Deliverable D7.1

A library of workflows and workflow nodes

Planned delivery date (as in DoA): M12

Actual submission date: 30/11/2018, month M12

Workpackage: WP7

Workpackage leaders: Carole Goble and Jean-Loup Faulon

Deliverable leader: Jean-Loup Faulon

Version: 1.0

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PU Public</td>
<td>X</td>
</tr>
<tr>
<td>CI Classified, as referred to Commission Decision 2001/844/EC</td>
<td></td>
</tr>
<tr>
<td>CO Confidential, only for members of the consortium (including the Commission Services)</td>
<td></td>
</tr>
</tbody>
</table>

This deliverable is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730976.
# TABLE OF CONTENTS

1 Summary 3
2 Introduction 3
3 Results 4
  3.1 IBISBA Workflow Nodes 4
    RetroRules node 5
    RetroPath2.0 node 6
    ScopeViewer node 7
    RP2paths node 9
    Metabolomics node 10
  3.2 Workflows 10
    Enumerating Muconic Acid Pathways workflow 11
    Metabolomics workflow 12
  3.3 A Workflow Library in IBISBAHub 13
  3.4 Workflows and Workflow nodes in CWL 13
    Script-based nodes as containers 13
    KNIME-based nodes as containers 14
  3.5 SOP development 14
4 Conclusion and future work 15
5 Partners involved in the work 16
6 References 16
1 Summary

The following IBISBA workflow nodes have been developed: RetroRules [1], RetroPath2.0 [2], ScopeViewer, RP2Paths, and Metabolomics [3]. The RetroRules node pulls reaction rules from the RetroRules.org SQL database and parses them into a RetroPath2.0 friendly format. The RetroPath2.0 node is a retrosynthetic algorithm that calculates the possible pathways between a given source and one or more sink molecules. ScopeViewer and RP2Paths are two nodes that take input from RetroPath2.0. The former presents all the metabolic pathways that were calculated and the latter enumerates all the individual pathways. Lastly, the Metabolomics node compares the output of RetroPath2.0 (run in a forward manner) and measured mass spectrometry data by assigning the theoretical mass of the metabolites generated to the mass spectrometry peaks.

To illustrate the node’s ability to be combined in novel manners, two different workflows have been worked out: a pathway enumeration workflow for the production of muconic acid, and a metabolomics workflow.

The pathway enumeration workflow is a classic example of metabolic engineering using the retrosynthetic power of RetroPath2.0. In this workflow, reaction rules are extracted from the RetroRules database and then used by RetroPath2.0 using muconic acid as the source and yeast metabolome as the sink, calculating the metabolic network that connects the two. This network may be inspected using the ScopeViewer node, and the individual pathways enumerated using the RP2Paths node.

The Metabolomics workflow runs RetroPath2.0 in a forward manner and leverages the promiscuous nature of enzymatic reactions to uncover new metabolites from a sink (in this case the metabolome of yeast). These results are compared with the measured mass spectrometry data that identifies peaks from the theoretical mass of molecules and labels those peaks that have not been defined.

We have started to register the above Workflow nodes and Workflows in the IBISBAHub centralised catalogue (to be reported in Deliverable 7.2), and have encoded some of Workflow nodes in the Common Workflow Language (CWL). In collaboration with partners of WP6, we have also started to collect Standard Operating Procedures (SOPs) and to register these in IBISBAHub.

2 Introduction

The main goal of the workpackage (WP7) is to investigate state-of-the-art IT infrastructure needed for multi-partner, multi-site, multi-component industrial biotechnology projects. The workpackage has four objectives and the deliverable is linked to the first.

- To use portable computational workflows to combine multiple computational resources and automate analytics using a state of the art workflow interoperability language to ensure generality across workflow management systems.

- To coordinate workflows with manual steps using programmatically encoded “Business Process Modelling” protocols. These BPM workflows are higher-level systematic descriptions of protocols and produce auditable provenance of the results of their execution.
To use rich metadata to describe and retain a common, related view of the components of a project: workflows, data, models, SOPs, samples, people etc. This work is done in collaboration with WP Network.

To provide an online repository and portal for storing, sharing and launching workflows, linked with related data and models, organised using the de-facto ISA standard (Investigation, Studies, Assays/Analysis) structure.

The nodes and workflows we are developing are to be used to execute and monitor the Design-Build-Test processes of WP6. There are essentially two types of nodes: computational for the Design step and the analysis of data acquired in the Test step, and experimental for the Build and Test steps. These steps conform to Standard Operating Protocols (SOPs) and are executed through human intervention by the infrastructure partners of IBISBA.

We report here a list of Workflow nodes (3.1) and Workflows (3.2), their implementation in IBISBAHub (3.3) the encoding of Workflows and Workflow nodes in CWL (3.4) and initial SOPs stored in IBISBAHub (3.5).

3 Results
The outcome products developed here are a series of IBISBA Workflows comprised of different IBISBA nodes. A node in this context is a strictly defined piece of code with well-defined inputs and outputs, which performs either data import, data processing, calculations, modelling or data visualisation. If the input/output data types and data structure of different nodes match, then they may be stringed together to form workflows (also called data pipelines). Due to the modularity of the nodes, multiple novel workflows may be produced from the same nodes. Here we present a series of nodes and workflows that have been developed for the purposes of metabolic engineering.

3.1 IBISBA Workflow Nodes
To support computational analysis, IBISBA is collecting and designing a toolkit of IBISBA Workflow Nodes, reusable steps that can be combined for a particular computational analysis to be executed as an IBISBA Workflow. In the simplest an IBISBA Workflow Node can be seen as wrapping a single existing tool or algorithm (e.g. open source software created by a third-party), but adding pre- and post-processing of data, as well as relevant configuration and visualization so that the IBISBA Workflow Nodes can be more readily combined on the fly.
A longstanding problem in retrosynthesis is the limited supply of readily accessible reaction rules [1], which allows one to build retrosynthetic maps to ultimately find pathways linking available materials (sink) to targets (source). To address this problem, the RetroRules database was developed and populated through the meta-analysis of multiple online biological pathways databases, including: MetaNetX [4], Brenda [5], BiGG [6], KEGG [7], MetaCyc [8], Reactome [9], Rhea [10], SABIO-RK [11] and The SEED [12]. The resulting RetroRules database contains more than 400,000 reactions rules, expressed in the SMARTS (SMIRKS) format. A reaction rule describes a mono-component reaction that is uniquely characterised by three key features:

1. The reaction ID from which the reaction rule has been computed.
2. A singular substrate ID of said reaction considering we are dealing with mono-component reactions.
3. A flag indicating whether the rule contains information on the stereochemistry of the reaction.

One feature that makes this database unique is the incorporation of reaction rules that describe the chirality of the molecular transformations as well as the atom-atom mapping. At the time of writing, this is the first openly accessible database to provide this type of information [1]. The node does not currently expose the descriptions of chirality; however, due to its active development of the node, more features such as chirality will be added.
RetroPath2.0 node

The RetroPath2.0 node was developed for the exploration and discovery of novel enzymatic networks. It was built as a KNIME workflow, using KNIME’s integration with the cheminformatics software package RDKit [13], to automate retrosynthetic and synthetic analysis from the application of generalised reaction rules.

This workflow may be used in a forward or in a reverse manner, depending on the needs of the user. The reverse application refers to default retrosynthetic exploration of a given sink compound to a defined target (Figure 2). The provided example illustrates this feature with the exploration of possible metabolic routes for the production of a muconic acid in yeast. For such a scenario the RetroPath2.0 node requires three input types:

1. A chemical species that we would like to reach using retrosynthesis, also called the source compound (Example: muconic Acid)
2. One or more sink compounds from which we begin (Example: yeast metabolites)
3. A list of rules that dictate the chemical transformations (the RetroRules)

The forward use of this algorithm refers to the “synthesis” of possible compounds that may be produced from a set of reaction rules (Figure 2). This forward use is typically used for exploratory means, where one would like to find all possible compounds that can be produced from a source. To do this, one needs to provide the following files:

1. A “source” file with one or more compounds to start the exploration with
2. A list of reaction rules (RetroRules)

RetroPath2.0 outputs the following files:

Figure 2: RetroPath2.0 input and output. The user may run RetroPath2.0 in a forward or reverse manner, where in both cases a list of reaction rules (rules.csv) and “source” compounds must be provided. If one wishes to perform retrosynthesis, a list of compounds to reach is required. If one wants to explore all possible compounds that one can produce applying the rules, the node may be run in a forward manner and a sink file is not needed (dashed lined box).
1. **results.csv**: contains all the calculated reactions including those that do not lead to the source compound
2. **scope.csv and scope.json** (both contain the same information in different formats): contains only the metabolic network that links to the source from one or more sink compounds.
3. **source-in-sink.csv**: contains the compounds that are specified in the source that are also in the sink. This is used for debugging.
4. Graphical representations of the chemical structures of the calculated intermediates as SVG files

**ScopeViewer node**

For a more user friendly output of RetroPath2.0, the ScopeViewer node was developed that graphically display the calculated metabolic network in an interactive fashion. Upon opening a static HTML page, the user uploads the scope.json output and all the associated SVG files generated by RetroPath2.0. An example output for identification of pathways for synthesis of muconic acid from metabolites available in yeast (sink) is shown in Figure 3 and Figure 4. The user can inspect individual reaction steps (circles) or compounds (squares), with the right hand panel displaying additional information that may be useful to select the most appropriate metabolic route.

![ScopeViewer output](image)

**Figure 3**: ScopeViewer output of designed pathways for the muconic acid production in yeast, showing the right hand side panel information when one selects a reaction
Figure 4: ScopeViewer output of designed pathways for the muconic acid production in yeast, showing the right hand side panel information when one selects species.
RP2paths node

Figure 5: RP2paths diagram. The output scope.csv from RetroPath2.0 is used as input and enumerates the pathways.

Figure 6: Example output for RP2paths where the source is muconic acid. This particular metabolic path contains two steps with three reactions and two cofactors. The reactions use the MetaNetX identifiers and the document contains hyperlinks that point to their appropriate webpage to provide the user with more details.

ScopeViewer enables a user to explore the whole metabolic space that is associated with a particular desired source. However, because RetroPath2.0 crosslinks the pathways sharing common molecular species, it can be difficult to identify individual metabolic pathways that link the source to a given (combination of) sink(s). To this end, the RP2paths node was developed to take the output source scope of RetroPath2.0 and enumerate all the possible combinations to return only single pathways (as illustrated in Figure 5). Figure 6 illustrates a single pathway output example from the muconic acid run, showing the structures of the intermediate steps and the co-factors for each step.
Mass spectrometry is a tool frequently used to study the metabolome of cells, while RetroPath2.0 may be used in a forward manner to compute the promiscuous nature of enzymes and calculate all the theoretical possible chemical species and their intermediates from a given source. The metabolomics node enables one to compare the theoretical molecular weight of a compound with a singular measured mass spectrometry peak.

The node is written in KNIME and uses the OpenMS package to parse the mass spectrometer file inputs [3].

3.2 Workflows
One of the advantages of defining the tasks as IBISBA Workflow Nodes is their interoperability and new ways in which they may be combined to generate innovative functions. To illustrate some of the novel ways one can link together IBISBA Workflow Nodes, we provide here two different examples. The first is the exploration of heterologous metabolic pathways for the production of muconic acid in yeast. The second is the use of RetroPath2.0 in a forward manner to generate possible metabolites that have not yet been characterised, and compare them with the mass spectrometry measured metabolome of an organism using the metabolomic node.
Enumerating Muconic Acid Pathways workflow

To explore possible heterologous metabolic pathways for the production of a compound of interest, we present the following workflow that combines the reaction rules from RetroRules, the pathway computation capabilities of RetroPath2.0, the visualisation of the results by ScopeViewer and finally the pathway enumeration of RP2paths.

The first step involves the extraction of the totality of rules that do not have any description of stereochemistry (stereochemistry still requires further testing) contained in the RetroRules database. These rules are then used as input to the RetroPath2.0 node, in combination with the source compound (muconic acid) and the sink compounds. In this workflow, the latter corresponds to the metabolome of yeast. Different sinks for different organisms or indeed different strains could be preloaded in the IBISBAHub for the more popular organisms for metabolic engineering.
Metabolomics workflow

Figure 9: Workflow for the identification of new peaks by using the promiscuous nature of the reaction rules for RetroRules2.0. The output is then fed to the metabolomics node to be compared to mass spectrometry output.

Mass spectrometry is a very powerful tool frequently used to study the metabolism of cells and verify that an engineered strain is actually producing the compounds of interest. Indeed, annotated databases of metabolites such as the Human Metabolome Database [14] provides useful services that enable one to identify individual metabolites and compare metabolomes. However, as of today, a significant number of spectral peaks remain unassigned. This high fraction of unassigned peaks might be due to several factors including isotope, adduct formation, ion fragmentation, and multimers. One possible factor that is rarely taken into consideration is the occurrence of metabolites through the promiscuous nature of enzymatic reactions.

To explore the latter, this workflow applies the RetroPath2.0 algorithm in a forward manner to explore the diversity of molecules possible from a given sink, calculates the molecular weight of the results to be compared with the measured mass spectrometry output using OpenMS.
3.3 A Workflow Library in IBISBAHub

The workflows and workflow nodes above need to be catalogued and registered in a searchable IBISBA collection, along with example data, narratives etc. They need to be curated, updated with new versions and made portable and interoperable to run over the infrastructures of IBISBA. Along with the SOPs developed by WP6, the WP7 workflow and workflow node library are registered in the IBISBAHub, our centralised catalogue for sharing the outcomes of IBISBA. A guide for SOPs is registered along with the first SOPs of the project. Preliminary extensions developed to the IBISBAHub to support workflows is reported below.

The IBISBAHub is to be fully reported in Deliverable D7.3 and the metadata used to describe them is to be fully reported in Deliverable 7.2.

3.4 Workflows and Workflow nodes in CWL

IBISBAHub will support computational workflows defined in the Common Workflow Language (CWL), providing interoperability and scaling across execution cloud instances, as well as rich support for structured metadata and file format declarations. It will support CWL descriptions and in due course CWL workflow execution.

An IBISBA Workflow is a particular combination of the IBISBA Workflow Nodes presented in the previous sections, which is then executed with explicit data inputs already uploaded to IBISBAHub and its results optionally deposited into IBISBAHub. To ensure compatible composition and software deployment of multiple flavoured IBISBA Workflow Nodes we use containers using Docker, which capture the complete runtime environment of the software without using virtualized machines. Docker images, typically distributed in the public Docker Hub, allow automatic deployment and upgrades on any local and cloud compute infrastructure. The Common Workflow Language community recommends distributing tools in such container images, as it makes the workflow definition portable, interoperable and reproducible without any manual installation.

Through analysis of our library, we have identified that the IBISBA Workflow Nodes are chiefly:

- Individual command line tools and Python scripts like RP2Paths, or
- KNIME workflows like RetroPath2.0

We have therefore adopted two methods for packaging the IBISBA Workflow Nodes as containers.

Script-based nodes as containers

We have wrapped the RP2Paths Python script as a Docker image, now available in the public Docker Hub as https://hub.docker.com/r/ibisba/rp2paths/, which we have tested to be reusable on a blank machine. This wrapping simplifies use of this IBISBA Workflow Node, which would otherwise require a detailed installation of multiple dependencies. The Node is described as CWL tool, which can reference the Docker image for deployment. We have also investigated packaging of RP2Paths as a BioConda package, which would simplify broader use without requiring container technology, however for more immediate results we have mostly focused on the Docker image with the addition of simple integration tests.
KNIME-based nodes as containers

As a majority of nodes are currently developed as KNIME workflows, we are developing a KNIME-as-CWL wrapper. Conceptually, this creates one Docker image per KNIME workflow, which is deployed and registered to the Docker Hub. Furthermore, the wrapped KNIME workflow is described in CWL to formalize inputs, outputs, and format metadata, resulting in a proper IBISBA Workflow Node.

- The KNIME-based node containers share a common “IBISBA KNIME Node” base image ([https://github.com/ibisba/knime-base](https://github.com/ibisba/knime-base)) that is a specific version of KNIME Analytics Platform for command line execution. Putting versions of KNIME Analytics Platform and additional extensions and integrations into a Docker image ensures that each IBISBA Workflow Node will use a specific and tested version of KNIME runtime, and can be maintained, updated or versioned independently.

- An additional “KNIME Workflow” image ([https://github.com/ibisba/knime-workflow-base](https://github.com/ibisba/knime-workflow-base)) on top of this base image provides scripts and tooling for wrapping a KNIME workflow into a Docker image. This container image can then be invoked using Docker without any external dependencies. This makes a KNIME version of an IBISBA Workflow Node.

Both image types are registered in the Docker Hub. This functionality is based on an existing prototype that has been extended to finalize Docker images to pre-install any KNIME Extensions or Integrations needed by the particular IBISBA Workflow Node.

Our initial approach is to manually create the IBISBA Workflow Node for RetroPath 2.0, as it is the precursor for calling the above mentioned RP2Paths node, which allows us to test a 2-step pipeline combining heterogeneous IBISBA Workflow Nodes early on. We customized the “KNIME Workflow” base image for RetroPath 2.0 to add its KNIME workflow definition. We generate the CWL for the specific inputs and outputs of the RetroPath 2.0 workflow manually.

3.5 SOP development

In order to describe the individual experimental and computational tasks involved in the Design, Build, Test, and Upscale phases of the synthetic biology meta-workflow, and to enable the collection of SOPs, we have developed a web-based application called TasCu ([https://tascu.vtt.fi](https://tascu.vtt.fi)). TasCu operates as a crowd-sourcing platform, in which the EU-IBISBA1.0 partners have the ability to Create, Read, Update, or Delete (CRUD) the individual steps of the meta-workflow (referred to as TasCu steps or DBTL steps).

The TasCu system has enabled the collection of SOPs and the small cohort of SOPs collected are available at [https://goo.gl/e9vEKR](https://goo.gl/e9vEKR) and are being uploaded to the IBISBAHub. A preliminary recommended format for describing the SOPs has also been developed. Once partners have had sufficient opportunities to test this and determine the appropriate level of detail for the SOPs, this format will become evolve into a standard.

The future work includes collection of more SOPs, standardizing their formats, defining metadata and registering the SOPs and metadata in the IBISBAHub.
4 Conclusion and future work

Many improvements are planned on the RetroRules IBISBA Workflow Node. In its current form, the node extracts all the reaction rules that do not describe the stereochemistry of the reaction. The limited abilities are not reflective of the great potential that this database has to offer. Work has started to implement an additional IBISBA Workflow Node that connects to an improved version of RetroRule’s REST API to enable the user to better extract the rules that are most relevant to the user’s needs.

One of the next logical steps in the presented nodes is the evaluation of the metabolic pathways. Figure 4 shows the ScopeViewer output of the muconic acid workflow (Figure 8) and in its current form, it is up to the user to select one of the metabolic pathways (that may be enumerates by RP2paths) to further study. Depending on the number of pathways and the parameterisation of the RetroPath2.0 tool in general, much larger metabolic maps may be generated that complicate the identification of one or more “best” pathways. To this end a new node called rpFBA is currently being developed that takes a whole cell model, adds the individual metabolic paths from RP2paths, and performs flux balance analysis with objective function the biomass and the product of interest. The goal is to be able to reduce the number of metabolic pathways that are presented to the user for manual curation.

IBISBA Workflow Nodes are also currently being developed that attempt to link between metabolic pathway enumeration and lab implementation; this is done by the automatic generation of the genetic constructs for the expression of the heterologous metabolic pathways in the organism of interest. Currently, SBOL (the Synthetic Biology Open Language) together with autoprotocol.org (an open standard for life science experimental design and automation) are being considered for the automatic generation of appropriate genetic constructs and experimental implementation respectively.

We are also developing a strategy for how IBISBA Workflow Nodes shall be maintained and further developed, while at the same time supporting production-use by IBISBA users.

We are aiming for automatic or semi-automatic generation of the corresponding CWL definition for invoking KNIME-based containers. We are planning to improve support for rich inputs (currently implemented as so-called QuickForms in KNIME Analytics Platform) to more readily support both desktop and command line invocation.

We have a development branch of IBISBAHub (separate from the in-production IBISBAHub). The development branch currently:

- Allows the registration of workflows defined on GitHub or on myExperiment
- Supports the visualisation of workflows using the CWL Viewer (https://view.commonwl.org/workflows)
- Allows the registration of workflow nodes that are defined on GitHub
- Allows the specification of workflow SOPs that relate to one or more workflows and specifies how to run them for a specific purpose in the context of IBISBA

The linkage to GitHub and myExperiment will be incrementally improved, for example to extract metadata, descriptions and versioning. The development branch capabilities are to be merged with the main production IBISBAHub and become live at the end of the first year of the project.
5 Partners involved in the work

Melchior du Lac, Thomas Duigou, Jean-Loup Faulon, Micalis Institute, INRA, FR

Carole Goble, Alan Williams, Stian Soiland-Reyes, UNIMAN, UK

Gopa Peddinti, Peter Blomberg, Jussi Jäntti, Merja Penttilä, VTT, FI

Martin Horn, Stefan Helfrich, KNIME, DE

Vitor Martins dos Santos, María Suárez, Rita Volkers WUR

6 References


