A Lightweight Supercomputing Web Portal for Inferring Phylogenetic Trees

David Johnson and Andrew Meade
School of Biological Sciences
University of Reading
Whiteknights
Reading, RG6 6BX, UK
{d.johnson, a.meade}@reading.ac.uk

Vassil N. Alexandrov
School of Systems Engineering
University of Reading
Whiteknights
Reading, RG6 6AY, UK
v.n.alexandrov@reading.ac.uk

Abstract

In this paper we describe a lightweight Web portal developed for running computational jobs on a IBM JS21 Bladecenter cluster, ThamesBlue, for inferring and analyzing evolutionary histories. We first discuss the need for leveraging HPC as a enabler for molecular phylogenetics research. We go on to describe how the portal is designed to interface with existing open-source software that is typical of a HPC resource configuration, and how by design this portal is generic enough to be portable to other similarly configured compute clusters, and for other applications.

1. Introduction

High performance computing (HPC) resources are not always easy to use for scientists as many of the tools for accessing, submitting and managing compute jobs is typically done through command-line interfaces (CLI). This makes for a difficult paradigm shift for many users who are trained and experienced to use only graphical user interfaces (GUI) afforded by modern operating systems.

To make analyses that run on HPC resources, such as compute clusters, supercomputers, and Grids, more accessible to scientists, Web-based GUI interfaces have been developed as the Web can be considered as a ubiquitously accessible platform. Web applications can be developed with relatively little dependence on underlying end-user technology as Web content presentation standards are widely adopted by browser software. As such, Web applications have been developed to hook into HPC resources to remove the need for users having to use CLI. We developed a Web portal that interfaces with our own HPC resource, ThamesBlue, to enable biologists to more easily leverage HPC using a specific phylogenetic inference software application, BayesPhylogenies.

2. HPC in Phylogenetics Research

Molecular phylogenetic inference, the creation of evolutionary histories from molecular data, is an activity universal to all biological disciplines and currently growing at a quadratic rate (see figure 1). The current rate of growth is over 3000 published articles using molecular phylogenetices per annum. This growth is fuelled by the recognition that phylogenies are a fundamental tool in biological data analysis, the advancement in molecular sequencing technology, universal access to online resources such as the National Cener for Biotechnology Information (NCBI) (with over 80 million distinct gene sequences) and the availability of suitable software and powerful hardware needed for the complex calculations. These factors have made phylogenetic inference a common activity.

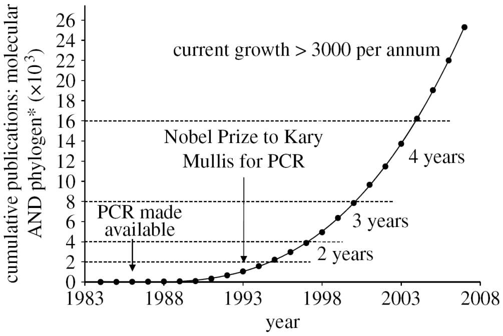


Figure 1. Growth of molecular phylogenetic research literature. [1]

The increase in molecular data as well as more accurate and statistically based analytical models has created an ever widening gap between data and methods which researchers would like to use and the availability of hardware and software needed to analyze the data in a reasonable amount of time. A typical analysis currently requires days or even weeks to complete and larger data sets may require months or even years. While scalable algorithms have been developed their uptake by the biological community is not common place. This is most likely due to a number of factors, such as, large start up and running costs associated with high performance computing, the technical expertise needed to install and maintain a system and the ability to effectively connect researches in one field to the resource.

Our Web portal aims to provide an easy to use interface for a specific phylogenetic application, BayesPhylogenies. BayesPhylogenies is software package that analyzes sequence data to infer phylogenetic trees using Bayesian Markov Chain Monte Carlo (MCMC) or Metropolis-coupled MCMC (MCMCMC) methods. It implements a number of different models of evolution including a 'mixture-model' that allows multiple models to be applied to sequence evolution. The details of BayesPhylogenies will not be discussed in this paper as it is described in full by Meade and Pagel in their 2004 paper [2].

3. Related Work

Accessing HPC resources can be simplified for users by providing GUI-based applications, and there are a number of projects that aim to provide user-friendly interfaces for scientists using HPC resources.

GridSphere aims to provide a standards based framework for producing Web portals to access HPC and Grid resources [3]. GridSphere provides portlet containers based on the JSR 168 Portlet API [4] with a view to providing a framework on which developers can produce small Web applications components that are portable between JSR 168 compliant application servers. The core components that are bundled with GridSphere include a set of ready-to-run portlets that implement basic functionality to allow users to hook into HPC resources. Although GridSphere provides the building blocks for constructing a Web portal to access compute resources, we decided that it was too complex for our purposes of developing our own application specific Web portal.

g-Eclipse is a graphical workbench for accessing HPC resources, in particular Grid and Cloud computing infrastructures [5]. The workbench is a GUI-based desktop application and aims to provide a set of generalized tools to access data and computing resources using a system independent abstraction of the Grid. g-Eclipse interfaces with specific systems through implementing system specific functionality to communicate with Grid or Cloud middleware. Although g-Eclipse could be set up to interface with our own cluster's configuration, it was determined that the workbench functionality is too generalized where the target users will not need most of the tools provided by the workbench. There is the possibility of packaging the workbench software into an application-specific interface, however we preferred a Web based approach to allow users more ubiquitous access via their Web browsers.

4. Portal Architecture

Although these existing solutions exist as described in the previous section, these frameworks aim to be generic solutions to interfacing with a wide variety of back-end systems. Our priority was to cater for a specific configuration to operate with our own HPC resource, ThamesBlue.

The ThamesBlue head-node configuration uses the Portable Batch System (PBS) submission and queueing middleware [6]. PBS is used for performing common tasks for submitting and monitoring jobs on compute clusters and typically a user will log in to a CLI session via a secure shell to run specific commands. To this end, we designed the Web portal to be able to log into the head-node and execute PBS commands on behalf of a user accessing our Web portal. This approach reduces the complexity of setting up and running a job from a user's perspective.

The Web portal itself leverages several open-source technologies to fulfil all of the major functionality desired for a user wanting to run a BayesPhylogenies analysis:

- *Apache Wicket* - To aid in rapid Web application development, we chose to use Wicket [7], an open-source Web application framework developed and maintiained by the Apache Software Foundation (ASF) [8]. Wicket applications are compiled as Java servlets that are deployed as self-contained archive files onto any Web application server that supports Java Enterprise Edition Web container.

- *Apache Derby* - To manage the data generated by users and by the BayesPhylogenies executable binary, the Web portal uses a relational database. Derby is a typical database implementation that accessed and manipulated using SQL [9], and like Wicket, is developed and maintained by the ASF.

- *Glassfish* - In order to deploy the portal Web application, we chose to host it on the Glassfish application server [10]. Glassfish is maintained by Sun Microsystems and includes a Web container in which to run servlet-based Web applications, and also includes Derby as part of the application server's standard services. By having a database embedded in the application server, we aim to reduce the number of dependencies of the portal's code on other projects.

The portal uses several other open-source Java and CLI utilities that will not be described in detail here which include *GanymedSSH2* [11] to provide secure shell client access via Java, *rsync* [12] to bridge users to job outputs lying on the PBS head-node, and *cron4j* [13] for scheduling background tasks such as executing polling the head-node for recent changes in job status'.

A key requirement for the Web portal are for the portal server to be able access the PBS head-node via a secure shell connection. In addition to this, the portal must be able to execute the rsync tool on the head-node. Apart from those two requirements, the portal Web application can be self-contained on any Glassfish application server. It has also been designed in such a way that if the compute resource becomes unavailable users can still access the last available information. Both the job details and any check-pointed output files are cached on the Web portal server. In contrast to traditional CLI access the tasks of checkpointing and checking the job status' would normally be done manually. Accessing the PBS head-node is through minute-by-minute polling to check job status' and to synchronize the output files using rsync, so a user's job information and outputs are backed up on the Web portal automatically.

5. Portal Usage

From a user perspective setting up a new BayesPhylogenies job is a straightforward four step process using a multipage wizard form. The steps are as follows:

1. Upload a Nexus data file.

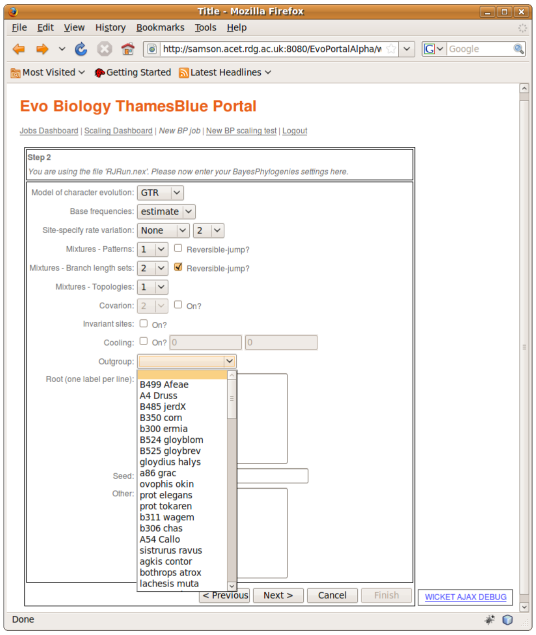
2. Configure the BayesPhylogenies model settings (illustrated in figure 2).

3. Configure the job settings.

4. Review settings and submit job to the cluster.

The wizard provides some degree of automation to aid the user in setting up an analysis job. At step 1, the Nexus input file is parsed to ensure that the file is in the correct format. Where a user may upload a Nexus file which has previously had a BayesPhylogenies control block appended to its content, the portal takes care of stripping the control block from the file in order to allow the user to configure the submission correctly using the wizard. In step 2, the user is presented with a form allowing configuration of the model settings that they wish to use. In step 3, the job settings are configured. This is where the user specifies the length of time they wish their analysis job to run, the frequency of check-pointing the output data, and the number of processors to run the job on.

Batch jobs can be set up as well at this stage, where the user can specify to submit multiples of jobs with identical model settings (but with different random seeds). The details of the settings are found in the help pages within the portal, and also in the BayesPhylogenies user manual.



Evo Biology ThamesBlue Portal

Jobs Dashboard | Scaling Dashboard | *New BP job* | New BP scaling test | Logout

Step 2

You are using the file 'RJRun.nex'. Please now enter your BayesPhylogenies settings here.

Model of character evolution: GTR
Base frequencies: estimate
Site-specify rate variation: None 2
Mixtures - Patterns: 1 □ Reversible-jump?
Mixtures - Branch length sets: 2 ☑ Reversible-jump?
Mixtures - Topologies: 1
Covarion: 2 □ On?
Invariant sites: □ On?
Cooling: □ On? 0 0
Outgroup:
Root (one label per line):
Seed:
Other:

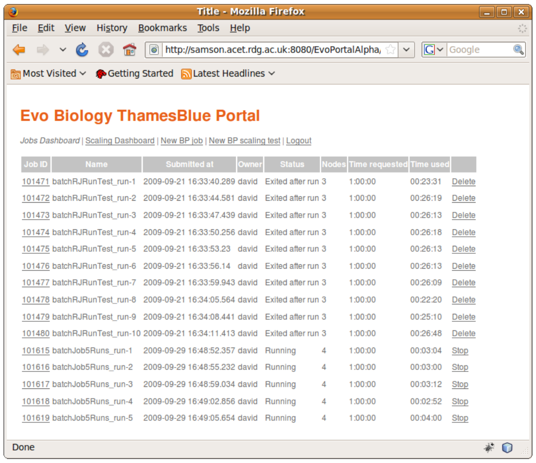
B499 Afeae
A4 Druss
B485 jerdX
B350 corn
b300 ermia
B524 gloyblom
B525 gloybrev
gloydius halys
a86 grac
ovophis okin
prot elegans
prot tokaren
b311 wagem
b306 chas
A54 Callo
sistrurus ravus
agkis contor
bothrops atrox
lachesis muta

Figure 2. Screenshot of the portal (prototype) showing an example of the BayesPhylogenies configuration form.

Users can monitor the status of their submitted jobs via a dashboard, as shown in figure 3 that shows an overview of all jobs they are running and have run in the past. Extra details including the check-pointed output files can also be accessed during runtime allowing for users to check the most recent iterations of their output trees, where if necessary the user can stop their job before it comes to its natural end or used all of its allocated run time. As BayesPhylogenies jobs can run up to several weeks or months, the portal is designed to email reminders to users of running jobs, and also notifies users when their job has stopped (either naturally or erroneously).

6. Conclusions

Molecular phylogenetics research demands new methods and resources to handle the growing pool of data needing to be processed in a timely manner. By providing a simple Web portal for biology researchers to access a phylogenetic inference software package (BayesPhylogenies) on a HPC resource (ThameBlue) we aim to promote the use of HPC amongst the local biology community. The value added by being able to infer evolutionary relationships in a matter of days and weeks rather than months or


Title - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://samson.acet.rdg.ac.uk:8080/EvoPortalAlpha,
Google
Most Visited
Getting Started
Latest Headlines
Done

Evo Biology ThamesBlue Portal

Jobs Dashboard | Scaling Dashboard | New BP job | New BP scaling test | Logout

Job ID	Name	Submitted at	Owner	Status	Nodes	Time requested	Time used	
101471	batchRJRunTest_run-1	2009-09-21 16:33:40.289	david	Exited after run	3	1:00:00	00:23:31	Delete
101472	batchRJRunTest_run-2	2009-09-21 16:33:44.581	david	Exited after run	3	1:00:00	00:26:19	Delete
101473	batchRJRunTest_run-3	2009-09-21 16:33:47.439	david	Exited after run	3	1:00:00	00:26:13	Delete
101474	batchRJRunTest_run-4	2009-09-21 16:33:50.256	david	Exited after run	3	1:00:00	00:26:18	Delete
101475	batchRJRunTest_run-5	2009-09-21 16:33:53.23	david	Exited after run	3	1:00:00	00:26:13	Delete
101476	batchRJRunTest_run-6	2009-09-21 16:33:56.14	david	Exited after run	3	1:00:00	00:26:13	Delete
101477	batchRJRunTest_run-7	2009-09-21 16:33:59.943	david	Exited after run	3	1:00:00	00:26:09	Delete
101478	batchRJRunTest_run-8	2009-09-21 16:34:05.564	david	Exited after run	3	1:00:00	00:22:20	Delete
101479	batchRJRunTest_run-9	2009-09-21 16:34:08.441	david	Exited after run	3	1:00:00	00:25:10	Delete
101480	batchRJRunTest_run-10	2009-09-21 16:34:11.413	david	Exited after run	3	1:00:00	00:26:48	Delete
101615	batchJob5Runs_run-1	2009-09-29 16:48:52.357	david	Running	4	1:00:00	00:03:04	Stop
101616	batchJob5Runs_run-2	2009-09-29 16:48:55.232	david	Running	4	1:00:00	00:03:00	Stop
101617	batchJob5Runs_run-3	2009-09-29 16:48:59.034	david	Running	4	1:00:00	00:03:12	Stop
101618	batchJob5Runs_run-4	2009-09-29 16:49:02.856	david	Running	4	1:00:00	00:02:52	Stop
101619	batchJob5Runs_run-5	2009-09-29 16:49:05.654	david	Running	4	1:00:00	00:04:00	Stop

Figure 3. Screenshot of the portal (prototype) showing the job monitoring dashboard.

years is potentially great, and without easy graphical and ubiquitous interfaces for users to access HPC facilities the widespread acceptance of using such resources is hindered. The production of our Web portal is a first step to overcoming this hindrence. We expect to provide access to our HPC resource in the near future, and to open-source the portal software to the biology community. More information on BayesPhylogenies and the Web portal can be found at http://www.evolution.reading.ac.uk

7. Acknowledgements

This project was supported in part by the University of Reading's Research Endowment Trust Fund (RETF). BayesPhylogenies was developed with support in part by the UK Biotechnology and Biological Sciences Research Council (BBSRC).

References

[1] M. Pagel and A. Meade, ”Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo,” in *Philosph. Trans. Roy. Soc. B: Biological Sciences*, vol. 363, no. 1512, pp. 39-55, 2008.

[2] A. Meade and M. Pagel, ”A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data,” in *Syst. Biol.*, vol. 53, pp. 571-581, 2004.

[3] J. Novotny et al, ”GridSphere: an advanced portal framework,” in *Proc. 30th Euromicro Conf.*, Rennes, France, pp. 412-419, 2004.

[4] ”JSR 168: Portlet Specification,” 2009 [Online]. Available: http://jcp.org/en/jsr/detail?id=168 [Accessed: Oct. 14, 2009]

[5] H. Gjermundrød et al, ”g-Eclipse - An Integrated Framework to Access and Maintain Grid Resources,” in *Proc. 9th IEEE/ACM Int. Conf. Grid Computing*, Tsukuba, Japan, 2008, pp. 57-64.

[6] R.L. Henderson, ”Job scheduling under the Portable Batch System,” in *Job Scheduling Strategies for Parallel Process.*, Lecture Notes in Comput. Sci. vol. 949, pp. 279-294, 1995.

[7] ”Apache Wicket,” 2009. [Online]. Available: http://wicket.apache.org [Accessed: Oct. 14, 2009].

[8] ”The Apache Software Foundation,” 2009. [Online]. Available: http://www.apache.org [Accessed: Oct. 14, 2009].

[9] ”Apache Derby,” 2009. [Online]. Available: http://db.apache.org/derby/ [Accessed: Oct. 14, 2009].

[10] ”Glassfish,” 2009. [Online]. Available: https://glassfish.dev.java.net [Accessed: Oct. 14, 2009].

[11] ”Ganymed SSH-2 for Java,” 2006. [Online]. Available: http://www.ganymed.ethz.ch/ssh2/ [Accessed: Oct. 14, 2009].

[12] A. Tridgell, ”Efficient Algorithms for Sorting and Synchronization,” Ph.D. dissertation, Australian Nat. Univ., Canberra, Australia, 1999.

[13] ”cron4j,” 2009. [Online]. Available: http://www.sauronsoftware.it/projects/cron4j/ [Accessed: Oct. 14, 2009].