Trends in Bio Data Integration: (Part 1) Scientific Workflows

Sarah Cohen-Boulakia
Université Paris Sud, LRI CNRS UMR 8623
On leave at INRIA Virtual Plants & Zenith, Inst. of Comput. Biology, Montpellier
Introduction

- Data Integration in the Life Science (DILS) is more important than ever
- Portals perform syntaxic integration and are frequently used
- Data warehouses are designed in several places. It remains the most frequently integration solution used in the Life Science community
- Faced with the increasing number of
  - data,
  - sources,
  - analytic tools,
  - and the increasing complexity of analysis pipelines...
  ... challenges are numerous...
The complexity of the questions to be answered has increased a lot
- Integration requires analysis and analysis requires integration
- Scientific workflows

The diversity of the sources has increased a lot
- Inclusion of quality as a first-class citizen
- Ranking of integrated search results

Solutions presented today and next time

The number of sources to be used has increased a lot
- Scalability of integration in number of sources
- One major goal of the Semantic Web, development of ontologies
This Tutorial

- **Part I – Data Integration workflows**
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - Workflows and reproducibility
  - Current challenges

- **Part II – Ranking Biological data**
  - Ranking criteria
  - Introducing ranking into integration solutions
    - Data warehouses
    - Portals

- **Part III – Conclusions**
Scientific Workflow Management System

- SWFS = WFS for scientific tasks
  - “Data analysis pipeline”
- Complex pipelines are broken into tasks and their connection
- Data flow driven
- Tasks can be executed locally or distributed
- SWFS manages scheduling, process control, logging, recovery, reproducibility, ...
- Equipped with graphical workflow designer
- Several systems available (Taverna, Galaxy, Kepler, ...)
Galaxy is an open source, web-based platform for data intensive biomedical research.

The Galaxy Team is a part of
- the Center for Comparative Genomics and Bioinformatics at Penn State,
- the Department of Biology and at Johns Hopkins University.

The Galaxy Project is supported in part by
- NSF,
- NHGRI,
- The Huck Institutes of the Life Sciences,
- The Institute for CyberScience at Penn State,
- and Johns Hopkins University...

Can be used with
- the free public server (usegalaxy.org)
- or other instances (several in France: Institut Curie, Institut Pasteur, Genouest, SouthGreen...)

Sarah Cohen-Boulakia, Université Paris Sud
Pages: documentation within Galaxy. To supplement publications or to present tutorials.

Workflows: define the steps in an analysis process. Workflows are analyses that are intended to be executed (one or more times) with different user-provided input Datasets.

Histories are analyses records in Galaxy that show all input, intermediate, and final datasets, as well as every step in the process and the settings used with each job executed.

Datasets represent individual files or jobs included within a History.

Data Libraries are collections of Datasets accessible. Designed for sharing datasets in between users or groups.
Designing and running a Galaxy workflow

- From scratch
  [http://screencast.g2.bx.psu.edu/flash/WorkflowFromScratch.html](http://screencast.g2.bx.psu.edu/flash/WorkflowFromScratch.html)
  - Start with an input data set (type)
  - Drag-and-drop tools into the working environment
  - Connect tools (green means compatibility)
  - Parametrize tools
  - Upload a data set
  - Run the workflow on the data set

- Extract workflows from histories (reverse engineering)
- ...
Other major workflow systems

- **Taverna** [http://www.taverna.org.uk/](http://www.taverna.org.uk/)
  - One of the first system, Univ. Manchester
  - Perfect to combine *Web services*
  - Linked to several integration projects/solutions: Bioconductor, BioMart...
  - Main positive points: the best concepts (*sub-workflows*, sharing with myExperiment,...), open project

- **VisTrails** [http://www.vistrails.org/](http://www.vistrails.org/)
  - Univ. Utah (New York)
  - Perfect to *visualize and interact* with data sets
  - Main positive point: keep track of the *evolution of the workflow*

- Main negative point for both systems: more advanced research prototypes than systems for use in production
And many others....!

- Kepler (https://kepler-project.org/, BioKepler)
- Pegasus (http://pegasus.isi.edu/, Cloud ++)
- Mobyle (http://mobyle.pasteur.fr/)
- OpenAlea (http://openalea.gforge.inria.fr, Plants ++)
- RapidMiner (https://rapidminer.com/)
- WINGS (http://www.wings-workflows.org/, semantics)
- KNIME (https://www.knime.org/)
- Cunieform (works on Hadoop YARN...)
- ...

Sarah Cohen-Boulakia, Université Paris Sud
This Tutorial

- **Part I – Data Integration workflows**
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - Workflows and reproducibility
  - Current challenges

- **Part II – Ranking Biological data**
  - Ranking criteria
  - Introducing ranking into integration solutions
    - Data warehouses
    - Portals

- **Part III – Conclusions**
Scientific Workflow Repositories

- Upload a scientific workflow
- Search, download & reuse existing scientific workflows
- Most specifically for single workflow system
Scientific Workflow Discovery

- Pose keyword query
- Search in textual annotations
- List of 10s or 100s of workflows
- Find appropriate workflows
- Reuse scientific workflow

Sarah Cohen-Boulakia, Université Paris Sud
myExperiment

- myExperiment.org
- Looking for workflows
  - By keywords
    - BioAID... workflow
    - Inspecting meta-data (author, favourited by, history...)
  - By authors
  - By group
  - ...
  → Demo
BioCatalogue

"The Life Science Web Services Registry"

URL: http://biocatalogue.org

- Registry of **web services** for the life sciences
- Discovery, annotation (provider and user) and uploading new services
  - >1,170 services, >3,000 used in Taverna workflows (BioMart, R, BioMoby, etc.)
Describing one service in BioCatalogue

- For each service
  - Categories (types, ontologies)
  - Annotations (keywords)
  - Country of origin, kind of services (SOAP...), statistics ...

![Image of WSNCBI Blast Service](http://www.ebi.ac.uk/Tools/webservices/WSNCBI Blast.wsdl?kiki)
Looking for services given an input format
This Tutorial

- **Part I – Data Integration workflows**
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - Workflows and reproducibility
  - Current challenges

- **Part II – Ranking Biological data**
  - Ranking criteria
  - Introducing ranking into integration solutions
    - Data warehouses
    - Portals

- **Part III – Conclusions**
Scientific workflow systems
[Element of solution 1 for reproducibility]

- Galaxy, Taverna, WINGS, ...
- Visual programming
  - Chaining tools
- Specification vs executions
- Provenance modules
  - Tracking, logging all data produced and consumed
Virtual machines
[Element of Solution 2 for reproducibility]

- Virtual machines capture the **programming environment**

- A hypervisor (piece of software, firmware or hardware) creates and runs virtual machines.

- A computer on which a hypervisor is running virtual machines = **host machine**

- Each virtual machine = **guest machine**

→ Can be huge (size)
- [https://www.docker.com/](https://www.docker.com/)
- **Aim:** package an application
  - with all of its dependencies
  - into a standardized unit for software development
- **Containers**
  - include the application and all of its dependencies
  - but share the kernel with other containers.
  - They
    - run as an isolated process in userspace on the host operating system;
    - are not tied to any specific infrastructure;
    - run on any computer, on any infrastructure and in any cloud.

➤ **Lighter solution than classical VM**
ReproZip

- ReproZip allows to **pack** any experiment along with all necessary
  - data files,
  - libraries,
  - environment variables and options.

- Contrary to VM, the aim is to store the **minimal amount of files** (based on dependencies)

- **Zip/Unzip** the **running environment**
Notebooks
[element of solution 3 for reproducibility]

IP[y]: IPython Interactive Computing

- Web-based interactive computational environment
- Combination of code execution, text, mathematics, plots and rich media into a single document
- Some systems export workflows as executable IPython papers...
- Very promising!
- Jupyter
Ten Simple Rules for Reproducible Computational Research (PlosOne)

1: For Every Result, Keep Track of How It Was Produced
2: Avoid Manual Data Manipulation Steps
3: Archive the Exact Versions of All External Programs Used
4: Version Control All Custom Scripts
5: Record All Intermediate Results, When Possible in Standardized Formats
6: For Analyses That Include Randomness, Note Underlying Random Seeds
7: Always Store Raw Data behind Plots
8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9: Connect Textual Statements to Underlying Results
10: Provide Public Access to Scripts, Runs, and Results

→ Several ways to follow them
→ More or less complex (from manually to fully automatically)
→ More or less time-consuming (repeat, reproduce, ....., reuse)
This Tutorial

Part I – Data Integration workflows
- What are scientific workflow systems
- Designing a workflow from scratch
- Repositories of workflows and web services (reuse)
- Workflows and reproducibility
- Current challenges and latest results

Part II – Ranking Biological data
- Ranking criteria
- Introducing ranking into integration solutions
  - Data warehouses
  - Portals

Part III – Conclusions
Study on workflow reuse....

- 36% of elements are re-used
  - These connect workflows quite densely
  - Can be exploited for repository IR

- Re-use rates have a Zipf-like distrib
  - Local: High re-use rates as-is
    - Web-Service: Authors have favorite services, unshared
    - Script & subworkflows: Authors have personal libraries

- True cross-author re-use is low: 3%
  - Authors have personal preferences & libraries
  - But don't use content from others

64% of processors used only once
89% used only by one author
How to improve reuse?

Help finding similar workflows

Make workflow structures less complex!

Plumbing workflows
How to improve reuse?

Help finding similar workflows

Make workflow structures less complex!

Plumbing workflows

Sarah Cohen-Boulakia, Université Paris
Scientific Workflow Discovery Improvement

Goal

- Group results by similar workflows
- Search by sample workflow
- Provide recommendations
  - Similar workflows
  - Replacements
  - Extensions
  - ...

Need: Similarity Measures
The Central Question

Functionally similar?

Sarah Cohen-Boulakia, Université Paris Sud
Example
workflows perform microarray analysis integrating various sources (pathway DB, probe mapping, PubMed)

- All three workflows may be used
  - entirely (which fits best?) or partly (from probes to pathways)
Similarity search for scientific workflows

[VLDB 2014]

- **Framework**
  - capture all the sim. search techniques
    - Structure-based
      - Graph struct. of the workflow
    - Annotation-based
      - Meta-data (description, tags...)

- **Goal of the study**
  - compare results obtained by all techniques
  - On various data sets
    - Taverna, Galaxy, VisTrails
Subtasks of Scientific Workflow Comparison

Module Comparison
Subtasks of Scientific Workflow Comparison

- Label
- Webservice Uri
- Scripts
- etc
Subtasks of Scientific Workflow Comparison

Module Comparison

- greedy
- maximum weight
Subtasks of Scientific Workflow Comparison

- Set of Modules
- Substructures
- Full Structure
Subtasks of Scientific Workflow Comparison

- Set of Modules
- Substructures
- Full Structure
Subtasks of Scientific Workflow Comparison

Module Comparison → Module Mapping → Topological Comparison → Normalization

Sarah Cohen-Boulakia, Université Paris Sud
Existing Approaches

- Module Comparison
- Module Mapping
- Topological Comparison
- Normalization

Stoyanovich et al.
Silva et al.
Bergmann et al.
Santos et al.
Goderis et al.
Friesen et al.
Xiang et al.

MCS = Maximum Common Subgraph      GED = Graph Edit Distance

Sarah Cohen-Boulakia, Université Paris Sud
## Existing Approaches

<table>
<thead>
<tr>
<th>Author et al.</th>
<th>Attributes</th>
<th>Mapping Method</th>
<th>Comparison</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stoyanovich et al.</td>
<td>single</td>
<td>-</td>
<td>modules</td>
<td>-</td>
</tr>
<tr>
<td>Silva et al.</td>
<td>multiple</td>
<td>greedy</td>
<td>modules</td>
<td>$</td>
</tr>
<tr>
<td>Bergmann et al.</td>
<td>semantic annot.</td>
<td>max. weight</td>
<td>modules &amp; edges</td>
<td>$</td>
</tr>
<tr>
<td>label edit dist.</td>
<td>max. weight</td>
<td>modules &amp; edges</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>Santos et al.</td>
<td>label matching</td>
<td>-</td>
<td>modules</td>
<td>-</td>
</tr>
<tr>
<td>label matching</td>
<td>-</td>
<td>MCS</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>Goderis et al.</td>
<td>label matching</td>
<td>-</td>
<td>MCS</td>
<td>-</td>
</tr>
<tr>
<td>label matching</td>
<td>-</td>
<td>MCS</td>
<td>'workflow sizes'</td>
<td></td>
</tr>
<tr>
<td>Friesen et al.</td>
<td>type matching</td>
<td>-</td>
<td>modules</td>
<td>-</td>
</tr>
<tr>
<td>type matching</td>
<td>-</td>
<td>MCS</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>type matching</td>
<td>-</td>
<td>graph kernels</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Xiang et al.</td>
<td>label matching</td>
<td>-</td>
<td>GED</td>
<td>-</td>
</tr>
</tbody>
</table>

**MCS** = Maximum Common Subgraph  
**GED** = Graph Edit Distance
## Existing Approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Attribute Handling</th>
<th>Similarity Measure</th>
<th>Workflow Size</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stoyanovich et al.</td>
<td>single attributes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Silva et al.</td>
<td>multiple attributes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bergmann et al.</td>
<td>semantic attributes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Santos et al.</td>
<td>label matching</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Goderis et al.</td>
<td>label matching</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Friesen et al.</td>
<td>type matching</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xiang et al.</td>
<td>label matching</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**What's best**

At each step?  
As a whole?

*MCS* = Maximum Common Subgraph  
*GED* = Graph Edit Distance
Expert Curated Similarity Corpus

FlowAlike — Scientific Workflow Similarity Evaluation

Reference workflow:

**EBI_Kalign**

Multiple sequence alignment using the Kalign tool. This workflow uses the EBI's WSKalign service (see http://www.ebi.ac.uk/Tools/webservices/services/kalign) to access the Kalign tool. The set of sequences to align and the molecule type (protein or nucleic acid) are the input, the other parameters for the search (see Job_params) are allowed to default.

**Note:** the WSKalign service used by this workflow is deprecated as of 21st September 2010 and should not be used in any new development. This service is will be retired during 2011. EBI's replacement Kalign services (REST or SOAP) should be used instead.

Are these 10 workflows similar to the reference?

**EBI_NCBI_BLAST_with_prompts**

Run a BLAST analysis using the EBI's WSNCFBlast service (see http://www.ebi.ac.uk/Tools/webservices/services/ncbistat). This workflow wraps the EBI_NCBI_BP workflow to provide a basic...

**EBI_InterProScan**

Note: the WSInterProScan web service used by this workflow is no longer available having been replaced by the EMBL-EBI's InterProScan (REST) (http://www.ebi.ac.uk/Tools/webservices/services/pfa/prscan_rest)

**BioQuali asynchronous workflow**

BioQuali: Network Compatibility and products variation inference in a biological network.

Sarah Cohen-Boulakia, Université Paris Sud
Expert Curated Similarity Corpus

**FlowAlike** — Scientific Workflow Similarity Evaluation

Reference workflow:

**EBI_Kalign**

Multiple sequence alignment using the Kalign tool. This workflow uses the EBI’s WSKalign service (see http://www.ebi.ac.uk/Tools/webservices/services/kalign) to access the Kalign tool. The set of sequences to align and the molecule type (protein or nucleic acid) are the input, the other parameters for the search (such as genome) are allowed to default.

Are these 10 workflows similar to the reference?

**EBI_NCBI_BLAST_with_prompts**

Run a BLAST analysis using the EBI’s WSNCEBIBlast service (see http://www.ebi.ac.uk/Tools/webservices/services/ncbi/blast). This workflow wraps the EBI_LIBREF_kflow to provide a basic...

---

Sarah Cohen-Boulakia, Université Paris Sud
Expert Curated Similarity Corpus

FlowAlike — Scientific Workflow Similarity Evaluation

Please choose a reference workflow to rate similar workflows against:

- 24 query workflows
- Each with 10 other workflows to compare to it
  - very similar
  - similar
  - related
  - unsure
  - dissimilar
- + Extended comparison lists for specific algorithms' results for 8 query workflows
- 15 experts (7 institutes) provided > 2400 ratings
  - classifying each pair of workflows
  - ranking workflow lists by similarity
Experts agreed on the similarity of workflow pairs

Annotation-based approaches
- Provide best results
- But only a few well-annotated workflows

Structural approaches
- Outperform annotation-based
  - Galaxy & VisTrails
- Graph edit distance is too expensive
- Module set provides good results
- Room for solutions in between
  - LayerDecomposition [eScience 2014]
    with J. Starlinger, U. Leser, S. Davidson, S. Khanna
  - Usable in real environments (myExperiment)
    [Future Generation Computer System 2016]
How to improve reuse?

Help finding similar workflows

Make workflow structures less complex!

Plumbing workflows

Sarah Cohen-Boulakia, Université Paris
DistillFlow

- Distilling workflow structures: Removing redundancy
- Collaboration with Taverna & BioVel
- BioVel (FP7)

- Virtual laboratory: Libraries of workflows for research on biodiversity
- Consortium of 15 partners (9 countries)

→ Improving reuse in BioVel
→ More generally: improving reuse in Taverna
Use case 1

3 processors duplicated! → Pure redundancy

Equivalence:

Same input

No redundancy
Use case 2

Workflow (ii) uses

merge

split

\[ [d_1, d_2] \]
Rewriting workflows

- Exploting the implicit iteration feature of Taverna
  - List of items with merge/split instead of single items with duplication

- Assumptions before merging several copies of a processor
  - Only copies with the exact same code
  - Only copies that do not depend on each other
  - Only deterministic processors (same input \(\rightarrow\) same output)

\(\rightarrow\) 2 anti-patterns and the corresponding rewriting
Anti-pattern (A)

$L_i$ can be one single value or a list of values

Corresponds to use case 1
Anti-pattern (B)

Processor P applies **cross product** to values on ports \(a_1\) to \(a_t\) and **dot product** to values on ports \(a_{t+1}\) to \(a_k\)

Corresponds to use case 2

Sarah Cohen-Boulakia, Université Paris Sud
Provenance in scientific workflows

- Provenance is highly important for users to interpret any scientific result
- Workflow systems are now equipped of *Provenance Modules* capturing the exact set of data used and consumed by the execution of each workflow step
- **Standards** to represent provenance information are now defined (W3C)
- One of the major challenge lies in dealing with the huge amounts of information
  - Example of solution with ZOOM*userviews which use the composition to hide (part of) the data
What is the difference between these two experiments? How these data have been generated? With which input data? Which tools? Which parameters?

Public sources
- Distributed
- Heterogeneous
- Network

Tools
- Distributed
- Heterogeneous
- Chained

Tools
- JAVA, Perl
- Web services
- Python
- Scripts

Workspace

Sarah Cohen-Boulakia, Université Paris Sud
Nodes: Modules + Inputs, Outputs

Edges: Possible dataflow

[Source: K. Sjölander, Bioinformatics 2004]
Workflow run

Specification

Nodes: Steps (executions of modules)

Edges: Actual dataflow (labelled with data object ids)
Workflow run: Provenance of d447?
(tree generated)
Workflow run: Provenance of d447 (immediate)

Specification
Workflow run: Provenance of d447 (deep)

Provenance overload!

Need to focus on relevant information
Composite modules

User view

Sarah Cohen-Boulakia, Université Paris Sud
Composition simplifies provenance!
Designing composite modules

- Composite modules are typically defined **by the workflow designer** to
  - Enable **reuse** between workflows
  - **Simplify** the view of the workflow according to what modules the **designer thinks are relevant** in the workflow

- However, users may have **different interests**, i.e. have different relevant modules

  ➔ **Several user views** of a given workflow should thus be considered, constructed according to each user’s interest
Each composite module takes the meaning of the relevant module it contains.
User views may differ

Adapt to user needs

Information generated by executions of M5 should be visible for Mary

Mary’s view

Sarah Cohen-Boulakia, Université Paris Sud
Grouping may be error-prone!

Grouping should **preserve the relationships between relevant modules**

The *annotation checking* module does not need input from the *run alignment* module!

**Monica’s view**

Sarah Cohen-Boulakia, Université Paris Sud
Constructing User Views (1/3)

- **Goal:** construct a set of composite modules forming a relevant user view given
  - A workflow definition
  - A set of relevant modules

- What properties should hold in the user view?

  (1) **At most one relevant module** per composite module

  ➔ The composite module takes the meaning of the relevant module it contains

M10 is seen as M3 (“run alignment”) + some formatting tasks
Constructing User Views (2/3)

- **Goal:** construct a set of composite modules forming a **relevant user view** given
  - A workflow definition
  - A set of relevant modules

- What properties should hold in the user view?

**2. The induced workflow should be complete w.r.t. precedence**

For every path between relevant modules in the specification there is a path between relevant composite modules in the user view.

In the induced workflow, M4 cannot be executed without having executed M1 and M5 (M3): this is not true in the original workflow specification.
Constructing User Views (3/3)

- **Goal:** construct a set of composite step-classes forming a **relevant user view** given
  - A workflow definition
  - A set of relevant modules

- What properties should hold in the user view?

*(3) The induced workflow should preserve precedence ➔ It should not introduce a new path between relevant modules*

In the induced workflow, there is a path from M1 to M5: violates precedence
Goals

- Help user **construct relevant user views**
  - Preserving the **relationships** between **relevant modules**
- Exploit **user views** to **reduce the provenance information** returned as answer to a query

Contributions

- **Model** for provenance and user views in scientific workflows
- **Algorithm (polynomial)** for generating **relevant** user views according to the user’s interests (minimal)
- **Provenance Reasoning system:** Querying provenance through user views
Provenance challenge

- First Provenance Challenge ([twiki.ipaw.info](twiki.ipaw.info))

- **Aims:** Understanding the capabilities of provenance-related systems (17)

- The challenge process
  - Workflow example (spec + run) provided
  - List of provenance queries to be answered
Terminology
- Nodes are **step-classes** (static)
- Edges capture the **flow of data** between step-classes
- An **execution** of a workflow generates a partial order of steps (dynamic)
  - Instances of step classes
- Each step has **input** and **output** data
Provenance

Public sources
- Distributed
- Heterogeneous
- Network

What is the difference between these two experiments?
How these data have been generated? With which input data? Which tools? Which parameters?

Tools
- Distributed
- Heterogeneous
- Chained

Tools
- JAVA, Perl
- Python
- Scripts
- Web services
...

Workspace

Sarah Cohen-Boulakia, Université Paris Sud
What’s the difference between these two runs of the same workflow?

Our problem is more than a “spot the difference” puzzle!
What’s the difference between these two runs of the same workflow?

Our problem is more than a “spot the difference” puzzle!

Mapping different objects in two figures is trivial.
What’s the difference between these two runs of the same workflow?

Mapping different fork or loop copies in two runs is nontrivial!

Our problem is more than a “spot the difference” puzzle!

Mapping different objects in two figures is trivial.
The problem of differencing runs is NP-hard on DAGs while polynomial time algorithms can be designed for Series-Parallel (SP) structures. Some approaches have considered such restrictions on workflow graph structures.
Definition of SP-graphs

G is SP iff \( \text{MaxRed}(G) = \text{BSP} \)

- \( \text{MaxRed}(G) \): iteratively performs series and parallel reductions on a given graph G

- \( \text{BSP} \): Basic Series-Parallel

Sarah Cohen-Boulakia, Université Paris Sud
G is SP iff
\[ \text{MaxRed}(G) = \text{BSP} \]

- **MaxRed(G):** iteratively performs series and parallel reductions on a given graph G
- **BSP:** Basic Series-Parallel

Sarah Cohen-Boulakia, Université Paris Sud
Paral red \((s,1)\)

3 Series red
\((2,3)(3,6)\)
\((2,4)(4,6)\)
\((2,5)(5,6)\)

Series red
\((s,1), (1,2)\)

Paral red \((2,6)\)

Series red
\((2,6)(6,t)\)

Paral red \((2,t)\)

Series red
\((s,2)(2,t)\)

Paral red \((2,t)\)

Series red
\((2,6)(6,t)\)
G is SP iff
MaxRed(G) = BSP

- MaxRed(G): iteratively performs series and parallel reductions on a given graph G
- BSP: Basic Series-Parallel

Is it Series-Parallel?
Is it Series-Parallel?

... Another definition of series-parallel graphs?

NO!
Another definition (Non SP-graphs)

G is non-SP iff MaxRed(G) contains $G_{\text{forbidden}}$

Intuitively, such graphs cannot be synchronized

$v$ and $w$ are called reduction nodes

Subgraph isomorphism is polynomial for SP graphs
SPFL-Workflow Model (PDiffView)

- Workflow Specification
  - A series-parallel graph overlaid with well-nested fork and loop subgraphs
  - Four kinds of executions: series, parallel, fork and loop

![Diagram of workflow model with series execution, parallel execution, fork execution, and loop execution.]
Valid Runs

- Derived from the specification by applying series, parallel, fork and loop executions recursively
Edit Operations

Path Insertion, Deletion, Expansion, Contraction

- **Elementary path**: each internal vertex has exactly one incoming edge and one outgoing edge, and the resulting graph is still valid with respect to the specification.
- Three motivating principles
  1. They preserve the validity of the run
  2. They are atomic
  3. They are complete

R₁ Delete 1→2→4

R₂ Insert 1→2→4

Expand 1→2→4
Problem statement

Given a pair of valid runs $R_1$ and $R_2$ of the same specification, and a cost function, compute a minimum cost edit script that transforms $R_1$ to $R_2$. The cost of this edit script is also known as the edit distance between $R_1$ and $R_2$. Polynomal-time algorithm designed in PDiffView for SPFL workflows.
Conclusion on workflows

Workflows plays a crucial role in biological data integration

Various areas of computer sciences are involved
  ◦ Databases (e.g., to query and store them)
  ◦ Software engineering (e.g., to optimize or rewrite them)
  ◦ Graph algorithmics (e.g., to query and compare them)
  ◦ ... and a lot of other optimization techniques

Very large spectrum of challenges
  ◦ From very theoretical (e.g., graph theory, equivalence of programs) to very technical and practical (user study, benchmarking on real data sets...)

Sarah Cohen-Boulakia, Université Paris Sud
To go further on workflows

* (SCB1) workflows citation
  - Paper: tapp14_paper_koop.pdf
  - Tool: myexperiment.org

* (SCB2) Ontologies associated with bioinformatics tools and workflows
  - Paper: sciencedirect.com
  - Tool: biocatalogue.org

* (SCB3) Notebooks and reproducible papers
  - Paper: shen_2014_nature.pdf
  - Tool: jupyter.org

* (SCB4) Reproducibility of bioinformatics experiments
  - Paper: plos.org
  - Tool: reprozip

* (SCB5) Provenance in scripts
  - Tool: yesworkflow-org

* (SCB6) Using workflows to perform phylogenetic analyses
  - Paper: biomedcentral
  - Tool: phylogeny@lirmm

* (SCB7) Using workflows to perform NGS analyses
  - Tool: Galaxy-for-NGS-Analysis

* (SCB8) Searching for patterns in scientific workflows
  - Paper: ...
  - Tool: FragFlow