An Interactive Workflow Generator to Support Bioinformatics Analysis through GPU Acceleration

Anuradha Welivita, Indika Perera, Dulani Meedeniya
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka

Outline

- → Introduction
- → Literature Review
- → Methodology
- → Implementation
- → Evaluation and Results
- → Conclusions

Introduction

Background

- ➜ Bioinformatics analyses play a significant role in Bioinformatics research.
- ➜ Carried out by constructing pipelines that executes multiple software tools in a sequential fashion.
- ➜ Workflow systems generated to simplify the construction of pipelines and automate analyses.

Research Problem

- Biological data is ever increasing
  - Hence, difficult to get results within reasonable period of time
- GPU accelerated computing has now become the mainstream for HPC applications
- But currently available solutions only provide distributed system support for parallelized computations
  - E.g. Galaxy, Taverna

Project Objectives

- ➜ An interactive workflow generation system
- ➜ Analyses through cloud based GPU computing resources
- ➜ Supporting specific requirements of bioinformatics software

Literature Review

Existing Techniques

- Scripting
  - A low level, a less abstract method of using basic scripting languages to generate workflows.
- Makefiles
  - A script having a set of rules defining a dependency tree declaratively.
- Scientific workflow management systems
  - Software that provides an infrastructure to set up, execute, and monitor workflows.

Evaluation of Existing Techniques

Technique	Advantages	Limitations	Examples
Scripting	→ Simple to construct → Openness → Ability to execute from command line → Extreme flexibility to manipulate pipelines	→ Not support for shared file systems Bash, Perl, Python → Development overhead → Hard to determine the exact point of failure → Difficult to reproduce analyses → Difficult to integrate new tools and databases	→ Bash → Perl → Python

Evaluation of Existing Techniques

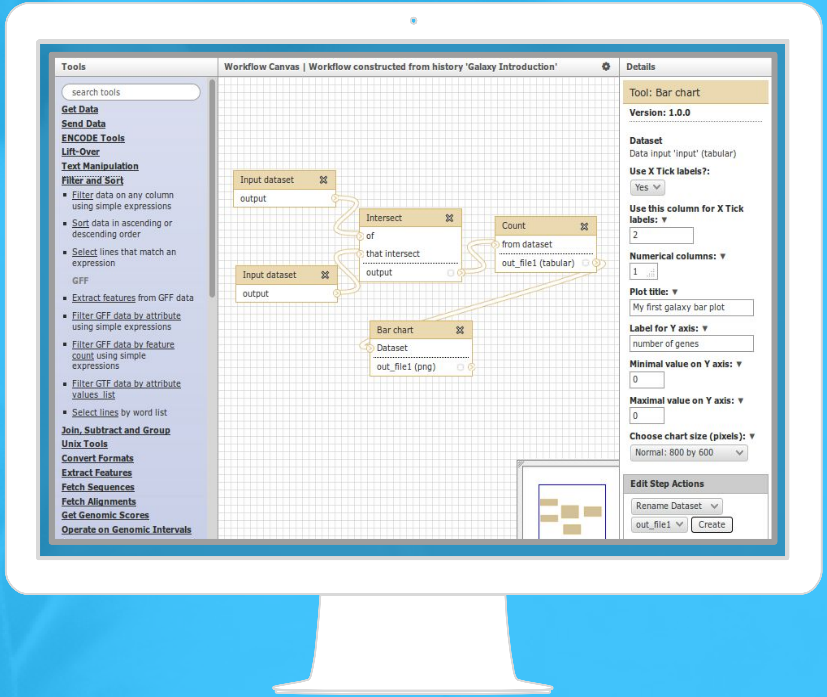
Technique	Advantages	Limitations	Examples
Makefiles	→ Simple to construct → Describe the data flow → Take care of dependency resolution → Commands can be executed in parallel → Cache results from previous runs → State dependencies among files & commands → Lazy processing	→ Not flexible compared to scripting Make, CMake → Single wild-card per rule restriction SCans → Cannot describe a recursive flow Makeflow → Require programming or shell experience Snakemake → Deceptive error messages → No support for multi-threaded/ multi-process jobs	→ Make → CMake → SCans → Makeflow → Snakemake

Evaluation of Existing Techniques

Technique	Advantages	Limitations	Examples
Scientific Workflow Management Systems	→ Interconnects components → Do not require programming experience → Enable reproducible data analysis → Can simply integrate with HPC systems → Allow execution on distributed resources	→ Require more effort → No authority to standardize for interoperability	→ Galaxy → Taverna → Bioconductor → BioPython → Nextflow Workbench

Scientific Workflow Management Systems

→ Galaxy



- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF

- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values_list
- Select lines by word list

Tool: Bar chart

Version: 1.0.0

Dataset
Data input 'input' (tabular)

Use X Tick labels?: Yes

Use this column for X Tick labels: 2

Numerical columns: 1

Plot title: My first galaxy bar plot

Label for Y axis: number of genes

Minimal value on Y axis: 0

Maximal value on Y axis: 0

Choose chart size (pixels): Normal: 800 by 600

Edit Step Actions

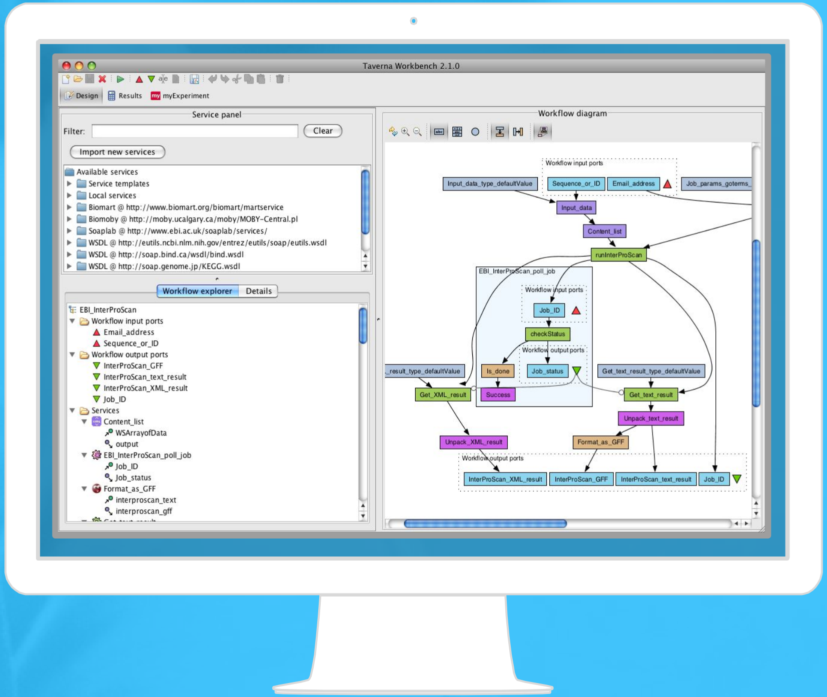
Rename Dataset
out_file1 Create

Galaxy

Pros	Cons
➔ Support reproducibility of results and transparency of workflow execution ➔ Available to users via a simple web interface ➔ Provides distributed computing support for calculations	➔ Does not enable GPU based computations

Scientific Workflow Management Systems

→ Taverna



Taverna Workbench

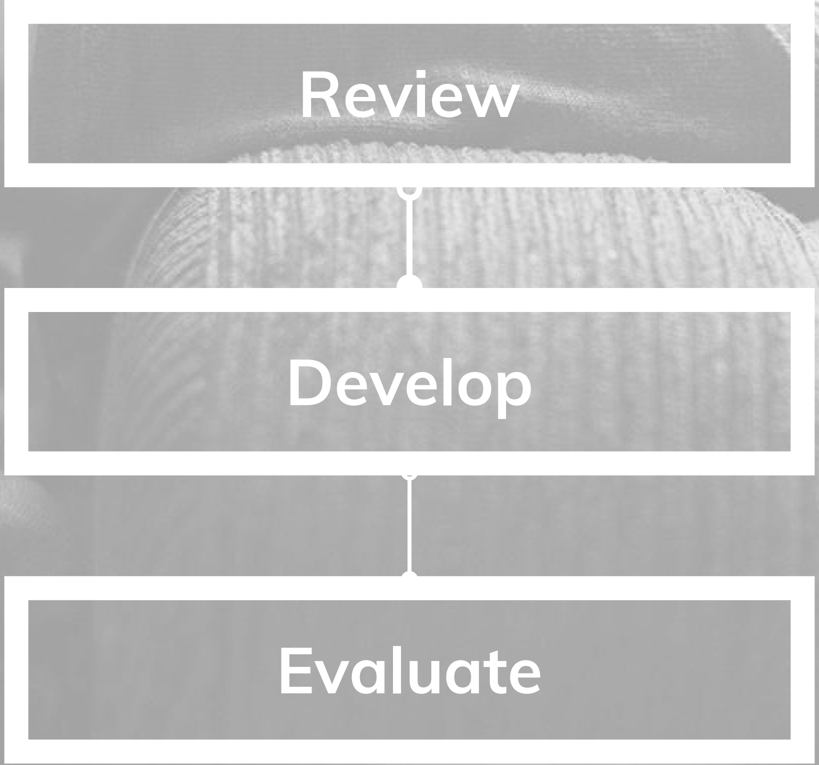
Pros	Cons
→ Enables integration of tools distributed across the internet → Provides a web based platform for sharing workflows → Provides distributed computing support for calculations	→ Being available only as a stand-alone application makes it less accessible by the community → Limited by the platform it runs on → Does not enable GPU based computations

Challenges and Limitations

- ➜ Large-scale data-intensive bioinformatics analyses pose significant challenges on performance and scalability.
- ➜ Currently available solutions only provide distributed system support for parallelized computations.
  - ○ E.g. Galaxy, Taverna
- ➜ But use of GPUs in the cloud can harness the power of GPU computation from the cloud itself and on demand.

Methodology

Process



Review

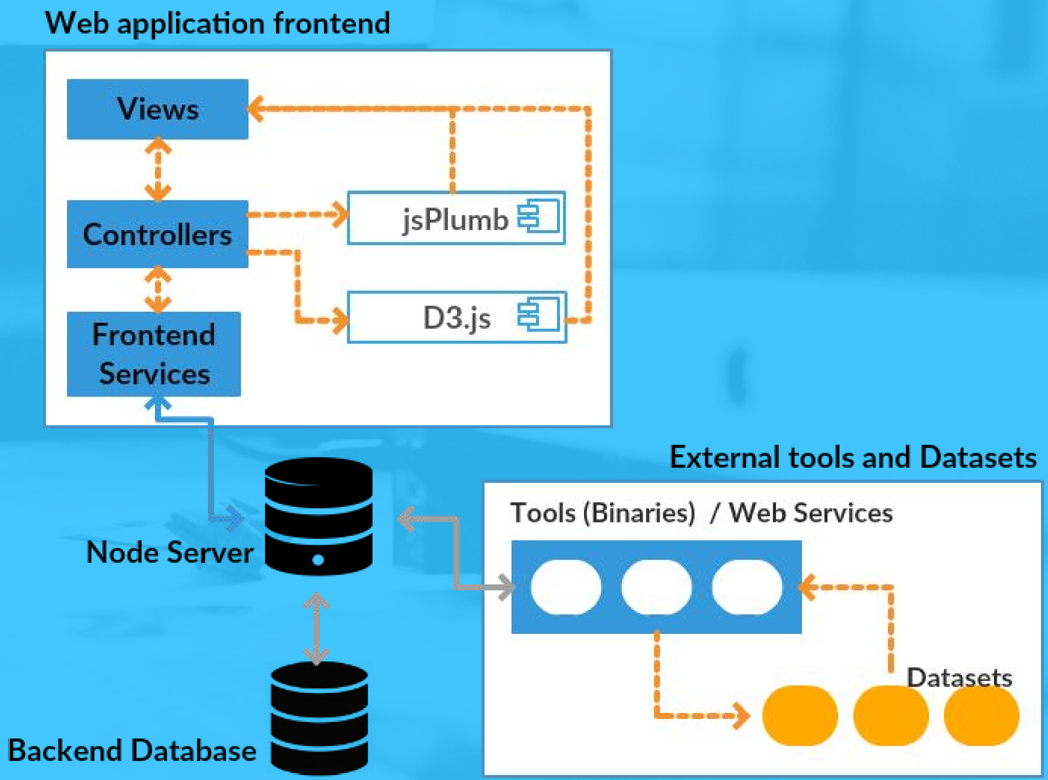
Develop

Evaluate

Web App Development

- ➜ SPA developed using JavaScript and NodeJS
- ➜ Front end development using AngularJS
- ➜ Hosted in an Amazon EC2 P2 instance
    - ○ A GPU accelerated cloud platform
    - ○ With up to 16 NVIDIA Tesla K80 GPUs
    - ○ Scalable and provides parallel computing capabilities

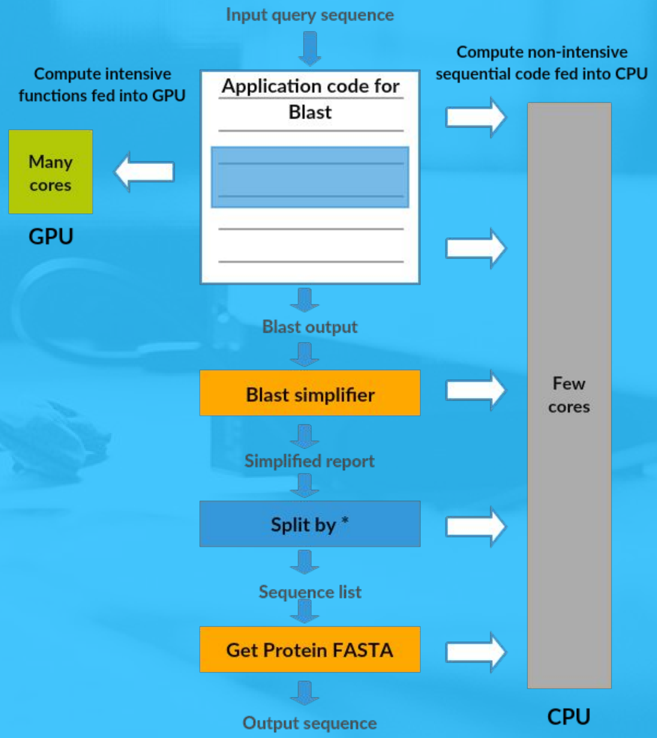
High Level Architecture



Implementation Details

Enhancing Performance

➜ How GPU acceleration works in a simple workflow



Input query sequence

Compute intensive functions fed into GPU

Application code for Blast

Compute non-intensive sequential code fed into CPU

Many cores

GPU

Blast output

Blast simplifier

Simplified report

Split by *

Sequence list

Get Protein FASTA

Output sequence

Few cores

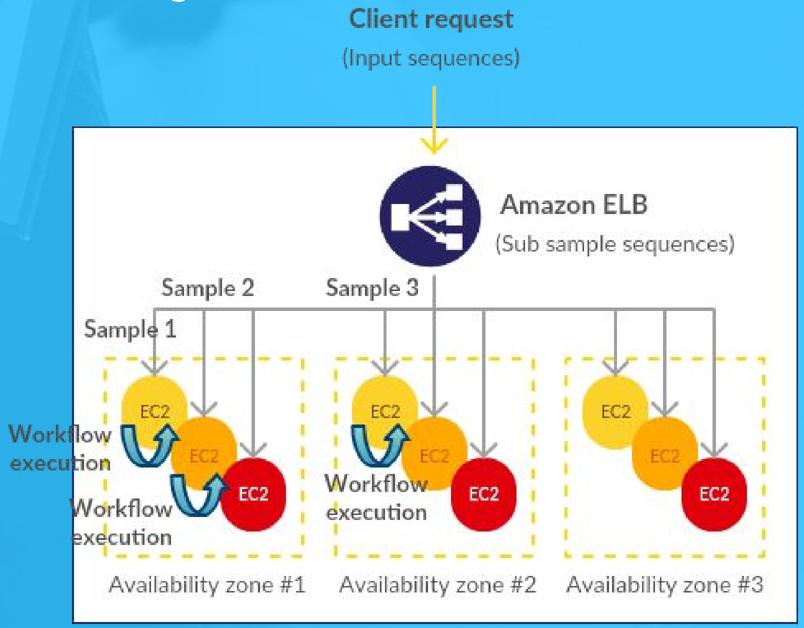
CPU

Enhancing Performance

- ➜ Amazon EC2 P2 virtual machine instances
  - ○ Amazon EC2 - a cloud based instance that enables hosting of HPC applications
  - ○ Amazon EC2 P2 - a type of Amazon EC2 cloud instance that supports computations on NVIDIA k80 GPUs

Enhancing Performance

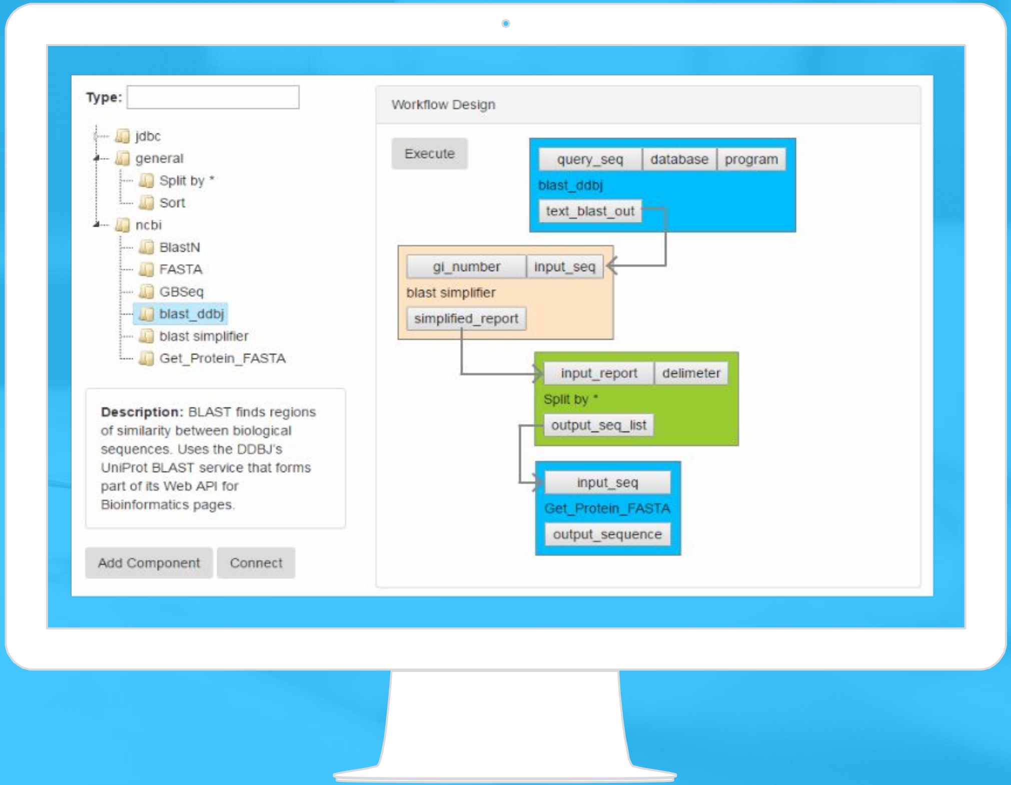
- ➜ On-demand scaling across a cluster of nodes
  - ○ Achieved through Amazon Elastic Load Balancing (ELB)



Implementation of Specific Requirements

- Interactive and Graphical Workflow Creation
- Module Extensibility
- Reporting
- Reproducibility
- User Management

1) Interactive & Graphical Workflow Creation

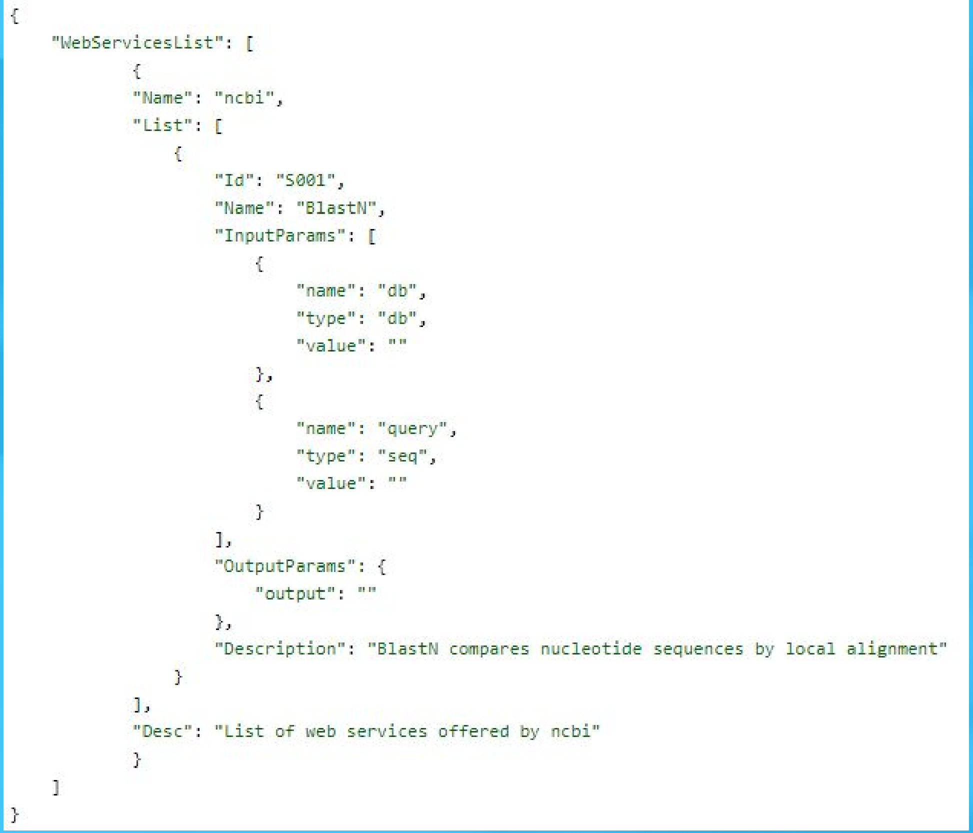


Description: BLAST finds regions of similarity between biological sequences. Uses the DDBJ's UniProt BLAST service that forms part of its Web API for Bioinformatics pages.

2) Module Extensibility

- Capability to add/remove data processing or analysing components
- A plugin architecture for service addition
- Add/remove tools & services via updating a JSON configuration file

2) Module Extensibility



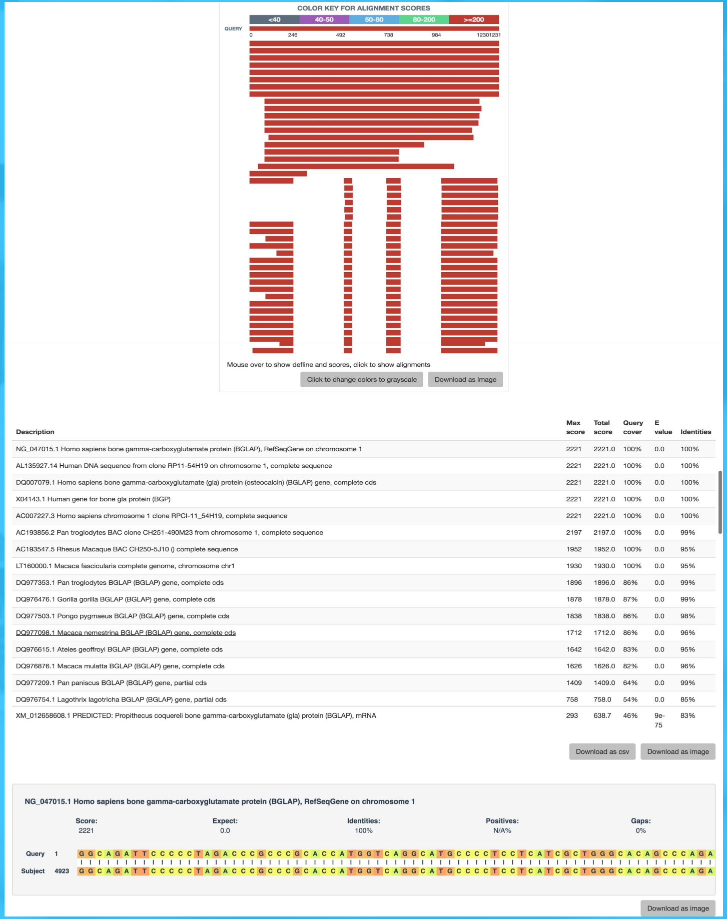
```
{
    "WebServicesList": [
            {
            "Name": "ncbi",
            "List": [
                {
                    "Id": "S001",
                    "Name": "BlastN",
                    "InputParams": [
                        {
                            "name": "db",
                            "type": "db",
                            "value": ""
                        },
                        {
                            "name": "query",
                            "type": "seq",
                            "value": ""
                        }
                    ],
                    "OutputParams": {
                        "output": ""
                    },
                    "Description": "BlastN compares nucleotide sequences by local alignment"
                }
            ],
            "Desc": "List of web services offered by ncbi"
            }
    ]
}
```

3) Reporting

- Helps maintaining details of the executed pipelines and summaries of analysis
- Automatic HTML report generation using PhantomJS
- D3.js to generate analysis specific visualizations

3) Reporting

- ➜ Analysis specific results visualization



Mouse over to show defline and scores, click to show alignments

Description	Max score	Total score	Query cover	E value	Identities
NG_047015.1 Homo sapiens bone gamma-carboxyglutamate protein (BGLAP), RefSeqGene on chromosome 1	2221	2221.0	100%	0.0	100%
AL135927.14 Human DNA sequence from clone RP11-54H19 on chromosome 1, complete sequence	2221	2221.0	100%	0.0	100%
DQ007079.1 Homo sapiens bone gamma-carboxyglutamate (gla) protein (osteocalcin) (BGLAP) gene, complete cds	2221	2221.0	100%	0.0	100%
X04143.1 Human gene for bone gla protein (BGP)	2221	2221.0	100%	0.0	100%
AC007227.3 Homo sapiens chromosome 1 clone RPCI-11_54H19, complete sequence	2221	2221.0	100%	0.0	100%
AC193856.2 Pan troglodytes BAC clone CH251-490M23 from chromosome 1, complete sequence	2197	2197.0	100%	0.0	99%
AC193547.5 Rhesus Macaque BAC CH250-5J10 () complete sequence	1952	1952.0	100%	0.0	95%
LT160000.1 Macaca fascicularis complete genome, chromosome chr1	1930	1930.0	100%	0.0	95%
DQ977353.1 Pan troglodytes BGLAP (BGLAP) gene, complete cds	1896	1896.0	86%	0.0	99%
DQ976476.1 Gorilla gorilla BGLAP (BGLAP) gene, complete cds	1878	1878.0	87%	0.0	99%
DQ977503.1 Pongo pygmaeus BGLAP (BGLAP) gene, complete cds	1838	1838.0	86%	0.0	98%
DQ977098.1 Macaca nemestrina BGLAP (BGLAP) gene, complete cds	1712	1712.0	86%	0.0	96%
DQ976615.1 Ateles geoffroyi BGLAP (BGLAP) gene, complete cds	1642	1642.0	83%	0.0	95%
DQ976876.1 Macaca mulatta BGLAP (BGLAP) gene, complete cds	1626	1626.0	82%	0.0	96%
DQ977209.1 Pan paniscus BGLAP (BGLAP) gene, partial cds	1409	1409.0	64%	0.0	99%
DQ976754.1 Lagothrix lagotricha BGLAP (BGLAP) gene, partial cds	758	758.0	54%	0.0	85%
XM_012658608.1 PREDICTED: Propithecus coquereli bone gamma-carboxyglutamate (gla) protein (BGLAP), mRNA	293	638.7	46%	9e-75	83%

NG_047015.1 Homo sapiens bone gamma-carboxyglutamate protein (BGLAP), RefSeqGene on chromosome 1

Score:	Expect:	Identities:	Positives:	Gaps:
2221	0.0	100%	N/A%	0%

```
Query    1     G G C A G A T T C C C C C T A G A C C C G C C C G C A C C A T G G T C A G G C A T G C C C C T C C T C A T C G C T G G G C A C A G C C C A G A
               | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Subject  4923  G G C A G A T T C C C C C T A G A C C C G C C C G C A C C A T G G T C A G G C A T G C C C C T C C T C A T C G C T G G G C A C A G C C C A G A
```

4) Reproducibility

- ➜ Challenges
  - ○ Original data may be modified or deleted by the researchers.
  - ○ Data may get corrupted by transfer processes.
  - ○ Versions of software tools change, services become unavailable or software used may become proprietary.

4) Reproducibility

- ➜ Solution
  - ○ Maintain an audit trail with technical metadata.
  - ○ A separate thread to record details each time the workflow is updated.
  - ○ Each step recorded in a backend relational database and accessed whenever the workflow needs to be reproduced.

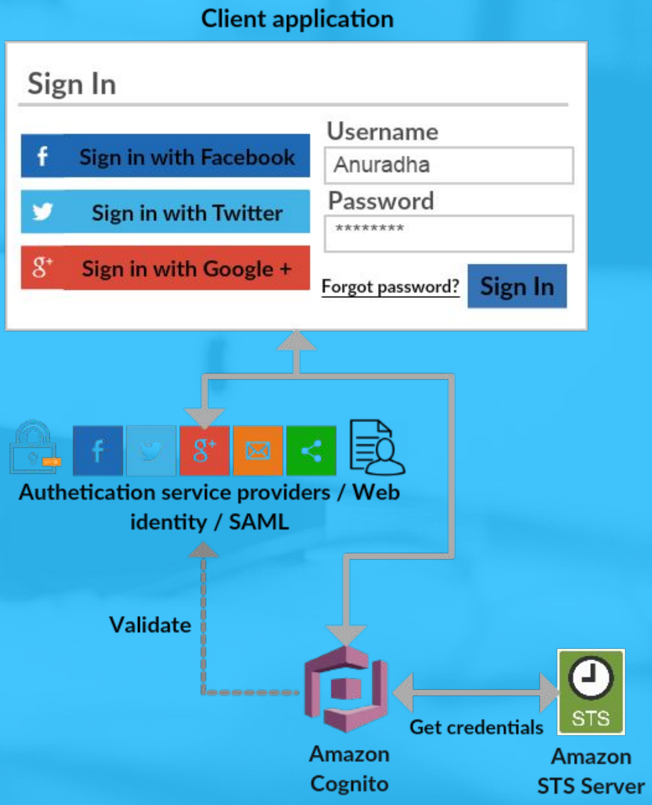
5) User management

- ➜ Importance of having a proper user management and authentication system,
  - ○ To track and share individual analyses
  - ○ To keep track of user data
  - ○ To process quotas

5) User management

➜ Amazon Cognito

- ○ User registration & authentication
- ○ Data synchronization



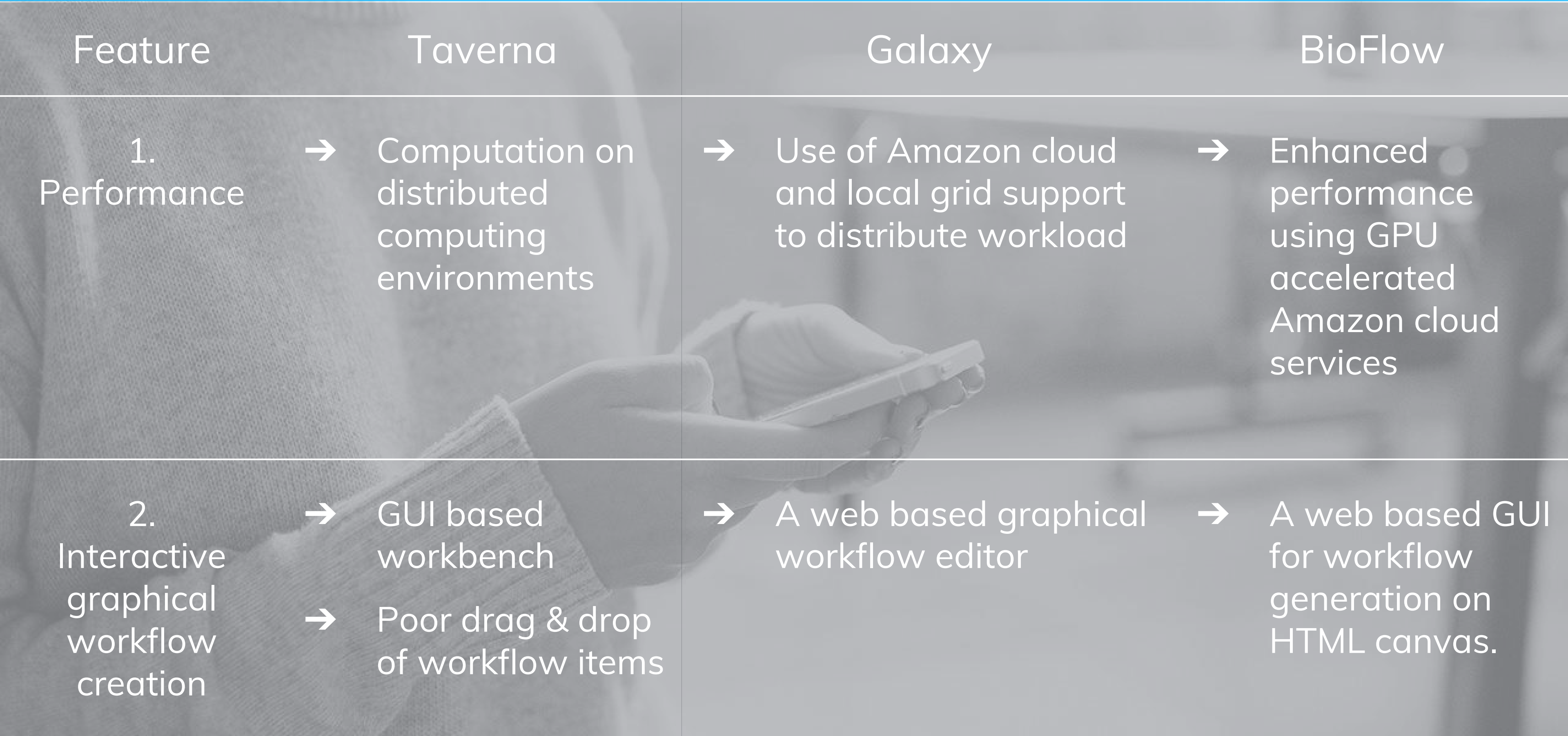
Client application

Sign In

Username: Anuradha

Password: ********

Summary of comparison of features with existing systems

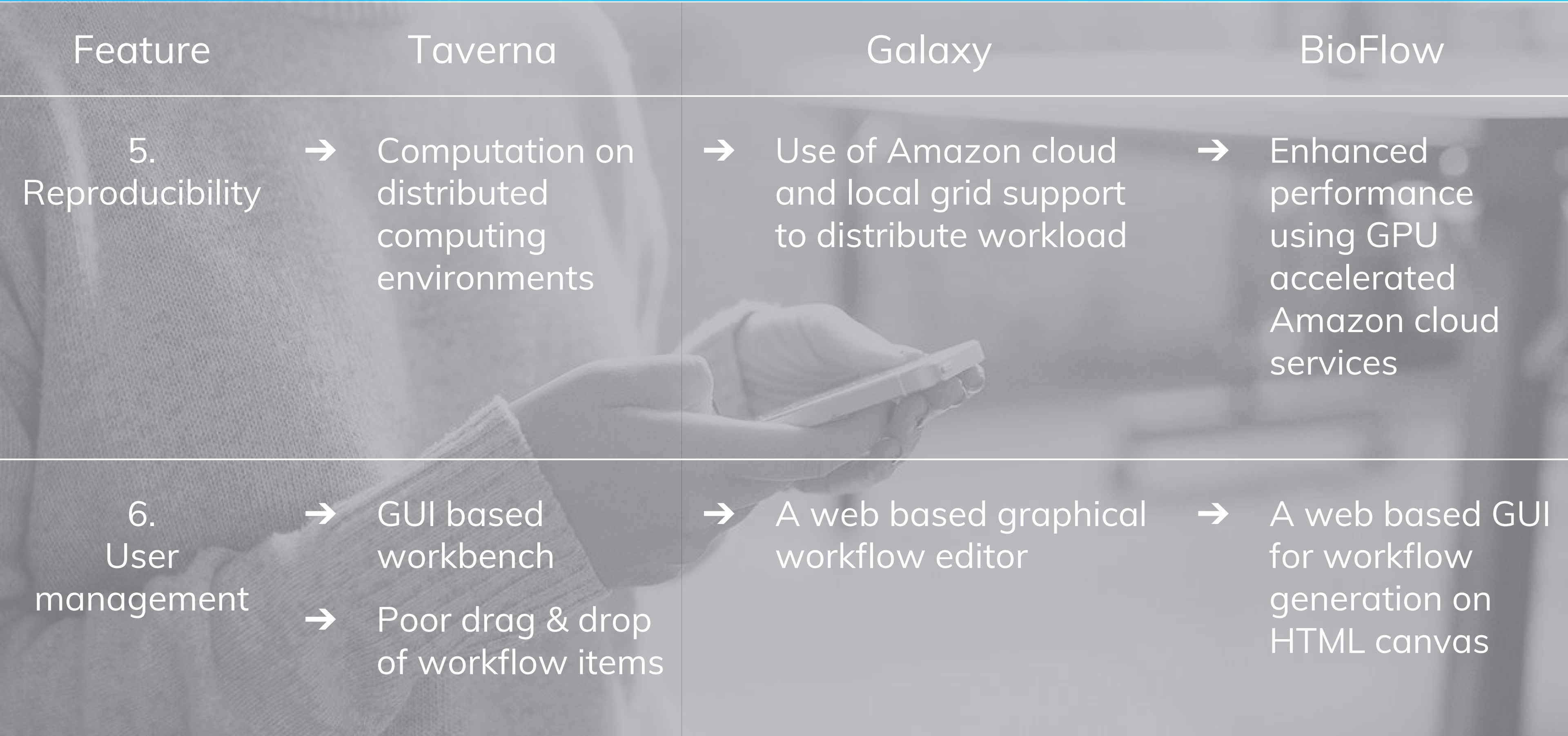


Feature	Taverna	Galaxy	BioFlow
1. Performance	→ Computation on distributed computing environments	→ Use of Amazon cloud and local grid support to distribute workload	→ Enhanced performance using GPU accelerated Amazon cloud services
2. Interactive graphical workflow creation	→ GUI based workbench → Poor drag & drop of workflow items	→ A web based graphical workflow editor	→ A web based GUI for workflow generation on HTML canvas.

Summary of comparison of features with existing systems

Feature	Taverna	Galaxy	BioFlow
3. Module Extensibility	→ Computation on distributed computing environments	→ Use of Amazon cloud and local grid support to distribute workload	→ Enhanced performance using GPU accelerated Amazon cloud services
4. Reporting	→ GUI based workbench → Poor drag & drop of workflow items	→ A web based graphical workflow editor	→ A web based GUI for workflow generation on HTML canvas.

Summary of comparison of features with existing systems



Feature	Taverna	Galaxy	BioFlow
5. Reproducibility	→ Computation on distributed computing environments	→ Use of Amazon cloud and local grid support to distribute workload	→ Enhanced performance using GPU accelerated Amazon cloud services
6. User management	→ GUI based workbench → Poor drag & drop of workflow items	→ A web based graphical workflow editor	→ A web based GUI for workflow generation on HTML canvas

Evaluation and Results

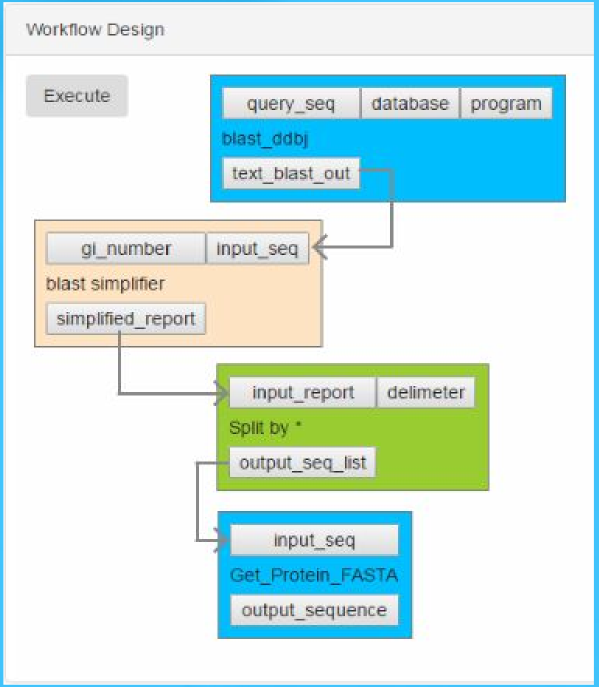
Performance Evaluation

- Workflow run on top of a GPU enabled Amazon EC2 Linux instance, having the following CPU and GPU specifications.

CPU	GPU
Intel(R) Core(TM) i5-2450M CPU	Nvidia GeForce GT 525M
2 cores, @ 2.50 GHz	96 CUDA Cores
4 GB RAM	1 GB RAM

Performance Evaluation

- ➜ Workflow executed by inputting different lengths of query sequences
- ➜ Used both remotely installed ncbi-blast and GPU-Blast
- ➜ GPU-Blast
  - ○ Accelerate gapped & ungapped protein sequence alignments



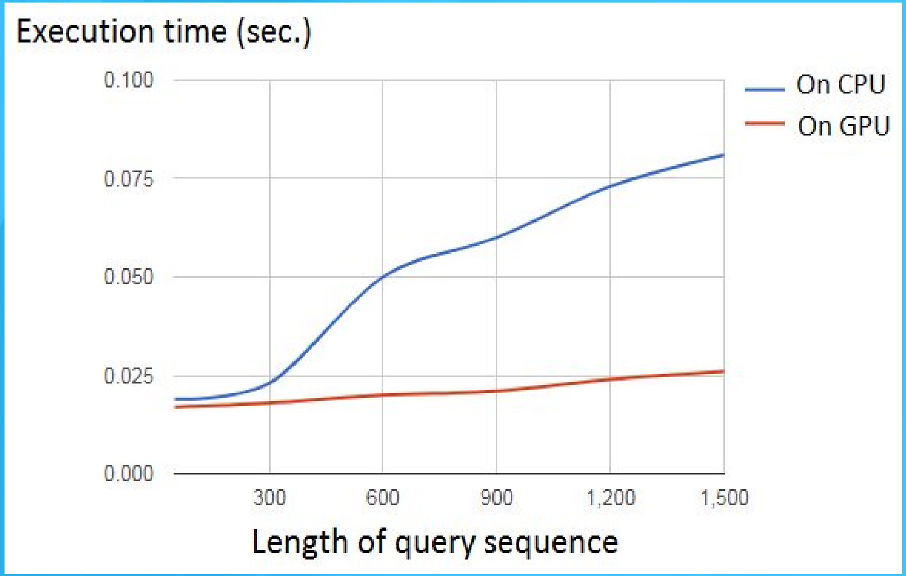
Performance Evaluation

➜ Evaluation results

Length of input seq.	Time on CPU (sec.)	Time on GPU (sec.)	Speedup ratio
50	0.019	0.017	1.12
300	0.023	0.018	1.28
600	0.050	0.020	2.50
900	0.060	0.021	2.86
1200	0.073	0.024	3.04
1500	0.081	0.026	3.11

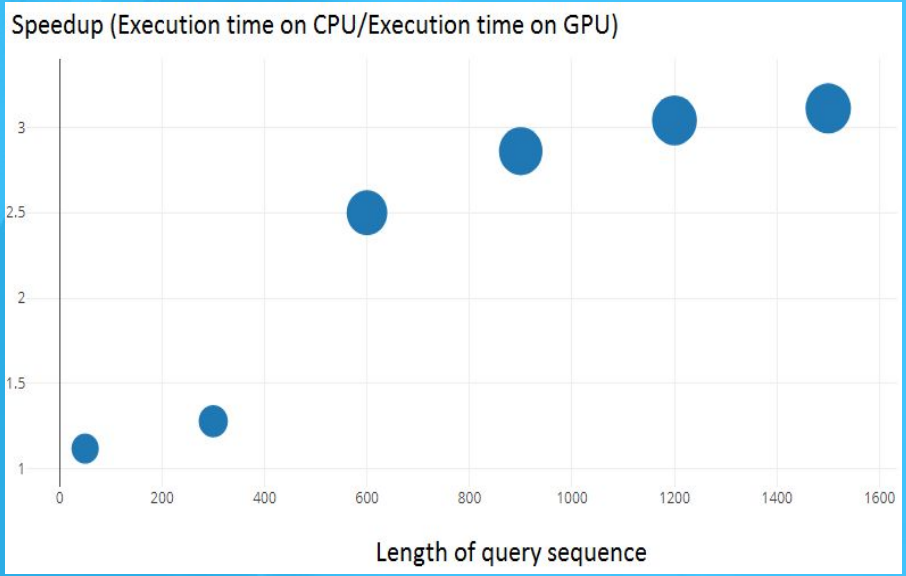
Performance Evaluation

- ➜ Comparison between executions times on CPU vs GPU



Performance Evaluation

- Avg. GPU speedup for different lengths of input query seq.



Performance Evaluation

- ➜ Observations
  - ○ When input length increases, speedup ratio also increases
  - ○ Significant increase in performance cannot be observed when the query length is small
  - ○ In long input queries, about 3 fold increase in performance can be obtained through GPU acceleration

Usability Evaluation

- ➜ System Usability Scale (SUS)
  - ○ Subjects: 10 subjects aged 20-30 having basic knowledge in computing and bioinformatics
  - ○ Compared against Taverna Workbench
- ➜ Open-ended interview

SUS Scores

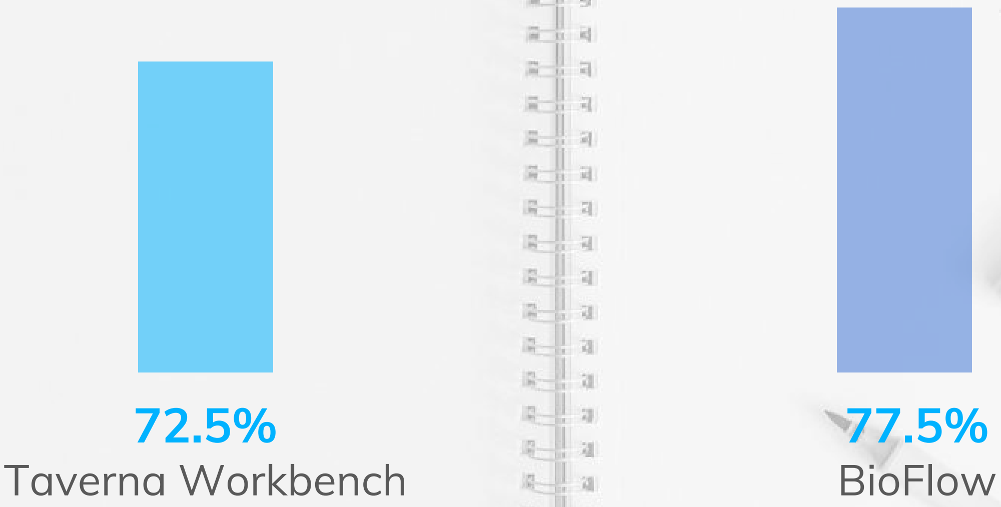


Diagram featured by http://slidemodel.com

Interview responses

→ Inability to drag individual components in Taverna system makes it inflexible to visualize the workflow the way we want.

→ As the Taverna system is desktop based, it has certain dependencies to be pre-installed in the user's local machine, which makes it cumbersome to use.

Conclusions

Research Outcomes

An interactive workflow generation software

Ability to process massive datasets in parallel

Significant increase in speed of execution

More desirable features of usability

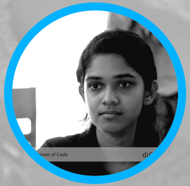
Attracts users with minimal programming experience

Project Demo

Future Extensions

- ➜ Exploring applicability of Amazon EC2 FPGA based computing instances to create custom hardware accelerations for the application
- ➜ Inclusion of more features,
  - ○ Sharing of workflows
  - ○ Pipeline comparison
  - ○ Citation support
- ➜ Development of comprehensive user support and interface enhancements

The Team



Anuradha Welivita

Lecturer (Contract)
Dept. of Comp. Sci. & Eng.
University of Moratuwa
Sri Lanka

Dr. Indika Perera

Senior Lecturer
Dept. of Comp. Sci. & Eng.
University of Moratuwa
Sri Lanka

Dr. Dulani Meedeniya

Senior Lecturer
Dept. of Comp. Sci. & Eng.
University of Moratuwa
Sri Lanka

Anuradha Wickramarachchi

Undergraduate
Dept. of Comp. Sci. & Eng.
University of Moratuwa
Sri Lanka

Vijini Mallawarachchi

Undergraduate
Dept. of Comp. Sci. & Eng.
University of Moratuwa
Sri Lanka

Thanks!

Any questions?

You can find me at:
@AnuradhaKW
anuradha@cse.mrt.ac.lk