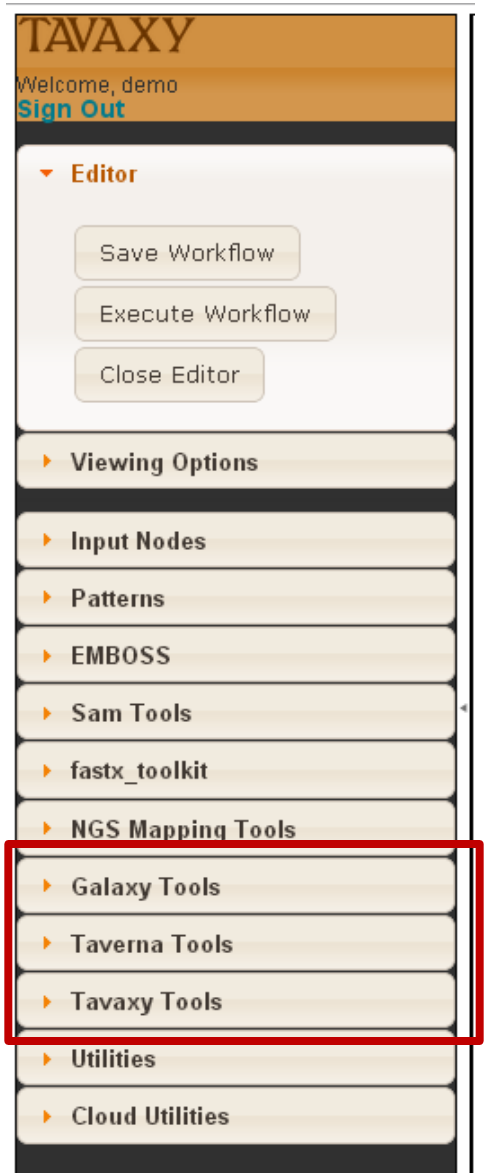
Bio-Jeti

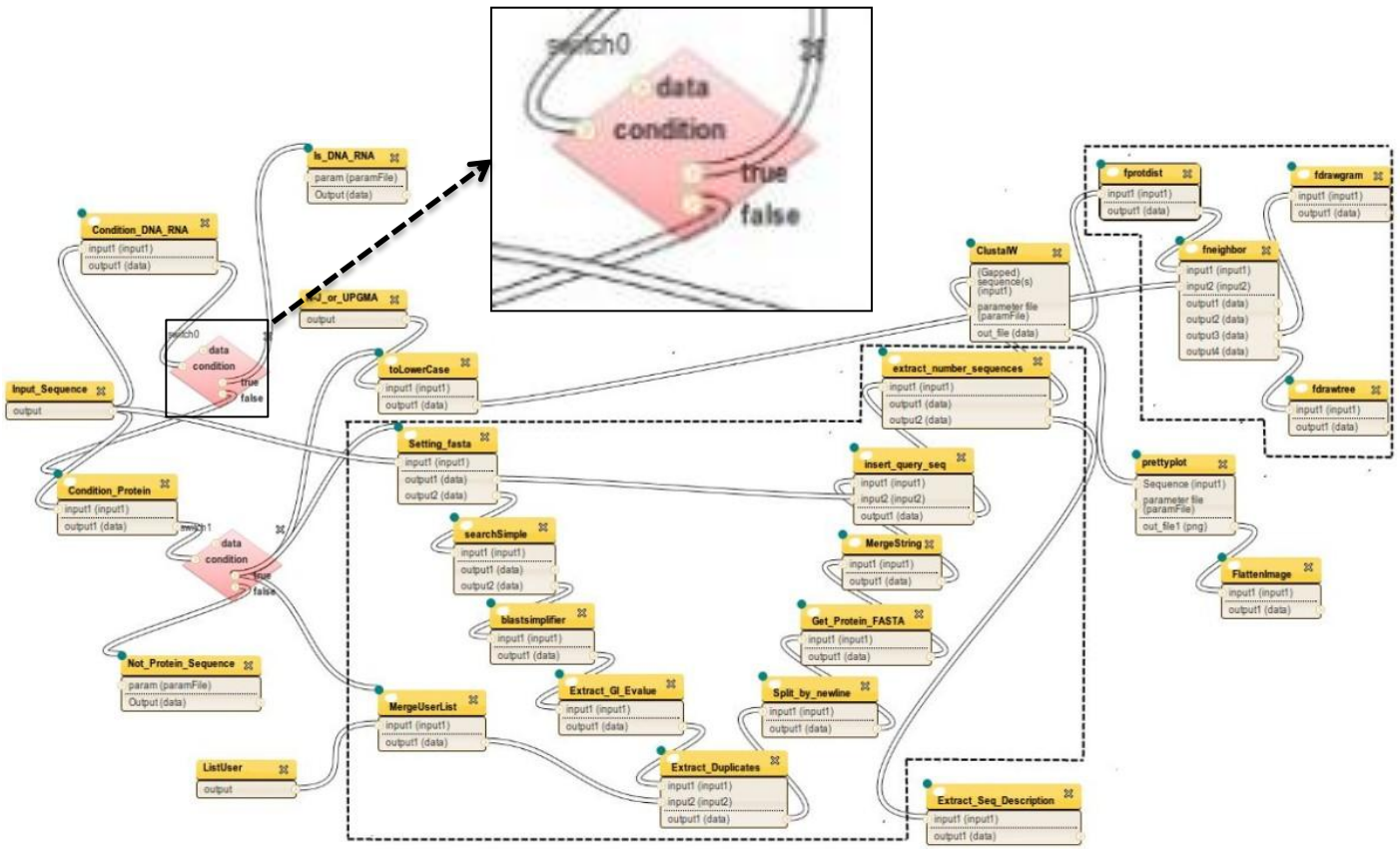


- Exploration de modèles et « model checking »
- Transformation de workflows en exécutables (Java et C++)
- Contient des structures de contrôle (boucles for)
- Bioinformatique

http://ls5-www.cs.tu-dortmund.de/projects/biojeti/index.php

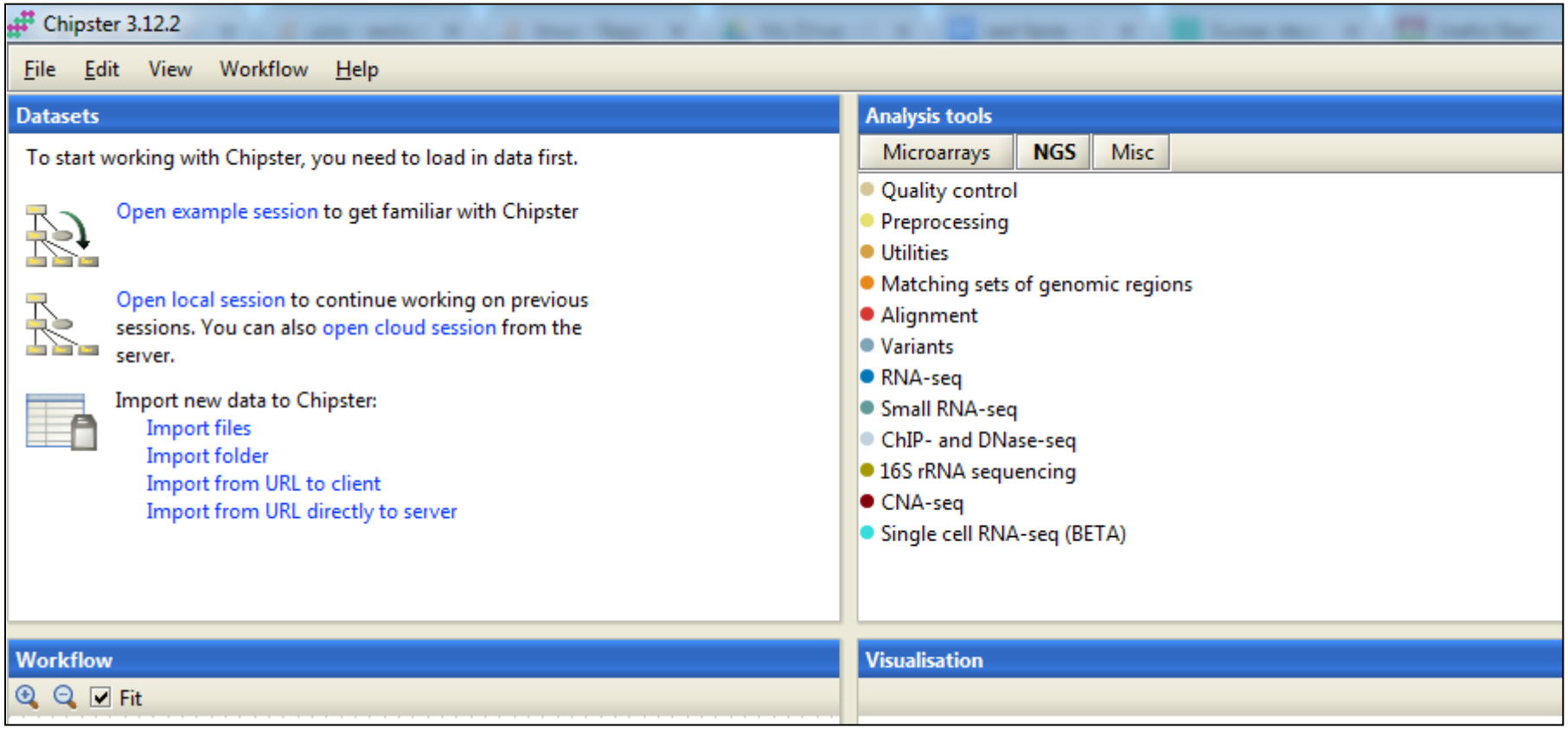
Tavaxy : lien entre Taverna et Galaxy





Abouelhoda M, Issa SA, Ghanem M. Tavaxy: integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics. 2012 May 4;13:77.

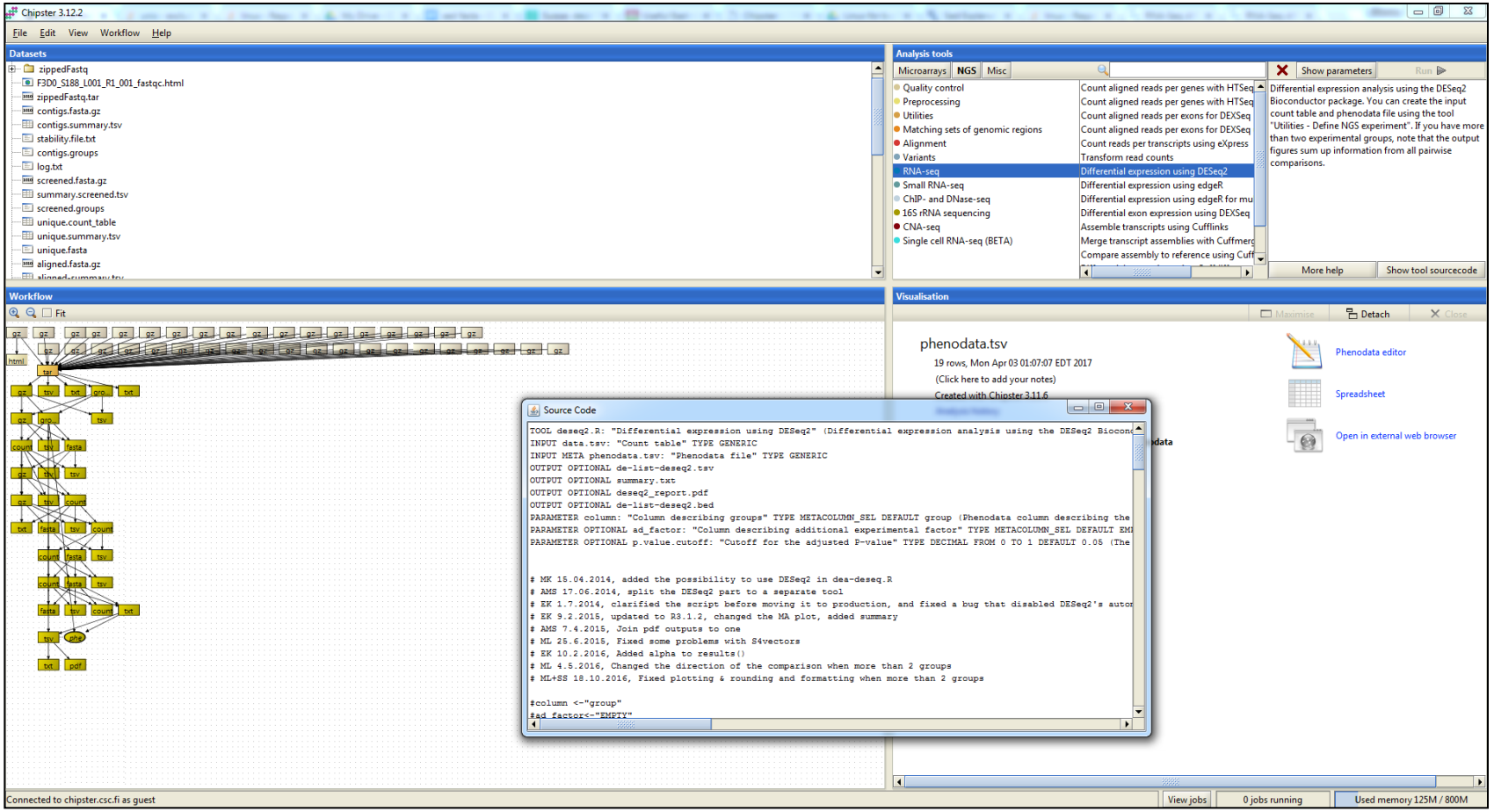
Chipster



https://chipster.github.io/chipster/

https://chipster.csc.fi

Chipster



Kallio, M. Aleksi, Jarno T. Tuimala, Taavi Hupponen, Petri Klemelä, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Käki, and Eija I. Korpelainen. "Chipster: user-friendly analysis software for microarray and other high-throughput data." *BMC genomics* 12, no. 1 (2011): 507.

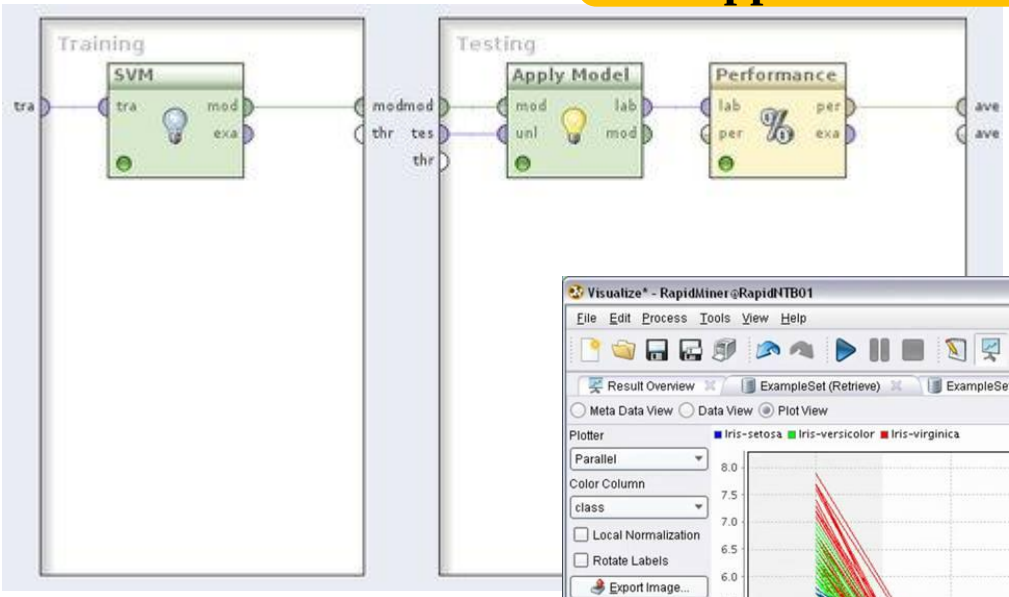
Exemple Chipster

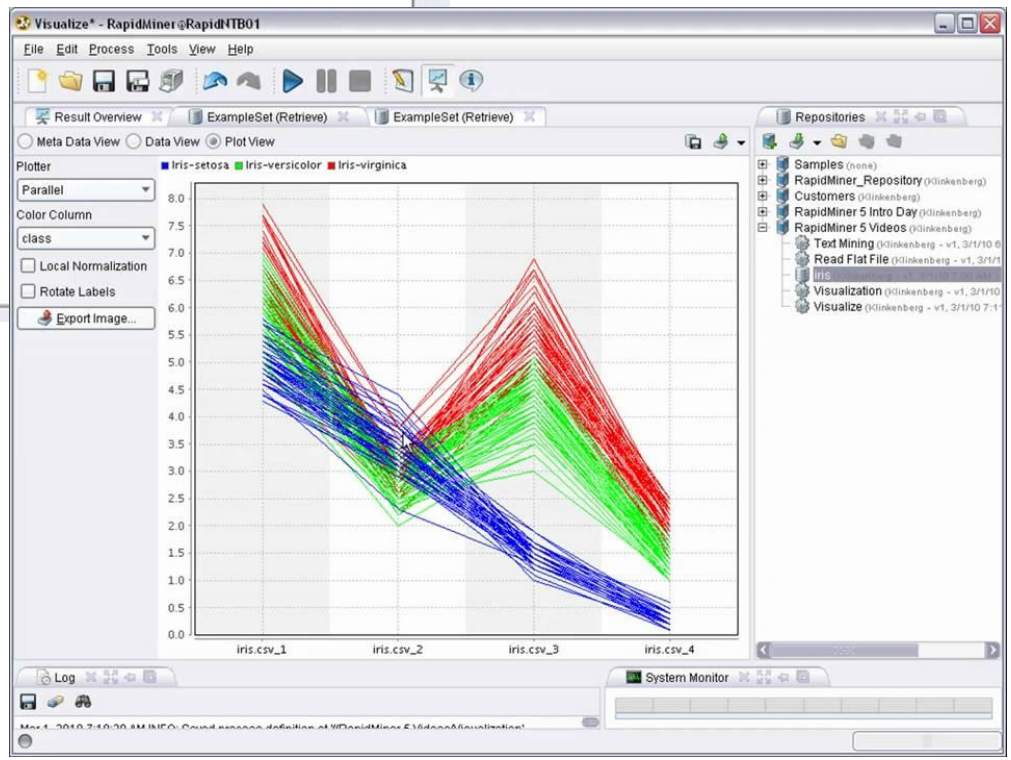
```
TOOL fastx-quality-filter.R: "Filter reads for quality with FastX" (Filters out reads
which contain quality scores below the user-specified criteria. This tool is based on
the FASTQ Quality Filter tool of the FASTX package.)
INPUT reads.fastq TYPE GENERIC
OUTPUT quality-filtered.fastq.gz
OUTPUT quality-filtered.log
PARAMETER quality: "Quality cut-off value" TYPE INTEGER FROM 1 TO 100 DEFAULT 20 (What
is the minimum quality score to keep.)
PARAMETER percentage: "Minimum percent of bases that must have that quality" TYPE
INTEGER FROM 1 TO 100 DEFAULT 90 (Percent of bases in sequence that must have quality
equal to or higher than the cut-off value.)
PARAMETER quality.format: "Quality value format used" TYPE [sanger: Sanger, illuminaold:
"Illumina GA v1.3-1.5"] DEFAULT sanger (What quality encoding is used in your FASTQ
file. Select Sanger if your data comes from Illumina 1.8 or later, SOLiD or 454.)

# EK 28.6.2011
# AMS 11.3.2014, gzip fastq outputs
# check out if the file is compressed and if so unzip it
source(file.path(chipster.common.path, "zip-utils.R"))
unzipIfGZipFile("reads.fastq")
# binary
binary <- c(file.path(chipster.tools.path, "fastx", "bin", "fastq_quality_filter"))
# command
quality.scale <- ifelse(quality.format == "sanger", "-Q 33", "")
command <- paste(binary, "-v", "-q", quality, "-p", percentage, quality.scale, "-i
reads.fastq -o quality-filtered.fastq > quality-filtered.log")

# run
system(command)
system("gzip *.fastq")
```

Locale | Langage Java | *Data-flow*
Application en AI et Machine Learning



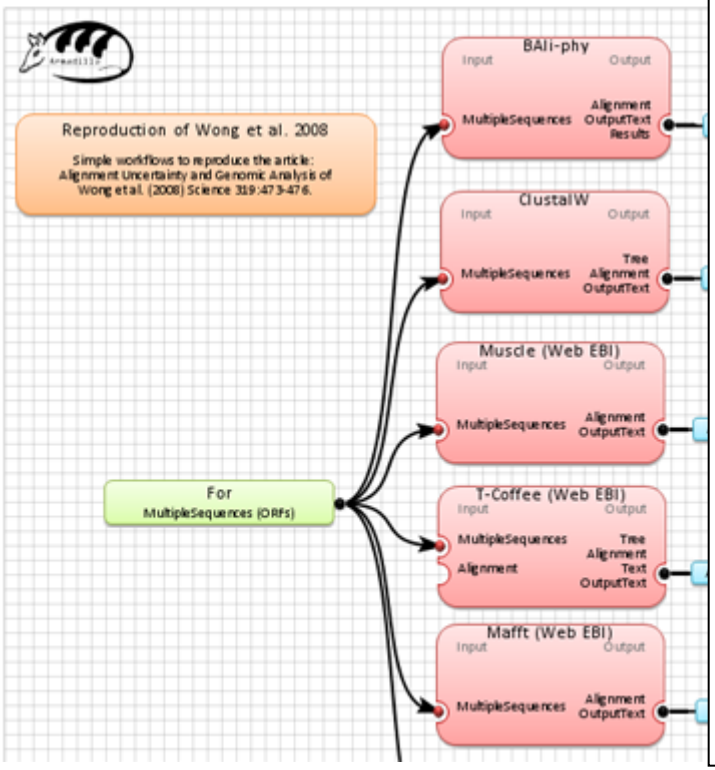


Téléchargement de Armadillo

Armadillo v1.1

adn.bioinfo.uqam.ca/armadillo
https://github.com/armadilloUQAM/armadillo2

HOMEPAGE | APPLICATIONS | TUTORIALS | WIKI | DOWNLOAD



OPEN ACCESS Freely available online

PLoS one

Armadillo 1.1: An Original Workflow Platform for Designing and Conducting Phylogenetic Analysis and Simulations

Etienne Lord[1], Mickael Leclercq[1], Alix Boc[2], Abdoulaye Baniré Diallo[1], Vladimir Makarenkov[1]*

1 Département d'informatique, Université du Québec à Montréal, Montréal, Canada, **2** Département de sciences biologiques, Université de Montréal, Montréal, Canada

Abstract

In this paper we introduce Armadillo v1.1, a novel workflow platform dedicated to designing and conducting phylogenetic studies, including comprehensive simulations. A number of important phylogenetic and general bioinformatics tools have been included in the first software release. As Armadillo is an open-source project, it allows scientists to develop their own modules as well as to integrate existing computer applications. Using our workflow platform, different complex phylogenetic tasks can be modeled and presented in a single workflow without any prior knowledge of programming techniques. The first version of Armadillo was successfully used by professors of bioinformatics at Université du Quebec à Montreal during graduate computational biology courses taught in 2010–11. The program and its source code are freely available at: <http://www.bioinfo.uqam.ca/armadillo>.

Citation: Lord E, Leclercq M, Boc A, Diallo AB, Makarenkov V (2012) Armadillo 1.1: An Original Workflow Platform for Designing and Conducting Phylogenetic Analysis and Simulations. PLoS ONE 7(1): e29903. doi:10.1371/journal.pone.0029903

Editor: Simon Joly, Montreal Botanical Garden, Canada

Received June 16, 2011; **Accepted** December 8, 2011; **Published** January 11, 2012

Funding: EL and AB are Natural Science and Engineering Research Council of Canada fellows (NSERC). VM and ABD hold NSERC discovery grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vladimir.makarenkov@uqam.ca

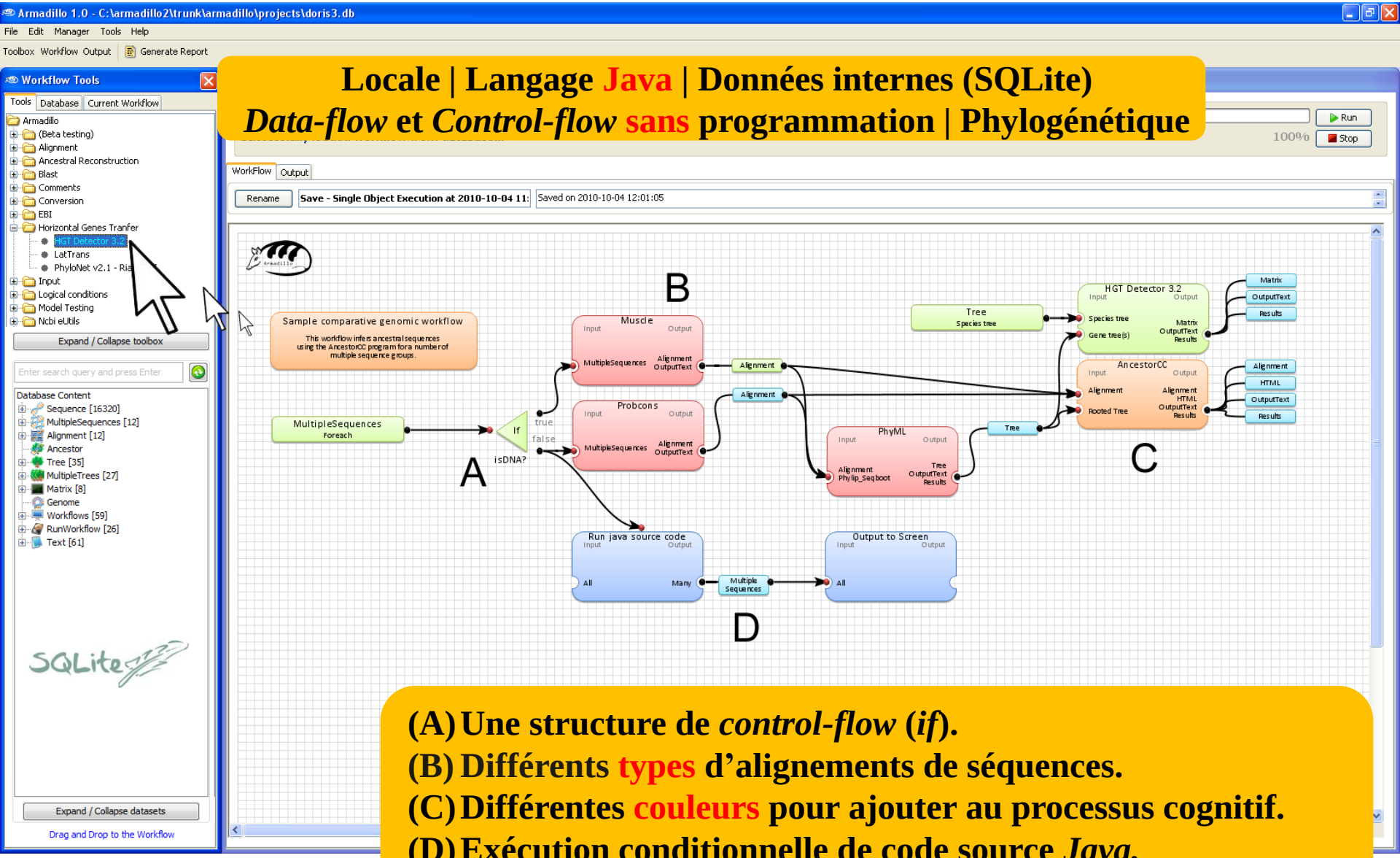
Introduction

Bioinformatics is a fast-evolving field that encompasses

Meanwhile, the increasing use of genomic and phylogenetic data fuels the need for pipeline managing software. For instance, Cicarelli et al. [9] developed an automatable procedure for

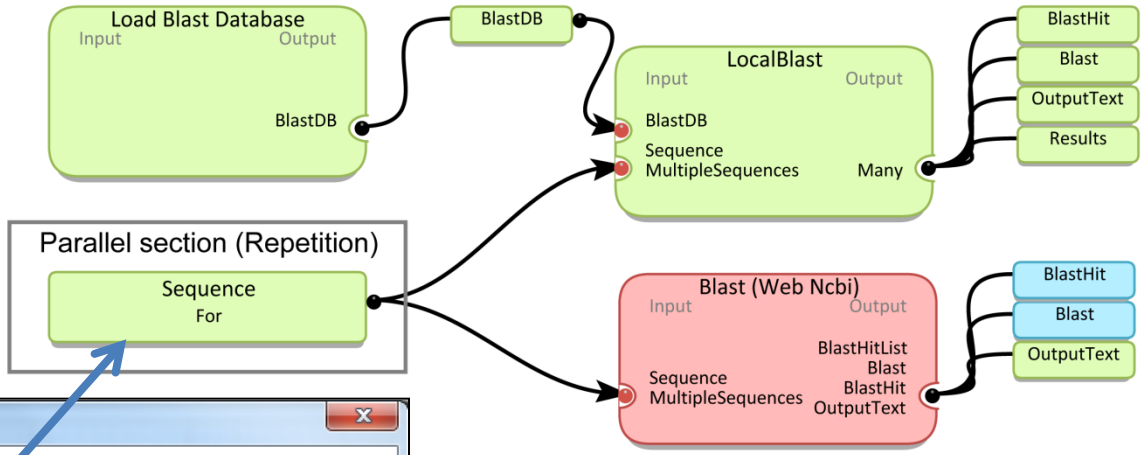
Armadillo v1.0 (adn.bioinfo.uqam.ca/armadillo)

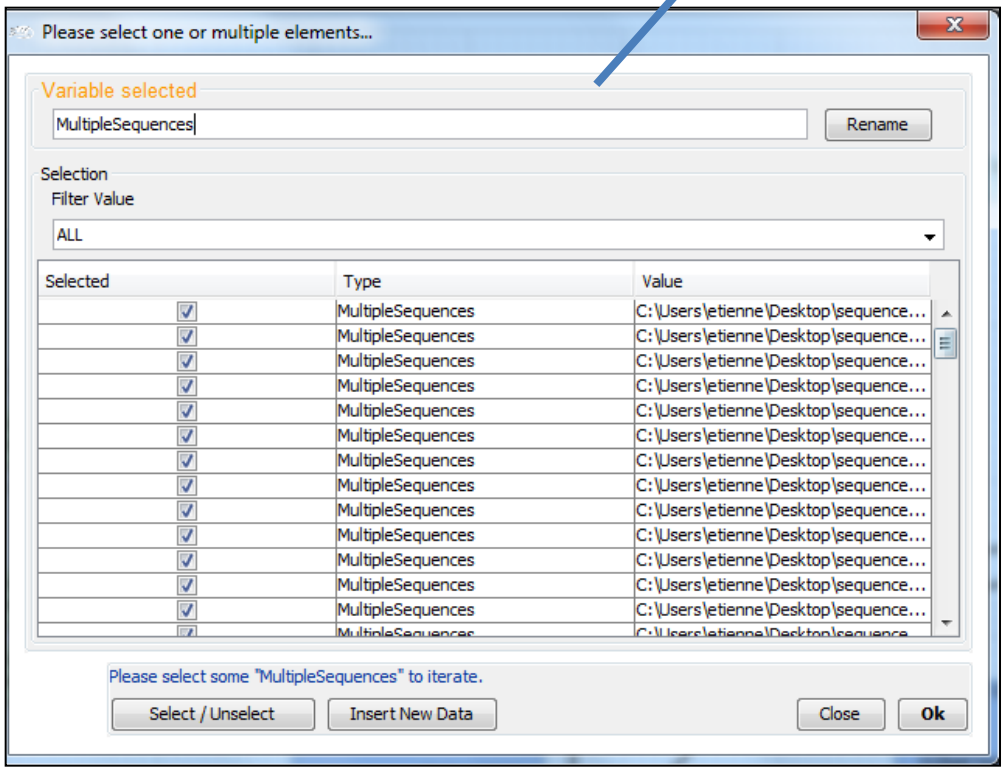
Locale | Langage Java | Données internes (SQLite)
***Data-flow* et *Control-flow* sans programmation | Phylogénétique**



(A) Une structure de *control-flow* (*if*).
(B) Différents types d'alignements de séquences.
(C) Différentes couleurs pour ajouter au processus cognitif.
(D) Exécution conditionnelle de code source *Java*.

Armadillo : structures de répétitions



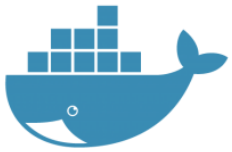


Types de données

Sequence, MultipleSequences,Tree, MultipleTrees, Blast, Alignment, Genome, Text

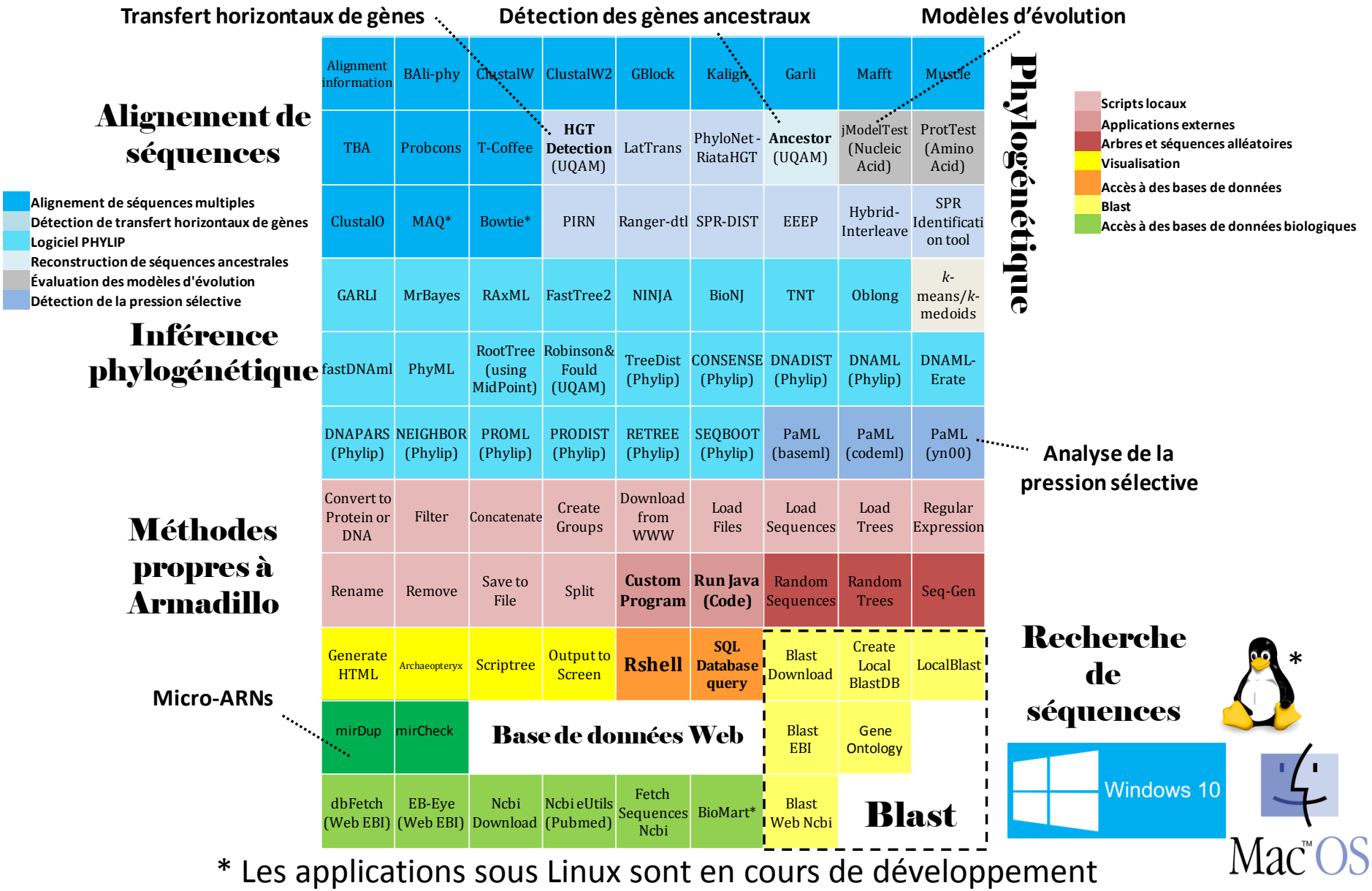
Dernière version

github.com/armadilloUQAM/armadillo2



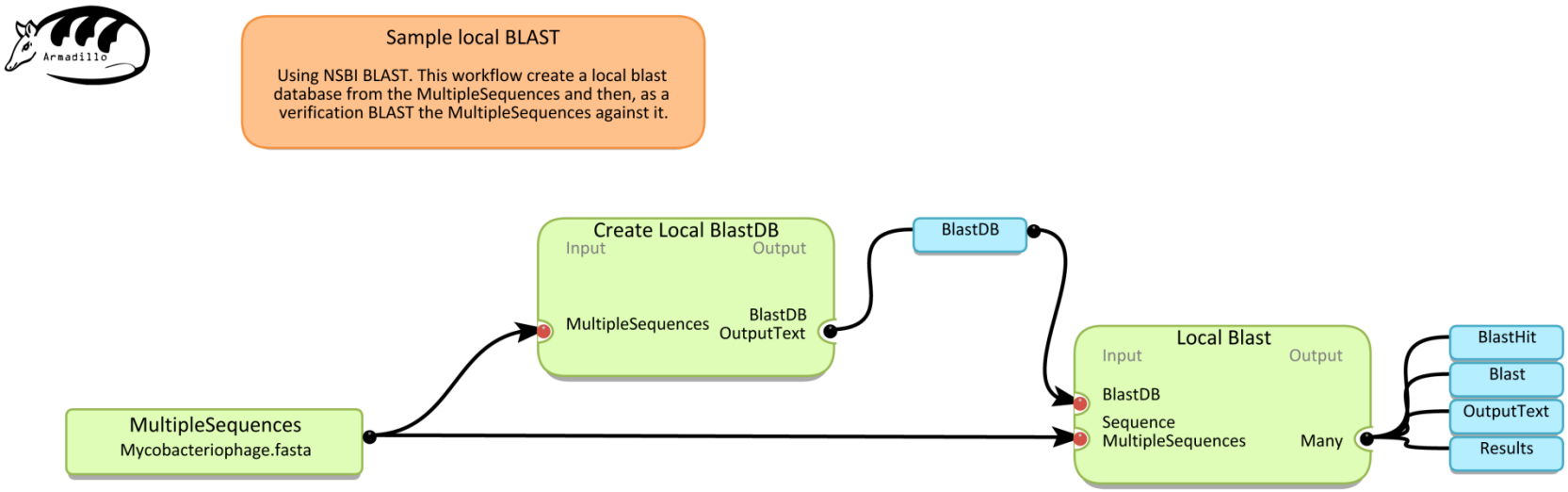
En développement, exécutables sous Docker.io

Armadillo : plusieurs applications incluses …

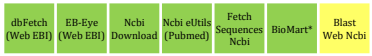


* Les applications sous Linux sont en cours de développement

Démo Armadillo



… et une plate-forme qui pourrait être adaptée à d'autres domaines.



Phylogénétique

En développement
Version *BeanShell*

Une sélection de commandes Linux

cd	ls	rm	cp		mv	chown	chmod	chgrp
pwd	mkdir	rmdir	scp	rcp	mmv	**su / sudo**	date	times
sleep	uniq	nice	slocate	locate	df	bzip2	tar	gzip
man	**help**	read	paste	find	du	sort	tsort	tree
top	ps	bg	cron	crontab	time	env	export	kill
let	declare	set	unset	**uuencode**	**uudecode**	alias	adduser	addgroup
printf	clear	echo	exec	**true**	**false**	wget	awk/ gawk	sed
command	function	if	until	eval	more	wc	grep/ egrep	head
case	for	while	continue	break	less	tail	cat	cut

Problématiques actuelles

Défis actuels (*Romano*, 2008)	2017 (solutions et recherches)
Abstraction limité des types de données	Ontologies bioinformatiques
Accès programmatiques aux données difficiles	Services REST/SOAP
Mauvaises performances des WfMS	Apache Hadoop, etc.
Ressources (applications) utilisables limités	Ontologies? Docker?
Mauvaise sémantique limite la réutilisation	Un langage de workflow commun: CWL
Difficultés à conserver les traces des exécutions	Ontologies?

Romano P. (2008) Automation of in-silico data analysis processes through workflow management systems. Brief Bioinform. 9(1):57-68

Common workflow languge : CWL

https://github.com/common-workflow-language/common-workflow-language

- Arvados Project (Curoverse)
- Broad Institute
- Cincinnati Children's Hospital
- Galaxy Project
- Harvard Chan School of Public Health
- Institut Pasteur
- Oregon Health & Science University
- Sanger Institute
- Seven Bridges
- Taverna
- UC Santa Cruz
- UC Davis

cwltool (reference implementation)
Rabix
Arvados
Galaxy
Parallel Recipes
Toil
CancerCollaboratory
Airflow (SciDAP)

Broad Institute CWL meetup 2015-11-13.pdf

CWL

- Modeled as user defined function (UDF)
- Typed input / output signature
- Inputs & outputs are fully specified
- Tool executions are isolated from one another & from parent process
- Well defined execution process
  1. Collect & validate inputs
  2. Map input file paths to locations inside container
  3. Build tool command line
  4. Build Docker invocation
  5. Execute
  6. Collect & validate outputs

Broad Institute CWL meetup 2015-11-13.pdf

CWL : SamTools (http://samtools.sourceforge.net/)

```
class: CommandLineTool
cwlVersion: draft-3
description: Sort by chromosomal coordinates
```

File type & metadata

```
requirements:
  - class: DockerRequirement
    dockerPull: scidap/samtools:v1.2-216-gdffc67f
```

Runtime environment

```
inputs:
  - id: input
    type: File
    inputBinding:
      position: 1

  - id: output_name
    type: string
    inputBinding:
      position: 2
```

Input parameters

```
outputs:
      - id: output
        type: File
        outputBinding:
          glob: $(inputs.output_name)
```

Output parameters

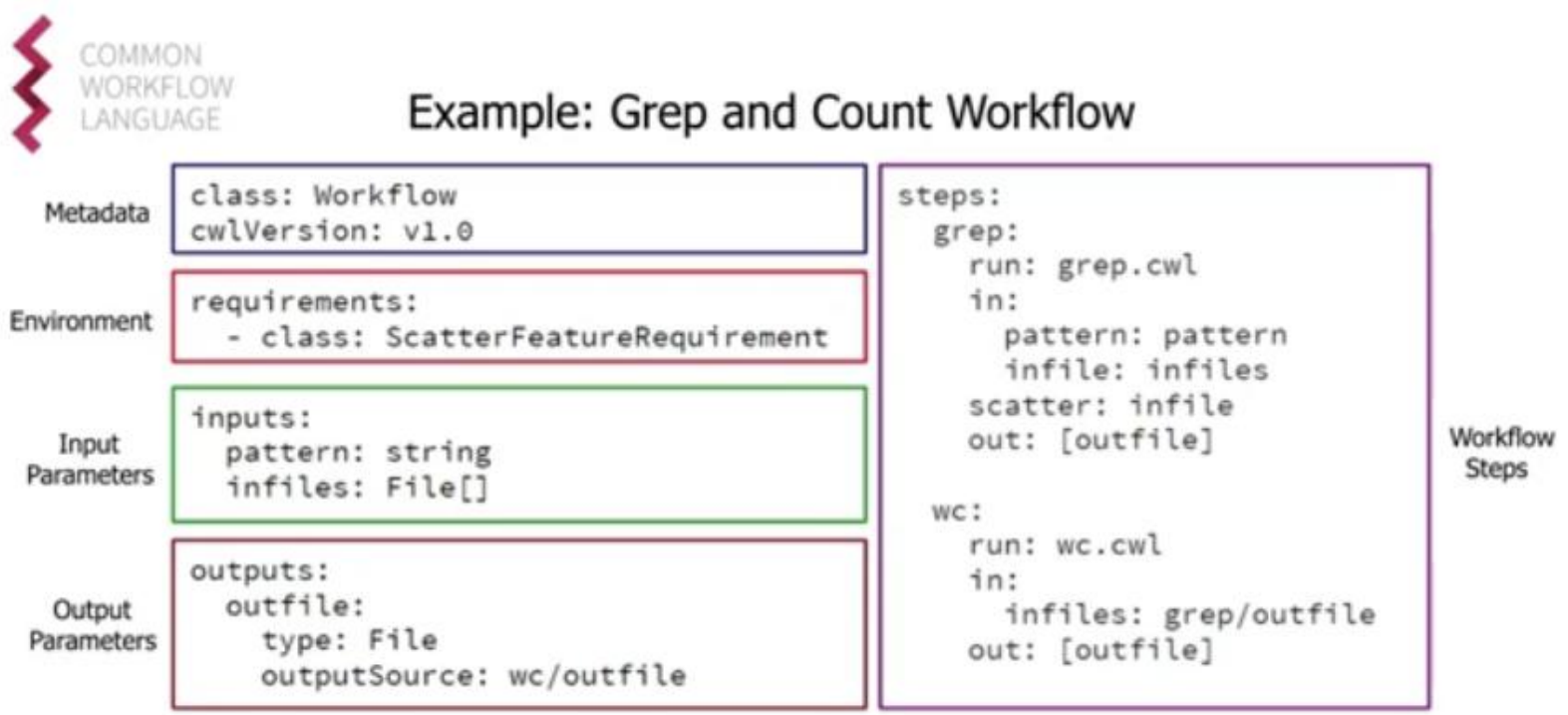
```
baseCommand: [samtools, sort]
```

Executable

Broad Institute CWL meetup 2015-11-13.pdf

Existing workflow system

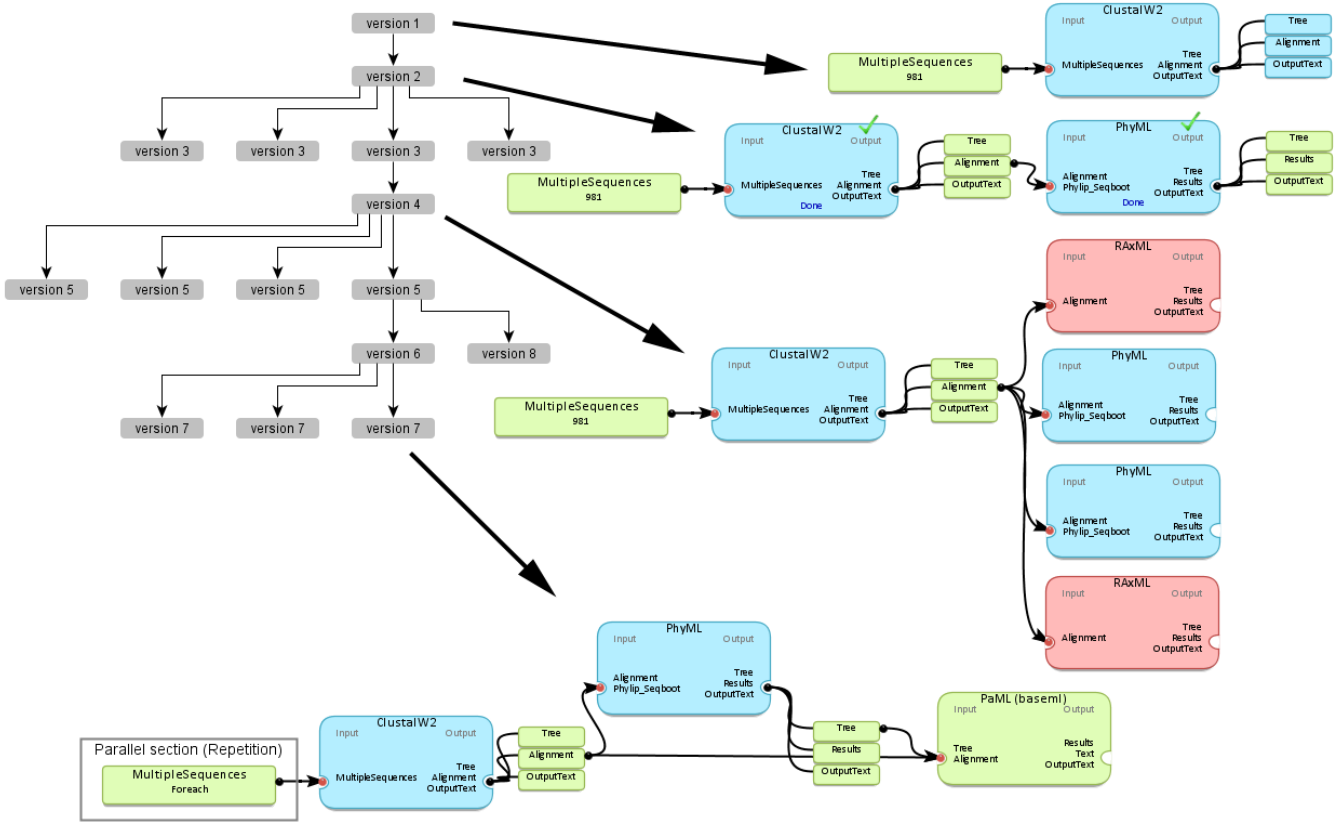
- https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems



https://github.com/common-workflow-language/workflows/blob/master/workflows/presentation-demo/grep-and-count.cwl

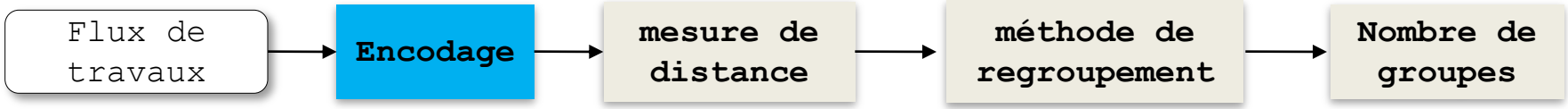
Problématique actuelle (2)

Problématique: une expérimentation *in silico* évolue et résulte en différentes versions d'un flux de travaux.
Comment les comparer?



Lord, E., Diallo, A. B., Makarenkov, V. (**2015**). Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms. *BMC Bioinformatics*, 16(1), 68.

Différents encodages des flux de travaux



Matrice binaire

Encoding of Type I	W1	W2	W3	W4	W5	Weights for Encoding of Type I
Blast (NCBI)	0	0	0	1	0	0.35
ClustalW2	0	1	0	0	1	0.49
HGT Detector 3.2	1	1	1	0	1	0.88
Muscle	1	0	0	0	1	0.41
PROTML (Phylip)	1	0	0	0	0	0.68
PhyML (1)	0	1	1	0	1	1.13
PhyML (2)	0	0	0	0	1	1.13
Probcons	0	0	1	0	0	0.55
Robinson&Foulds distance	0	0	0	1	0	0.25
SEQBOOT	1	0	0	0	0	0.14
Seq-Gen	0	1	0	1	0	0.43

-- Vecteur de temps moyens

Disperser les tâches similaires pour des **exécutions plus rapides**

Encoding of Type II	W1	W2	W3	W4	W5	Weights for Encoding of Type II
Blast (NCBI)	0	0	0	1	0	0.10
ClustalW2	0	1	0	0	1	0.10
HGT Detector 3.2	1	1	1	0	1	1.00
Muscle	1	0	0	0	1	0.10
PROTML (Phylip)	1	0	0	0	0	0.10
PhyML	0	1	1	0	2	0.10
Probcons	0	0	1	0	0	0.10
Robinson&Foulds distance	0	0	0	1	0	0.10
SEQBOOT	1	0	0	0	0	0.10
Seq-Gen	0	1	0	1	0	0.10

-- Vecteur de poids

Grouper les flux de travaux similaires en **fonction de mots-clés**

Matrice d'occurrences

Différents encodages des flux de travaux

Matrice d'occurrences

Encoding of Type III	W1	W2	W3	W4	W5	Weights for Encoding of Type III
Blast (NCBI)	0	0	0	1	0	0.35
HGT Detector 3.2	1	1	1	0	1	0.88
Robinson&Foulds distance	0	0	0	1	0	0.25
ClustalW2→PhyML	0	1	0	0	1	1.62
Muscle→PhyML	0	0	0	0	1	1.54
Muscle→SEQBOOT (Phylip)	1	0	0	0	0	0.55
PROTML (Phylip)→HGT Detector 3.2	1	0	0	0	0	1.56
PhyML→HGT Detector 3.2	0	1	1	0	2	2.01
Probcons→PhyML	0	0	1	0	0	1.68
SEQBOOT (Phylip)→PROTML (Phylip)	1	0	0	0	0	0.82
Seq-Gen→Blast (NCBI)	0	0	0	1	0	0.78
Seq-Gen→ClustalW2	0	1	0	0	0	0.92

Paires de tâches

-- Vecteur de temps moyens

Disperser les tâches similaires pour des **exécutions plus rapides**

Encoding of Type IV	W1	W2	W3	W4	W5	Weights for Encoding of Type IV
Blast (NCBI)	0	0	0	1	0	0.10
HGT Detector 3.2	1	1	1	0	1	1.00
Robinson&Foulds distance	0	0	0	1	0	0.10
ClustalW2→PhyML	0	1	0	0	1	0.10
Muscle→PhyML	0	0	0	0	1	0.10
Muscle→SEQBOOT (Phylip)	1	0	0	0	0	0.10
PROTML (Phylip)→HGT Detector 3.2	1	0	0	0	0	1.00
PhyML→HGT Detector 3.2	0	1	1	0	2	1.00
Probcons→PhyML	0	0	1	0	0	0.10
SEQBOOT (Phylip)→PROTML (Phylip)	1	0	0	0	0	0.10
Seq-Gen→Blast (NCBI)	0	0	0	1	0	0.10
Seq-Gen→ClustalW2	0	1	0	0	0	0.10
INPUT_Sequences	1	0	1	0	1	1.00
INPUT_Tree	1	1	1	2	0	1.00
OUTPUT_Blast (NCBI)	0	0	0	1	0	1.00
OUTPUT_Matrix	1	1	1	1	1	1.00
OUTPUT_MultipleTrees	0	0	0	1	0	1.00
OUTPUT_OutputText	1	1	1	2	1	1.00
OUTPUT_Results	1	1	1	1	1	1.00

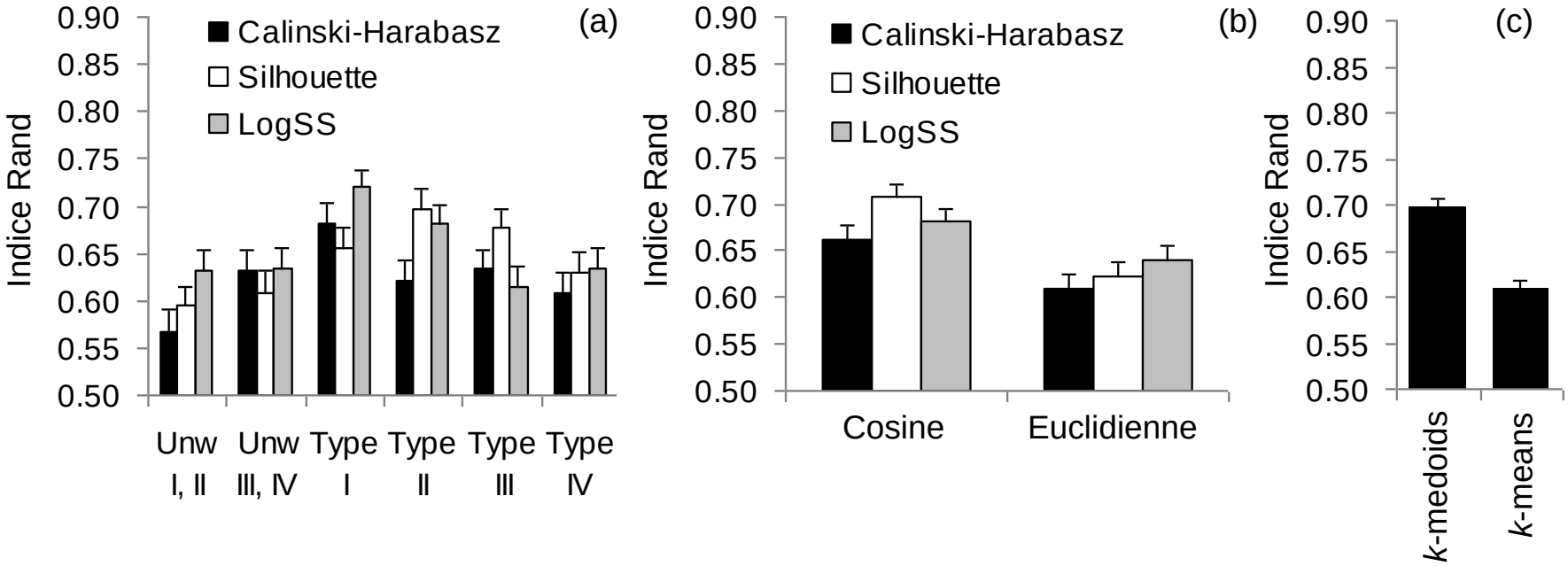
-- Vecteur de poids

Encourager la **réutilisation des données** (génomiques)

Matrice d'occurrences

Conclusions pour le regroupement par partitionnement

(a) Effet des critères d'optimisation pour les encodages non pondérés (les deux premiers jeux de barres) et pondérés (quatre derniers jeux de barres);
(b) Effet de la mesure de distances;
(c) Effet de l'algorithme de partitionnement appliqué;

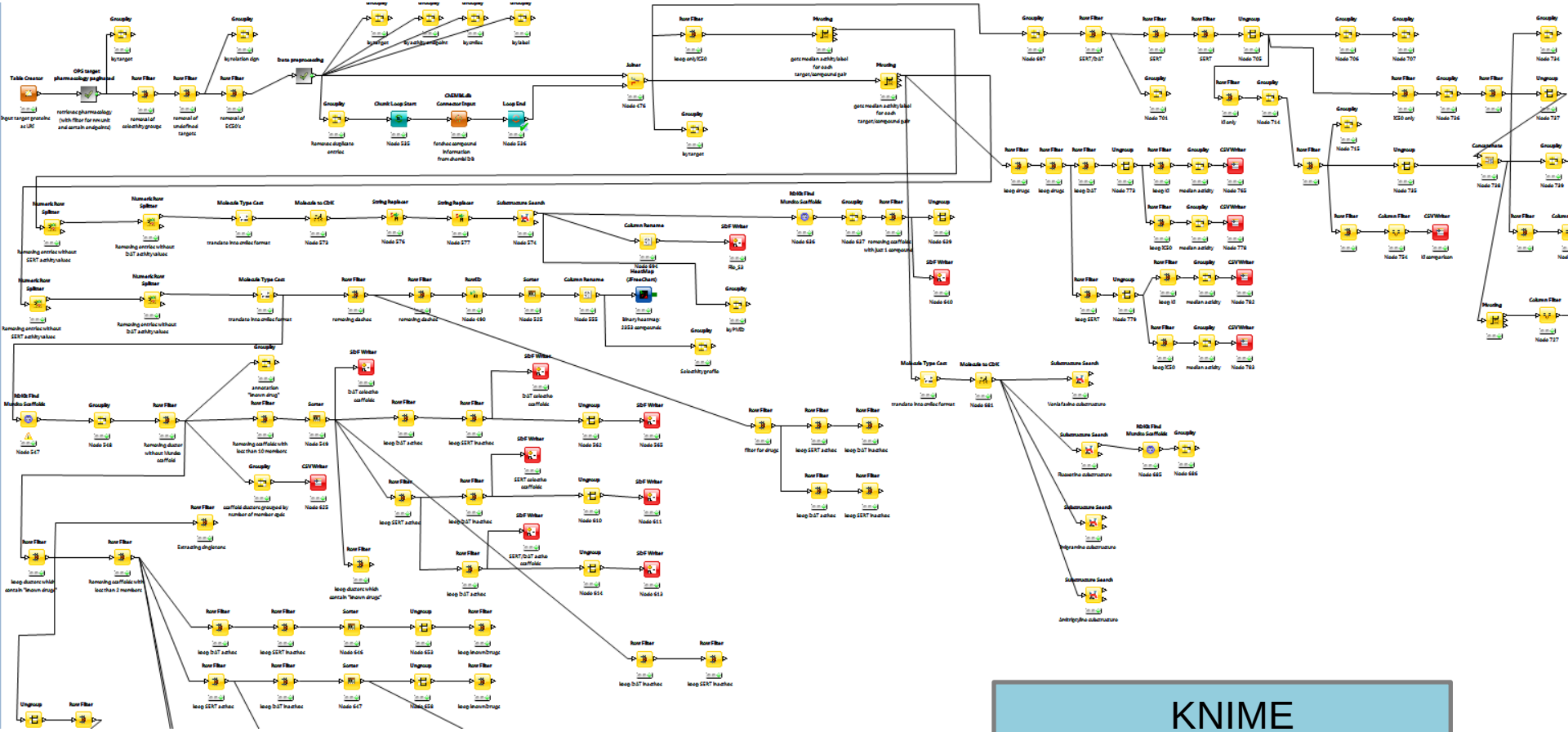


La distance cosine pondérée, utilisée avec l'algorithme *k*-medoids, l'encodage de Type I et l'indice Silhouette, montre la meilleure performance.

N.B. **Les plus grandes valeurs de l'indice Rand sont les meilleures.**

Problématique actuelle (3)

Workflow spaghetti



KNIME

https://www.knime.org/

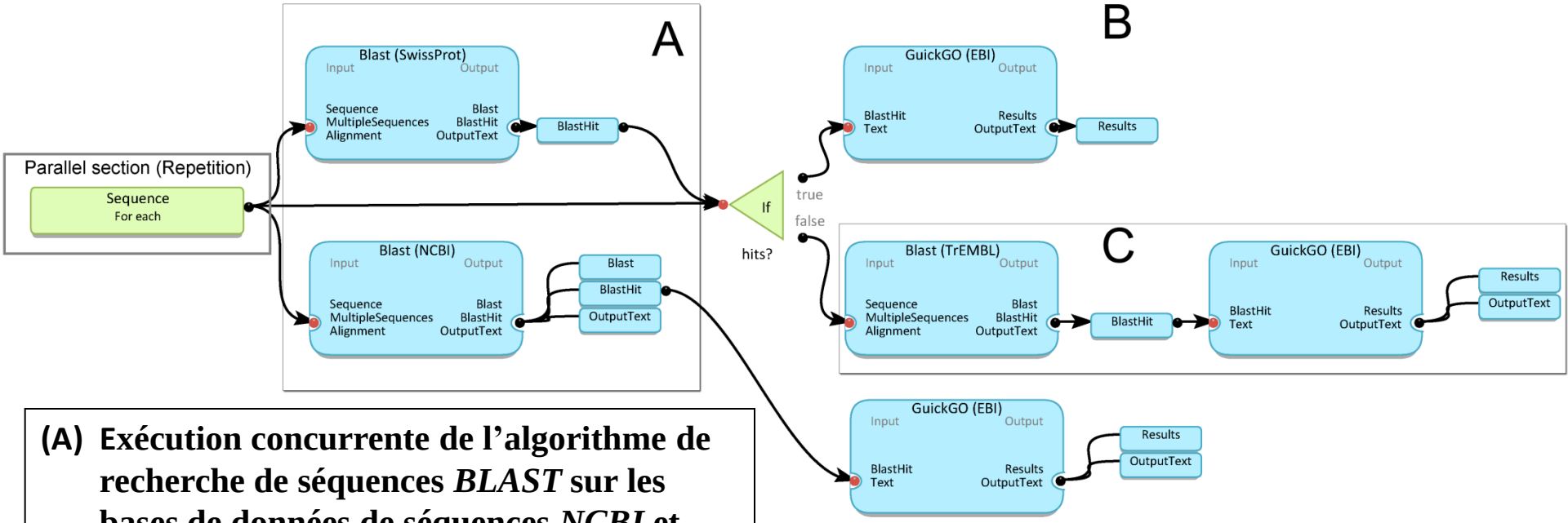
Flux de travaux bioinformatiques

Agharbaoui,Z., Leclercq, M., Remita,M.A., Badawi, M., **Lord, E.**, Houde,M., Danyluk, J., Diallo, A.B., Sarhan, F. (**2015**) An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC Genomics* 16:339

Remita MA, **Lord, E.**, et al. **(2015)** WMP: A novel comprehensive wheat miRNA database, including related bioinformatics software.
doi: http://dx.doi.org/10.1101/024893

Exemple : Flux de travaux utilisés pour annoter les séquences cibles

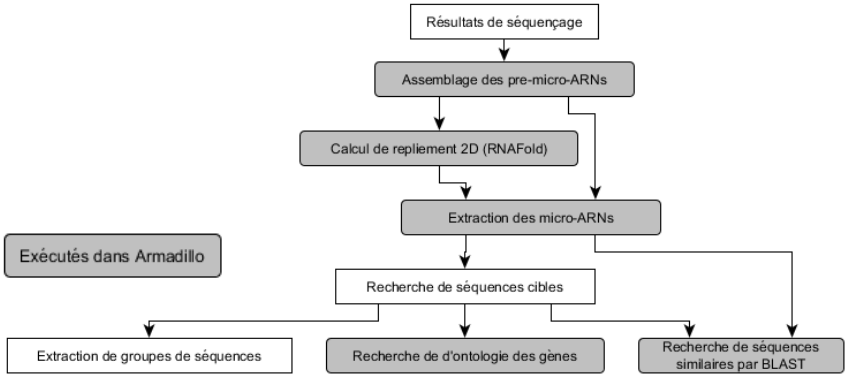
Création d'une banque de données de petits ARNs chez le blé.
http://wheat.bioinfo.uqam.ca/



(A) Exécution concurrente de l'algorithme de recherche de séquences *BLAST* sur les bases de données de séquences *NCBI* et *SwissProt*.

(B) Exécution conditionnelle de la recherche ontologique si on a des résultats en (A).

(C) Recherche par la méthode *BLAST*, mais sur la base de données *TReMBL* et recherche ontologique sur ces résultats, s'il n'y a pas de résultats en (A).

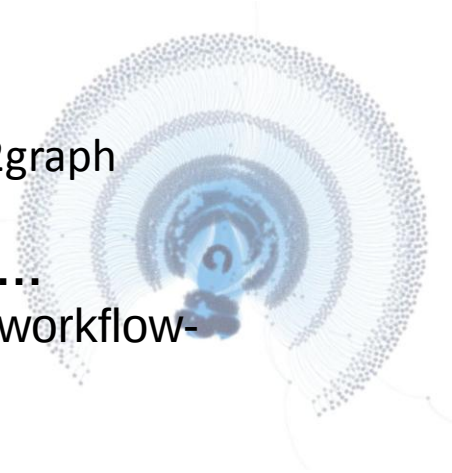


Autres paradigmes : les langages de type *Pipe*

Langages	Exemple	Liens
Switf	`foreach protein in proteinList { runBLAST(protein); }`	http://swift-lang.org/main/
GNU Make	**output.txt:** bin/analyse input.txt bin/analyse input.txt > output.txt	bost.ocks.org/mike/make/ www.bioinformaticszen.com
MakeFlow*	**part1 part2 part3:** input.data split.py ./split.py input.data	ccl.cse.nd.edu/software/makeflow/

Apache Spark	`val text_file = spark.textFile("hdfs://...")`
	`val word_counts = text_file.flatMap(lambda line: line.split()) .map(lambda word:(word,.reduceByKey(lambda a, b: a+b)`
	`val count = word_counts.count()`
		http://spark.apache.org/

Pierre Lindenbaum
(http://plindenbaum.blogspot.ca/)
https://github.com/lindenb/makefile2graph



Encore: une liste des workflows en bioinformatiques…
https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

Autres paradigmes : Big Data

Apache Apex™ Home Documentation Powered By Apex Roadmap Community Github Download

Apache Apex™

Enterprise-grade unified stream and batch processing engine.

*Now with **event-time windowing** and **high-level API**.*

Enterprise Grade

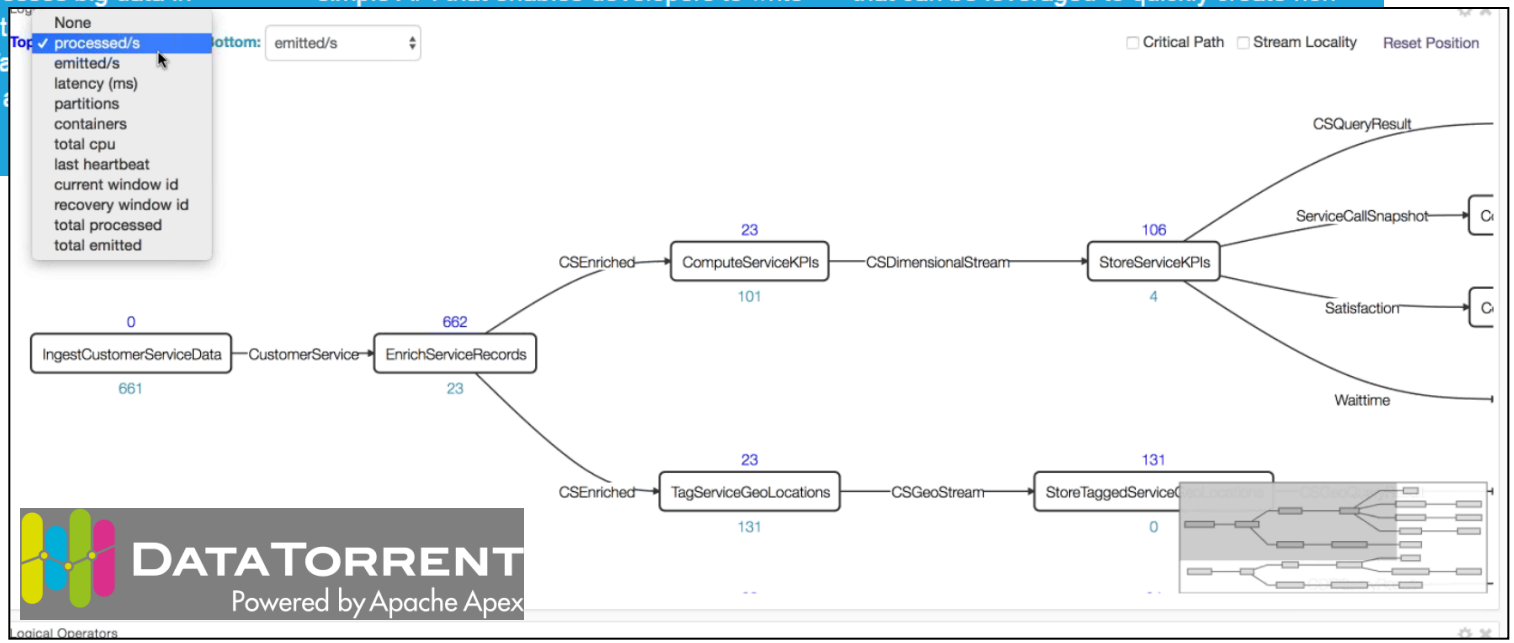
Apex is a Hadoop YARN native platform that unifies stream and batch processing. It processes big data in-motion in a way that
highly performant, fa
secure, distributed,

Low Barrier-to-Entry

Write your business logic and leave all operability to the platform. It provides a simple API that enables developers to write

Modular

The Apex platform comes with Malhar, a library of operators (units of functionality) that can be leveraged to quickly create non-

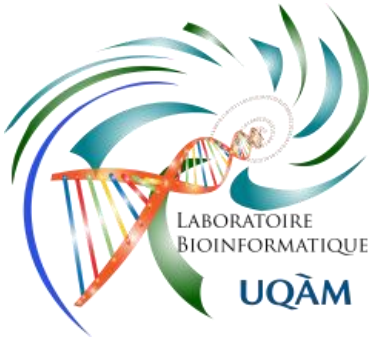


Conclusions

- La plate-forme *Armadillo* ne requiert pas de programmation.
- Elle ne nécessite pas d'applications externes et permet le partage de workflows.
- La plate-forme *Armadillo* pourrait être adaptée à d'autres domaines scientifiques.
- Il reste beaucoup de recherches à effectuer dans le domaine des flux de travaux: ontologie, dispersion des tâches, meilleure abstraction, etc.
- Une meilleure intégration des services webs serait aussi souhaitée ainsi que l'intégration du CWL.

Remerciements

Vladimir Makarenkov, Université du Québec à Montréal
Abdoulaye Baniré Diallo, Université du Québec à Montréal
François-Joseph Lapointe, Université de Montréal



UQÀM

Université de Montréal

Fonds de recherche sur la nature et les technologies Québec

NSERC CRSNG

Références

Lord, E., Leclercq, M., Boc, A., Diallo, A.B., et Makarenkov, V. (2012). Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. *PloS One, 7*(1), e 29903.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., et Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, *18*(10), 1067-1100.

Oinn, T., Li, P., Kell, D. B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turl, D., et Zhao, J. (2007). Taverna/myGrid: aligning a workflow system with the life sciences community. Dans *Workflows for e-Science* (p. 300-319). Springer London.

Sharma, A., Rai, A., et Lal, S. (2013). Workflow management systems for gene sequence analysis and evolutionary studies - A Review. *Bioinformation*, *9*(13), 663-672.

Silva, V., Chirigati, F., Maia, K., Ogasawara, E., Oliveira, D., Braganholo, V., Murta, L., et Mattoso, M. (2011). Similarity-based workflow clustering. *Journal of Computational Interdisciplinary Sciences*, *2*(1), 23-35.