B-Fabric: On-the-Fly Integration of Life Sciences Applications

Can Türker, Fuat Akal, Dieter Joho, Christian Panse, Simon
Barkow-Oesterreicher, and Ralph Schlapbach

Functional Genomics Center Zurich (FGCZ), UZH / ETH Zurich
Winterthurerstrasse 190, CH–8057 Zurich, Switzerland
`<tuerker|akal|cp|sb|schlapbach>@fgcz.ethz.ch`

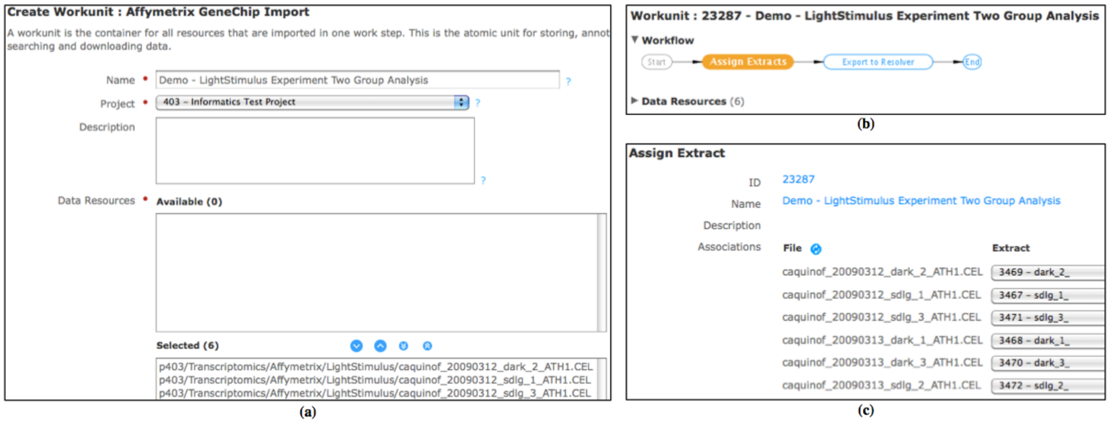
1 Demonstration

In this demonstration, we sketch some central features of the B-Fabric life sciences data management system [1]. As an example scenario, we use a scientist who is working on a plant named Arabidopsis Thaliana with the goal to figure out the effect of certain genes and the effect on light on it. For this purpose, he registers his samples and extracts with B-Fabric, loads his data into B-Fabric and defines his experiment. Afterwards, he runs his experiment and stores the results in B-Fabric. To complement this scenario, retrieval and administrative issues are being demonstrated as well. Due to space limitations and to avoid clutter, only the screen shots of the system related to the core of this demo are presented in the paper. The remaining parts are verbally explained.

Register Samples/Extracts. Users register their samples and extracts through intuitively designed forms. Data entering is facilitated by providing as many drop-down menus as possible to select annotations from the system vocabularies and by dynamically drawing forms according to selected annotation values. In addition, users typically register several samples and extracts where only a few attributes differ. In order to further ease the registration of them, cloning as well as batch registration of samples and extracts are supported.

Annotation Management. B-Fabric provides extensible vocabularies for the different annotations. If a user does not find a needed annotation in the corresponding drop-down list, the user can create a new one. All annotations created by users must be reviewed by an expert (in our case by an FGCZ employee). The expert checks the annotation and releases it if it is correct.

Annotation reviewing can be a tedious task due to similarly written versions of the same annotation. In such cases, B-Fabric automatically detects similar annotations and recommends merging them. B-Fabric allows similar annotations to be merged easily to maintain the annotation consistency system wide. If the expert decides to merge two annotations, B-Fabric provides a form where he can easily select the attributes of the resulting merged annotation. When the two annotations merged, B-Fabric automatically associates the samples which were previously associated with the misspelled annotation.


Create Workunit : Affymetrix GeneChip Import

A workunit is the container for all resources that are imported in one work step. This is the atomic unit for storing, annot
searching and downloading data.

Name * Demo - LightStimulus Experiment Two Group Analysis
Project * 403 – Informatics Test Project
Description
Data Resources * **Available (0)**

Selected (6)

p403/Transcriptomics/Affymetrix/LightStimulus/caquinof_20090312_dark_2_ATH1.CEL
p403/Transcriptomics/Affymetrix/LightStimulus/caquinof_20090312_sdlg_1_ATH1.CEL
p403/Transcriptomics/Affymetrix/LightStimulus/caquinof_20090312_sdlg_3_ATH1.CEL

(a)

Workunit : 23287 - Demo - LightStimulus Experiment Two Group Analysis

(b)

Assign Extract

ID 23287
Name Demo - LightStimulus Experiment Two Group Analysis
Description
Associations

File	Extract
caquinof_20090312_dark_2_ATH1.CEL	3469 – dark_2_
caquinof_20090312_sdlg_1_ATH1.CEL	3467 – sdlg_1_
caquinof_20090312_sdlg_3_ATH1.CEL	3471 – sdlg_3_
caquinof_20090313_dark_1_ATH1.CEL	3468 – dark_1_
caquinof_20090313_dark_3_ATH1.CEL	3470 – dark_3_
caquinof_20090313_sdlg_2_ATH1.CEL	3472 – sdlg_2_

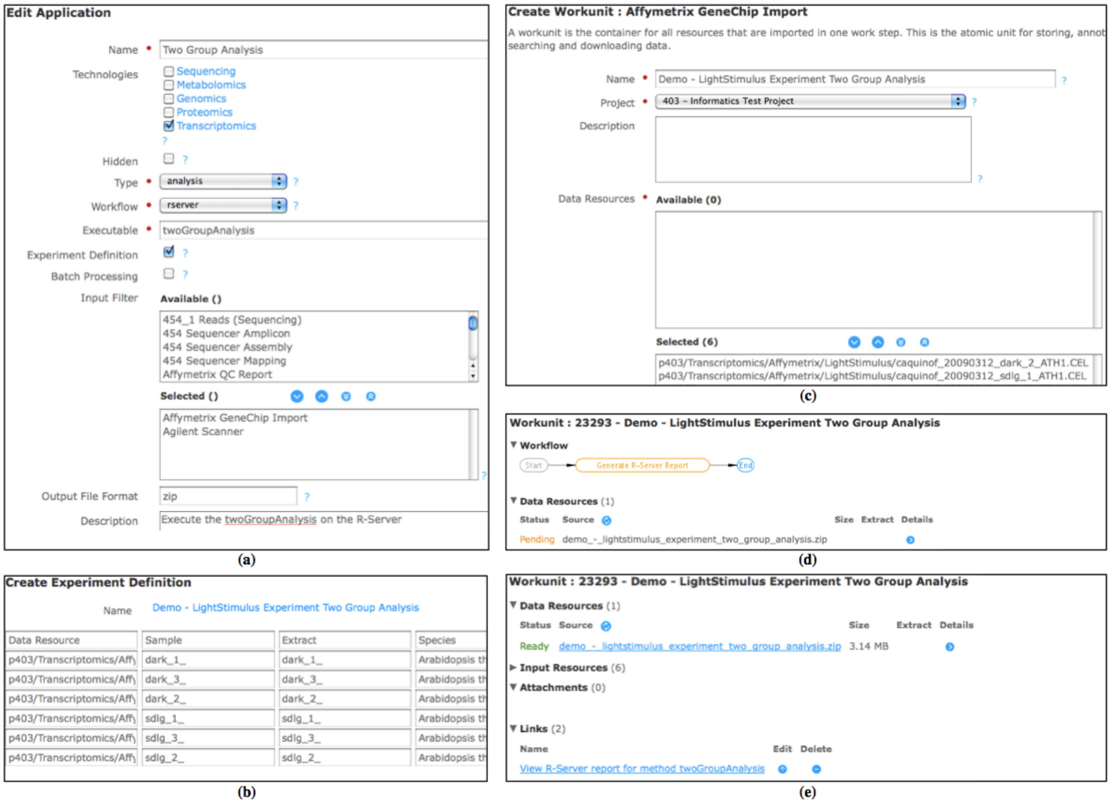
(c)

Fig. 1. a) Create Workunit b) Assign Extracts Workflow c) Assign Extracts

Task Orientation. B-Fabric is a task-oriented system that reminds its users about open tasks, awaiting to be performed next. For instance, as soon as a new annotation is added to the vocabulary, a new task to release this annotation appears in the task list of the corresponding expert.

Data Import. B-Fabric supports two ways to import data: 1) physically copying and 2) linking data files. To import data from a data source, a proper data provider must be configured. The B-Fabric deployment at FGCZ allows importing data files from local file systems as well as several instruments available at FGCZ. New data providers can be added to the system easily. With the configuration of a data provider the selection of the data files in corresponding data stores can be restricted to the ones that are potentially relevant for the user. This is a crucial feature since the number of the data files can be huge. An import results in a workunit. A workunit thus represents a unit of logically related data files. Figure 1(a) shows the screen where a workunit is created by fetching files from the Affymetrix GeneChip instrument, which is an instrument already known to B-Fabric. B-Fabric implements the data import via workflows. With the initiation of a data import, the corresponding workflow becomes visible to the user (Figure 1(b)). The next step to be taken by the user is highlighted in the graphical representation of the workflow. In data import workflow, for instance, the user must assign extracts to the imported files. The workflow-driven approach of B-Fabric is very useful in practice to reduce human mistakes and avoid skipping steps. Assigning extracts to data resources also comes with some intelligence in B-Fabric. When the scientist goes to the assign extracts screen, he gets already the best matches between data resources and extract names (Figure 1(c)). Typically he just needs to press the save button and continue.

Application Integration. Integration of external functionality into B-Fabric is done via *application registration*. First, a connector is written for a certain type of application, e.g., for running *R* scripts on an *Rserve* system. Then, a small



Edit Application

Name • Two Group Analysis

Technologies
☐ Sequencing
☐ Metabolomics
☐ Genomics
☐ Proteomics
☑ Transcriptomics

Hidden ☐

Type • analysis

Workflow • rserver

Executable • twoGroupAnalysis

Experiment Definition ☑

Batch Processing ☐

Input Filter

Available ()
454_1 Reads (Sequencing)
454 Sequencer Amplicon
454 Sequencer Assembly
454 Sequencer Mapping
Affymetrix QC Report

Selected ()
Affymetrix GeneChip Import
Agilent Scanner

Output File Format zip

Description Execute the twoGroupAnalysis on the R-Server

(a)

Create Experiment Definition

Name Demo - LightStimulus Experiment Two Group Analysis

Data Resource	Sample	Extract	Species
p403/Transcriptomics/Affy	dark_1_	dark_1_	Arabidopsis th
p403/Transcriptomics/Affy	dark_3_	dark_3_	Arabidopsis th
p403/Transcriptomics/Affy	dark_2_	dark_2_	Arabidopsis th
p403/Transcriptomics/Affy	sdlg_1_	sdlg_1_	Arabidopsis th
p403/Transcriptomics/Affy	sdlg_3_	sdlg_3_	Arabidopsis th
p403/Transcriptomics/Affy	sdlg_2_	sdlg_2_	Arabidopsis th

(b)

Create Workunit : Affymetrix GeneChip Import

A workunit is the container for all resources that are imported in one work step. This is the atomic unit for storing, annot searching and downloading data.

Name • Demo - LightStimulus Experiment Two Group Analysis

Project • 403 – Informatics Test Project

Description

Data Resources • **Available (0)**

Selected (6)
p403/Transcriptomics/Affymetrix/LightStimulus/caquinof_20090312_dark_2_ATH1.CEL
p403/Transcriptomics/Affymetrix/LightStimulus/caquinof_20090312_sdlg_1_ATH1.CEL

(c)

Workunit : 23293 - Demo - LightStimulus Experiment Two Group Analysis

▼ **Workflow**

▼ **Data Resources** (1)

Status	Source	Size	Extract	Details
Pending	demo_-_lightstimulus_experiment_two_group_analysis.zip			

(d)

Workunit : 23293 - Demo - LightStimulus Experiment Two Group Analysis

▼ **Data Resources** (1)

Status	Source	Size	Extract	Details
Ready	demo - lightstimulus experiment two group analysis.zip	3.14 MB		

▶ **Input Resources** (6)

▼ **Attachments** (0)

▼ **Links** (2)

Name	Edit	Delete
View R-Server report for method twoGroupAnalysis		

(e)

Fig. 2. a) Application Registration b) Create Experiment Definition c) Run Experiment d) Run Experiment - Pending State e) Run Experiment - Ready

interface is defined to describe how the application gets its input (Figure 2(a)). Finally, the scientist writes the application (shown as Executable in the figure) in any language. This on-the-fly coupling of external applications is a crucial feature of B-Fabric, which allows fast evolution of the system. Once an application is registered, an experiment can be created to run this application. As an example, Figure 2(b) shows the definition of the experiment that will be conducted on the Arabidopsis Thaliana plant as mentioned earlier. Defining an experiment consists of selecting data resources, samples, extracts, and an arbitrary number of attributes (e.g. species and treatment in the example) that will be used as input for the application. Figure 2(c) shows how easily a previously registered application (Two Group Analysis) can be invoked to conduct the desired experiment. This step requires a name for the resulting workunit which contains the result files of the application along with specific parameters regarding the experiment, e.g. reference group. Once the experiment is started, a corresponding workflow is initiated. The graphic presentation of the workflow is also used to show what is happening underneath in the system. The example workflow

(Generate R Server Report) is quite simple and consists of a single step (see Figure 2(d)). Note that B-Fabric supports arbitrary complex workflows based on its underlying workflow engine. When the experiment is done, the scientist can view the experiment results by clicking the related link on the screen (see Figure 2(e)). The results of the experiment is also presented to the user as a zip file so that they can easily be transferred to another medium.

Full-text Search. B-Fabric provides full-text search capabilities. A search may vary from certain attributes of certain objects to the content of readable attachments and data resources. The system provides quick search boxes on the main screen as well as more refined advanced search form. Searches done by the user are kept in the search history during his session and can be executed easily by selecting a search query from the search history. A query can also be saved for future reuse. A later invocation of such a saved query will of course include all objects satisfying the query at run-time. Another important feature of B-Fabric is that search results can be exported into files.

Miscellaneous Functions. In addition to all major aspects presented above, B-Fabric provides some additional functionality. Especially, B-Fabric supports a view on the main data objects in a networked fashion. Users can simply browse bidirectionally through all objects linked together. In addition, all data manipulation operations (create/update/delete) are logged in the system such that the user can remember what he did in the past and the system can be monitored. Last but not least, B-Fabric provides a bunch of administrative functions to manage objects, workflows, errors, and to maintain the system.

Final Remark. B-Fabric is running in daily business at FGCZ since the beginning of 2007 [2]. Here are some figures about the FGCZ deployment as of July 2010:

Users	1766	Samples	4127
Projects	850	Extracts	5033
Institutes	285	Data Resources	49436
Organizations	74	Workunits	29616

Acknowledgement: The further development of B-Fabric is partially supported by SWITCH (Swiss Academic and Research Network) within the AAA/Switch project "Generalizing B-Fabric towards an Infrastructure for Collaborative Research in Switzerland" (UZH.5, June 2009-May 2011).

References

1. Türker, C., Stolte, E., Joho, D., Schlapbach, R.: B-Fabric: A Data and Application Integration Framework for Life Sciences Research. In: Data Integration in the Life Sciences, DILS 2007, LNCS 4544, pp. 37–47. Springer-Verlag, 2007.
2. B-Fabric. `http://www.bfabric.org/`