

Architektura dużych projektów bioinformatycznych

Bartek Wilczyński

bartek@mimuw.edu.pl

<http://www.mimuw.edu.pl/~bartek>

**Wykład 6. - O przepływach pracy –
Kepler, Taverna i Galaxy**
25. IV. 2018

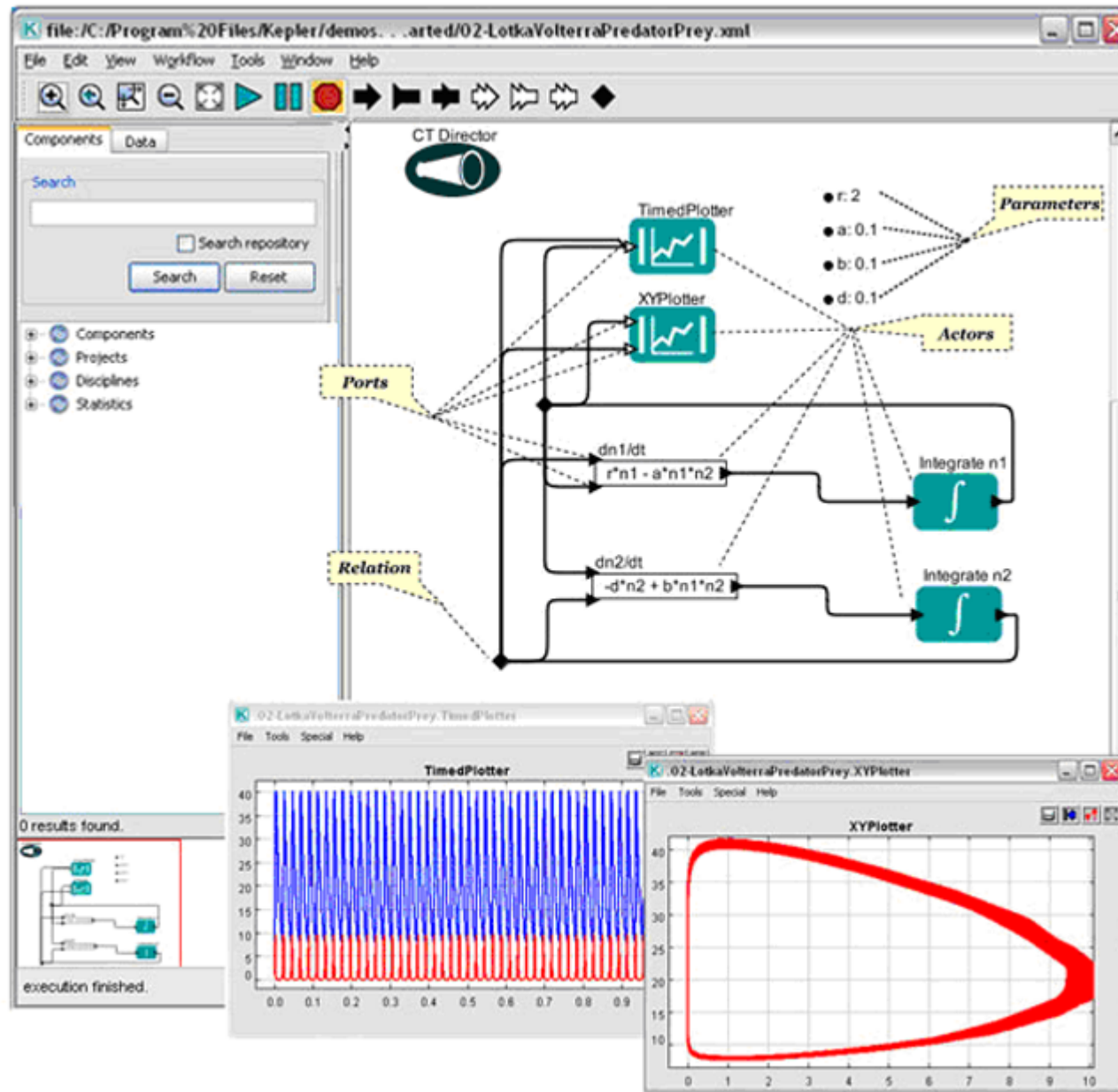
Plan na dziś

- Usługi sieciowe i projekty typu grid
- O pomysłach na przepływy pracy – Kepler (2002), Taverna workbench (2004), i wiele innych
- Pomysły na wymianę workflow'ów - myExperiment, i usług – biocatalogue
- Porównanie podejść – rozproszonego z użyciem aplikacji klienta (Taverna itp) i klient-serwer przez przeglądarkę (Galaxy)

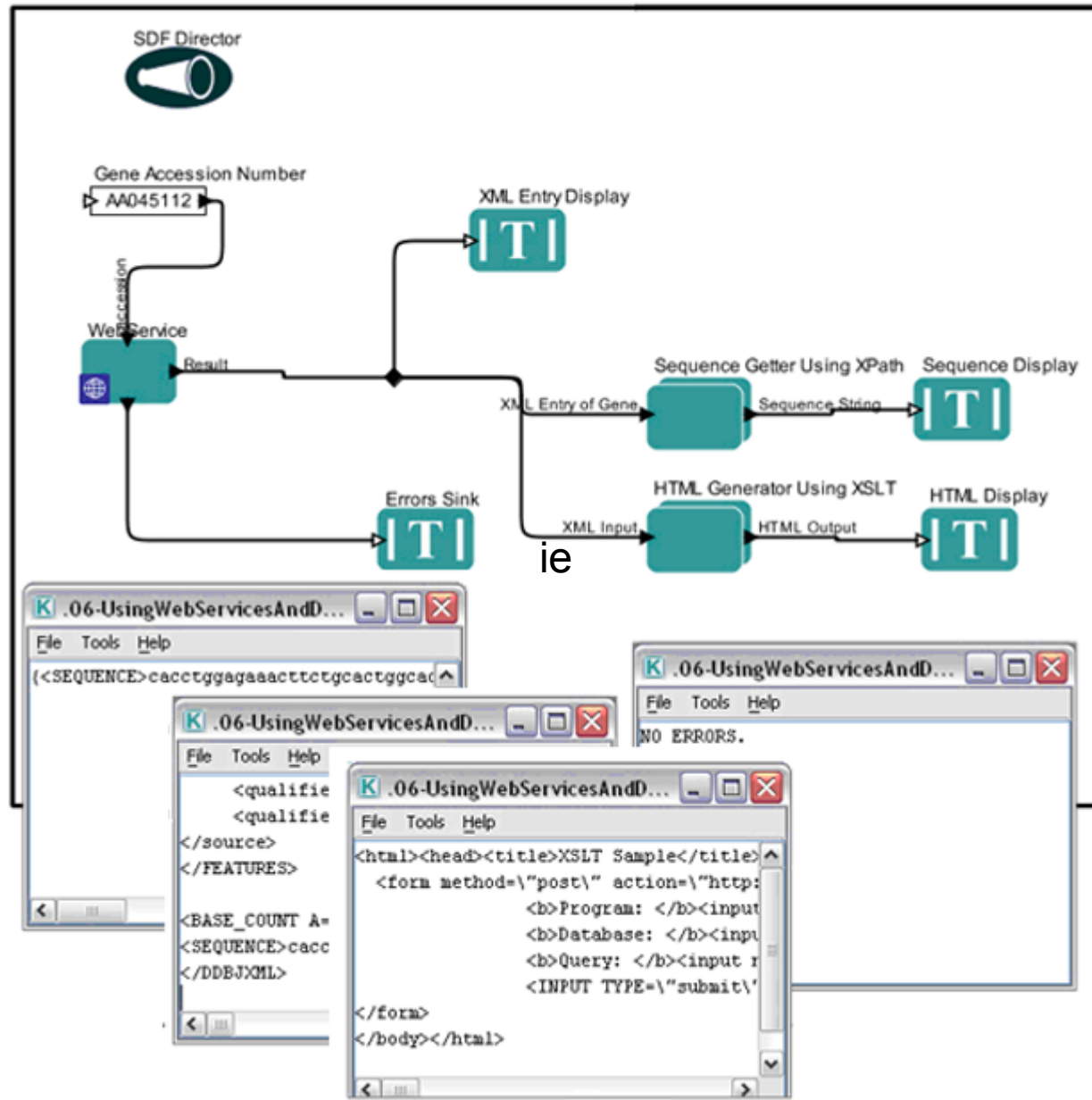
Projekt Kepler

- Pomyślany jako interfejs do pakietu Ptolemy (1999), służącego do złożonego modelowania matematycznego procesów fizycznych
- Napisany w Javie
- Pierwsze wersje w roku 2002
- Obecnie w wersji stabilnej
- Stale organizowane szkolenia dla naukowców i nauczycieli
- Open Source(BSD), wspierany przez granty NSF

Model Lotki-Volterra w Keplerze



Pobieranie sekwencji DNA z usługi sieciowej



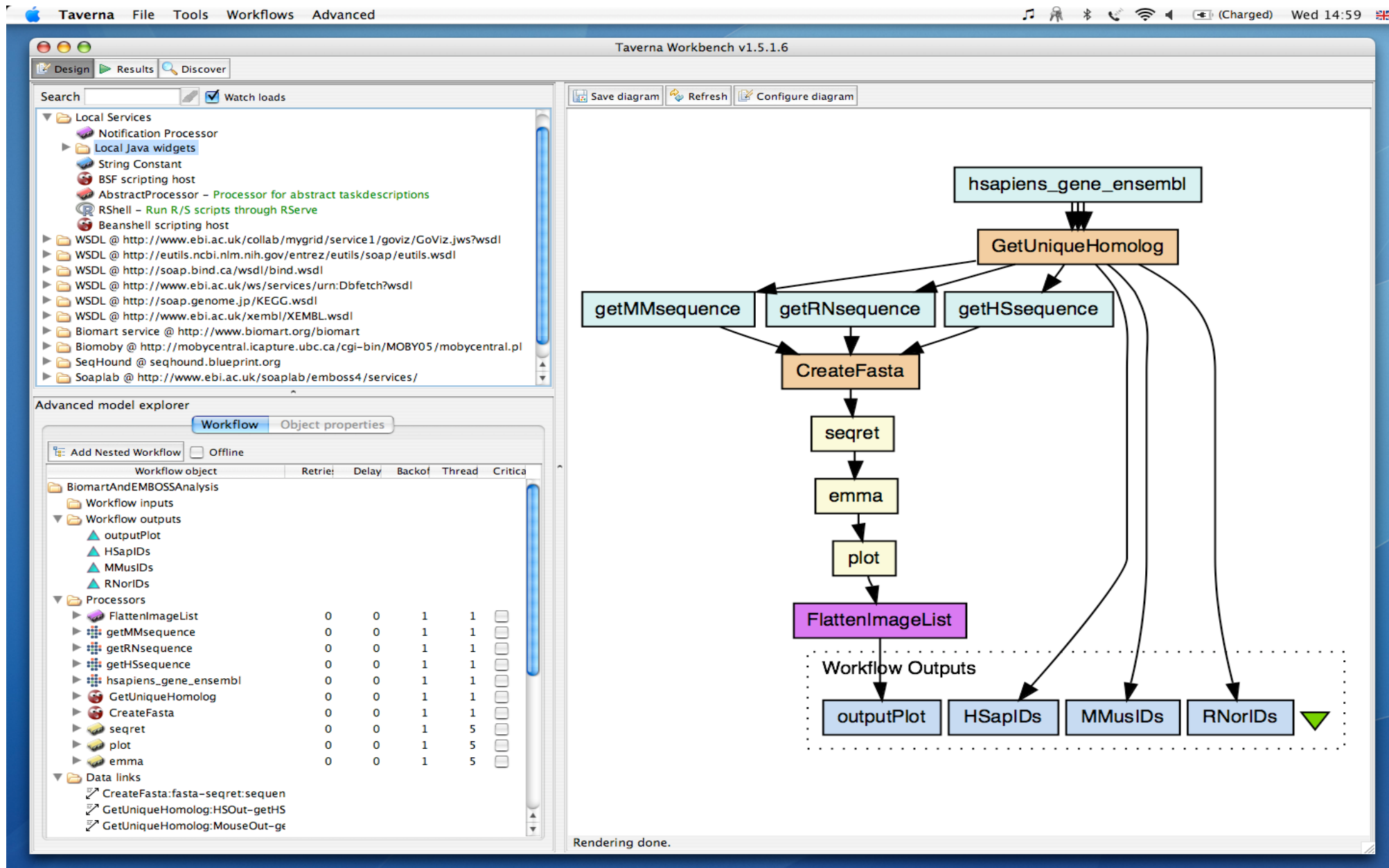
Cechy Kepler'a

- Środowisko uruchomieniowe Ptolemy (równoległe działające systemy)
- Interfejsy do czujników eksperymentalnych
- Wersje na GRID
- BioKepler dla bioinformatyków
- Raczej dla eksperymentatorów lub nauczania
- Silne wsparcie w środowisku fizyków doświadczalnych

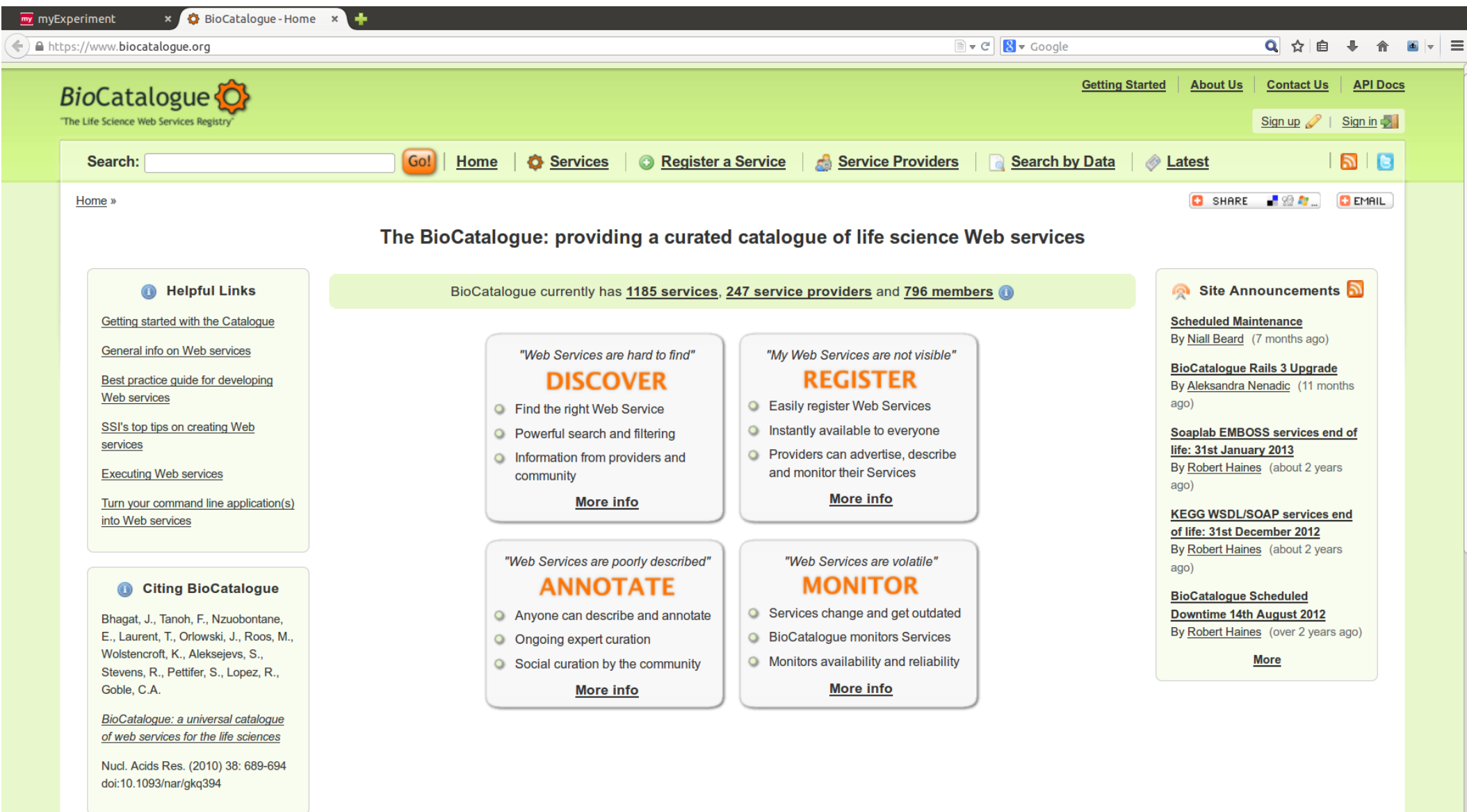
Taverna Workbench

- Rozpoczęta jako oddzielny projekt w roku 2002
- Wsparcie grantami badawczymi, obecnie konsorcjum
- Licencja LGPL
- Zasadniczo oparta na pomysle reproducible research
- Oryginalnie pomyślana jako interfejs do webserwisów (bez obliczeń lokalnych)
- Wersje dla astronomów, bioinformatyków, itp.
- Obecnie również do obliczeń lokalnych, interfejs do R, Galaxy itp

Pobieranie sekwencji z Ensembl



BioCatalogue



MyExperiment



myExperiment makes it easy to **find**, **use** and **share scientific workflows** and other **Research Objects**, and to build **communities**.

AllSearch

First time visitor? Try these videos:

 [Project Introduction](#)

 [Bioinformatics Case Study](#)

Use myExperiment to...

 [Find Workflows](#)

 [Share Your Workflows and Files](#)

 [Create and Find Packs of Items](#)

 [Find People and Make Friends](#)

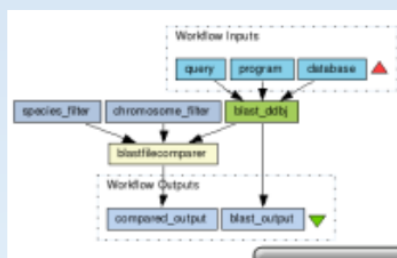
 [Create and Join Groups](#)

 [Build your Profile and Reputation](#)

 [Tag and Rate things](#)

 [Write Reviews and Comments](#)

Explore



Find Workflows

[About myExperiment](#)

[Join the Mailing List](#)

[myExperiment Publications](#)

[For Developers](#)

[Give us Feedback](#)

[The BioCatalogue Project](#)

Register

or Login:

Username or Email:

Password:

Remember me: ☐

Or use OpenID:

(eg: name.myopenid.com)

Login

[Forgot Password?](#)

Galaxy

- Środowisko do uruchamiania narzędzi bioinformatycznych
- Architektura klient-serwer, ze wsparciem dla historii danych i analiz
- Pozwala na zapisywanie historii i dzielenie się nimi w ramach serwera
- Edytor “przepływów”, pozwalający na uruchamianie tych samych analiz na wielu zbiorach danych
- Zachowuje elementarne zabezpieczenia danych
- Pozwala na publikację wyników analiz

Publiczny serwer galaxy

Galaxy

https://usegalaxy.org

my experiment

Using 0%

New high performance job execution options are available! See [the wiki](#) for more information.

Tools

search tools

Get Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity

NGS TOOLBOX BETA
Phenotype Association
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: RNA-seq
NGS: Picard (beta)
NGS: Variant Analysis
NGS: VCF Manipulation
snpEff
BEDTools
EMBOSS

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Want help?
Get answers.

 **Biostars**
GALAXY EXPLAINED

Tweets

Follow

Galaxy Project @galaxyproject 14 Nov
A case study for cloud based high throughput analysis of NGS data using Globus Genomics system, Bhuvaneshwar et al. [bit.ly/112U5OV](#)
Expand

CRS4 Orione @CRS4Galaxy 13 Nov
Seven Mycoplasma hyosynoviae strains assembled using Orione @CRS4Galaxy [goo.gl/XMMci1](#)
Retweeted by Galaxy Project
Expand

Galaxy Project @galaxyproject 13 Nov
RNA-Seq Analysis of Differential Gene Expression in Electroporated Chick Embryonic Spinal Cord, Viecei & Yan, JoVE [bit.ly/112kZ9U](#)
Expand

Tweet to @galaxyproject

History

search datasets

Unnamed history
0 bytes

This history has been deleted

This history is empty. You can [load your own data](#) or [get data from an external source](#)

PENN STATE

JOHNS HOPKINS UNIVERSITY

TACC

iPlant Collaborative™

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

This instance of Galaxy is utilizing infrastructure generously provided by the [iPlant Collaborative](#) at the [Texas Advanced Computing Center](#), with support from the [National Science Foundation](#).

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](#).

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own [local Galaxy instance](#) or run [Galaxy on the cloud](#).

Workflow editor

Galaxy

http://main.g2.bx.psu.edu/workflow/editor?id=6ad322b8c4f7fdec

Analyze Data Workflow Data Libraries Lab Admin Help User

Workflow Canvas | Clone of 'metagenomic analysis' shared by 'anton@bx.psu.edu'

Tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Evolution
- Metagenomic analyses
- EMBOSS
- NGS Toolbox Beta
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Peak Calling
- Workflow control
- Inputs

Workflow Canvas

Input dataset

output

Select high quality segments

Reads

Quality scores

output1 (fasta)

FASTA-to-Tabular

Convert these sequences

output (tabular)

Add column

to Query

out_file1

Tabular-to-FASTA

Tab-delimited file

output (fasta)

Megablast

Compare these sequences

output1 (tabular)

Concatenate queries

Concatenate Query

Query 1 > Select

out_file1

Join two Queries

Join

with

out_file1

Filter

Filter

out_file1

Fetch taxonomic representation

Show taxonomic representation for

out_file1 (taxonomy)

Find lowest diagnostic rank

for taxonomy dataset

out_file1 (taxonomy)

Draw phylogeny

Draw phylogram for

out_file1 (pdf)

Summarize taxonomy

Summarize taxonomic representation for

out_file1 (tabular)

Compute sequence length

Compute length for these sequences

output (tabular)

Details

Tool: Filter

Filter

Data input 'Input' (tabular)

With following condition

c5/c15 > 0.5

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Double equal signs, ==, must be used as 'equal to' (e.g., c1 == 'chr22')

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

- Columns are referenced with c and a number. For example, c1 refers to the first column of the dataset.

Repozytorium narzędzi - toolshed

Galaxy Tool Shed - Mozilla Firefox

Galaxy Tool Shed

toolshed.g2.bx.psu.edu

Google

Galaxy Tool Shed

Repositories Help User

1857 valid tools on May 24, 2012

Search

- Search for valid tools
- Search for workflows

Repositories

- Browse by category
- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	17
Computational chemistry	Tools for use in computational chemistry	3
Convert Formats	Tools for converting data formats	23
Data Source	Tools for retrieving data from external data sources	6
Fasta Manipulation	Tools for manipulating fasta data	22
Genomic Interval Operations	Tools for operating on genomic intervals	17
Graphics	Tools producing images	14
Metagenomics	Tools enabling the study of metagenomes	4
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	32
Ontology Manipulation	Tools for manipulating ontologies	5
Proteomics	Tools enabling the study of proteins	0

Możliwość tworzenia własnych “opakowań” dla istniejących narzędzi

- Pisząc dość krótki plik “opakowujący” (Ang. Wrapper) można dołączyć dowolny program działający na serwerze do własnego galaxy
- Łatwo też można zmodyfikować istniejące wrappery, aby dostosować istniejący program do własnych potrzeb
- Jest to dość proste, choć jeśli dotyczy aplikacji graficznych, może być bardziej skomplikowane

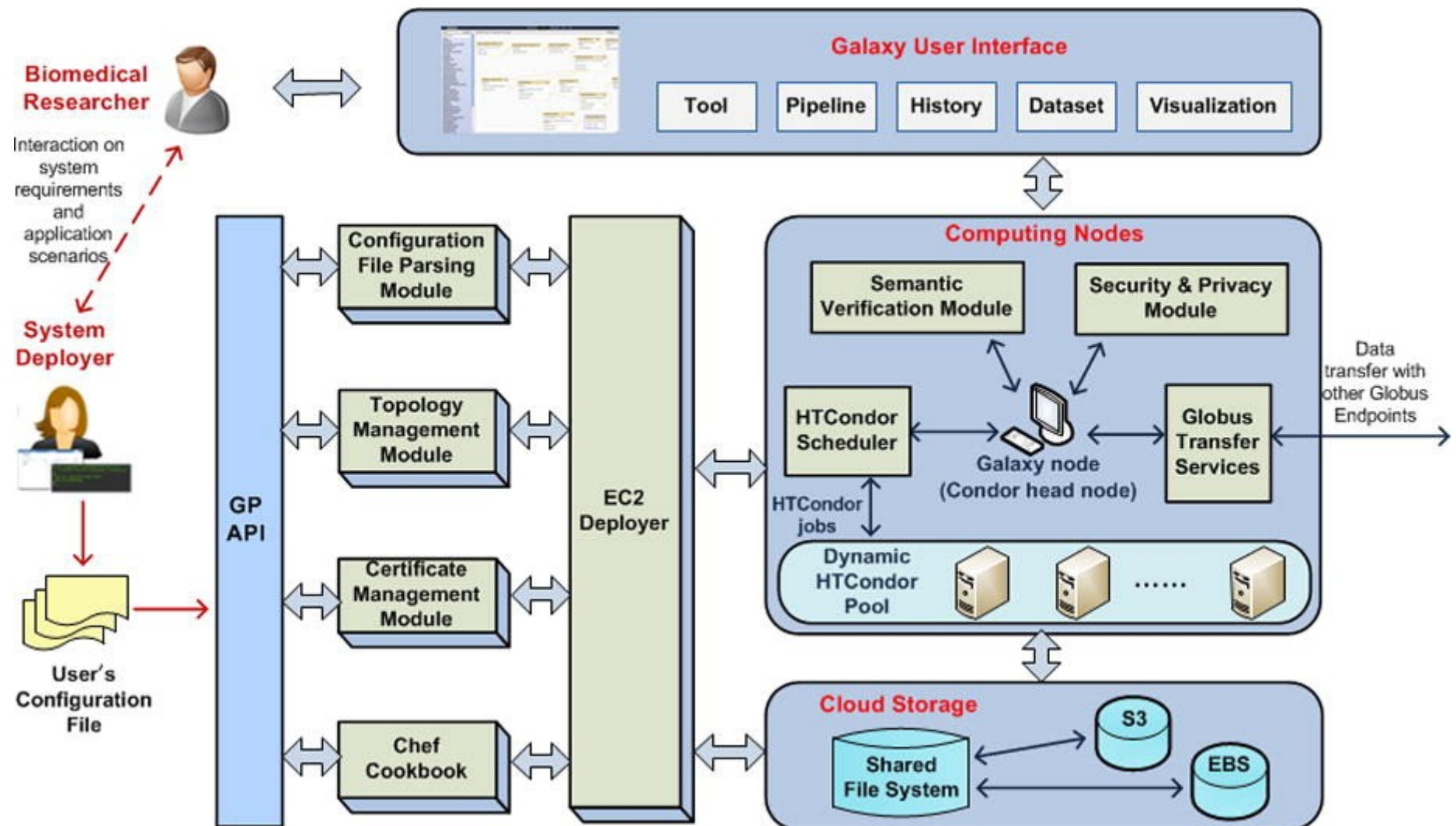
Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses

Bo Liu ^a✉, Ravi K Madduri ^{b, c}, Borja Sotomayor ^b, Kyle Chard ^b, Lukasz Lacinski ^b, Utpal J Dave ^b, Jianqiang Li ^d, Chunchen Liu ^a, Ian T Foster ^{b, c}

[Show more](#)

<https://doi.org/10.1016/j.jbi.2014.01.005>

[Get rights and content](#)



Podsumowanie

- Wiele projektów pomyślanych jako narzędzie do “reproducible research”
- W zasadzie jedyny model finansowania to granty badawcze
- Bardzo powolna adopcja rozwiązań, które wymagają dodatkowego wysiłku, lub są ograniczone
- Duży sukces galaxy w środowisku bioinformatycznym, bo tu może stanowić ułatwienie
- Nadal niełatwe zdobywanie popularności wśród specjalistów