Bioinformatics Application Note:

CONTEXT – A Phylogenomics Dataset Browser

Joe Parker[1,2] and Stephen J. Rossiter[2]

1. Kitson Consulting, Bristol, UK; Present address: Jodrell Laboratory, Royal Botanic Gardens, Kew, UK

2. School of Biological and Chemical Sciences, Queen Mary University of London,

UK.

Word count: 555

Corresponding Author:
Joe Parker
Jodrell Laboratory,
Royal Botanic Gardens, Kew,
TW9 3DS, UK
Tel. +44 20-8332-5063
Fax +44 20-8332-5197
joe.parker@kew.org

Project email: joe+CONTEXT@kitson-consulting.co.uk

Abstract

Summary. Quality control (QC) in large phylogenomic datasets is a key requirement for reliable and reproducible research in evolution, adaptation, speciation and taxonomy. CONTEXT is a browser for high-throughput visualisation and comparative QC of phylogenomic datasets, consisting of a Java API and an executable binary jarfile with graphical user interface (GUI). The tool allows users to rapidly and easily visualise thousands of multiple sequence alignments and hundreds of phylogenies using a GUI to identify outliers which could affect downstream analyses. CONTEXT calculates a variety of downstream statistics on alignments and phylogenies including entropy, informativeness, imbalance, signal:noise and size.

Motivation. Comparative genomics studies have become increasingly common, but these analyses are sensitive to the quality and heterogeneity of input datasets (multiple sequence analyses and phylogenies). Currently few tools exist to readily compute descriptive statistics, or to visualise large numbers of input datasets. CONTEXT is a phylogenomics dataset browser which facilitates these analyses in a lightweight application. It allows any user to rapidly visualise, inspect, score, and sort input datasets to identify outlying datasets which may need additional processing, filtering, or masking from further analyses.

Results. The application has been successfully implemented on a variety of infrastructures. A variety of common input data formats including FASTA, Phylip/PAML, Nexus, and Newick conventions are automatically read and parsed.

Availability and implementation. The API is implemented in native Java code, available online at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline. The executable binary can be downloaded at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/tree/master/trunk/bin. The project page is at https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/blob/master/CONTEXT.md

Contact. joe.parker@kew.org

Introduction

Features and implementation

The API elements contain resources for phylogenomics such as input/output and parsing utilities; trimming, pruning and validation methods for alignments and phylogenies; statistics for evaluating alignments, phylogenies, likelihood fits and dN/dS values; UI elements including two main GUI platforms; post-

processing including linear regression and descriptive parametric statistics on large distributions of small floating-point numbers.

Evaluation

In operation, CONTEXT was able to display up to XXX alignments of YYY taxa and ZZZ sites on a SSS system with RRR RAM requirements. Example usage statistics shown in Table 1.

CONTEXT has been successfully tested on Java Virtual Machines at 1.6 and above on the following operating systems / hardware / CPU clock / RAM:

Ubunte MATE / Raspberry Pi 2 Model B+ / ARM v7 @ 0.9GHz
Windows 7 / Toshiba Portege

	OS	**Arch**	**CPU type, clock GHz**	**cores**	**RAM Gb**	**HDD Gb**
pandanus	Ubuntu / Biolinux	i686	Xeon E5620 @ 2.4	4	33 / ? / ?	1000 @ ATA 7200rpm
2	Ubuntu / MATE	ARM	ARMv7 @ 0.9	1	1 / ? / ?	8 @ SD
toshiba	Windows 7	x64	Core i7 @ 2.4	4	3 / ? / ?	128 @ SSD
MBP	Mac OSX 10.9.5	x64	Core i7 @ 2.2	4	8 / 1333 / DDR3	250 @ SSD
EC2 m4.10xlarge	Ubuntu 15.04?	x64	Xeon E5-2670 @ 2.5	16	122 / ? / ?	320 @ SSD
EC2	Ubuntu	x64	Xeon E5-2680 @ 2.8	32	60 / ? / ?	2x320 @ SSD

Table 1

Roadmap and versioning

CONTEXT is currently supplied at **Version 0.8 prerelease.**

Acknowledgements

This work has been funded by BBSRC at QMUL, specifically the methodological innovations for convergence detection methods correctly controlling for false positives (essential in genomic datasets) and a core API to implement these and facilitate handling genomic sequence data, carried out

principally by Dr. Parker (with input from Prof. Rossiter (PI), Drs. James Cotton & Elia Stupka (Co-I) and Dr. Tsagkogeorga (PDRA)) under BBSRC # BB/H017178/1.

Figures / data / tables

Table 1: Example system resource usage. The RAM usage (in megabytes) and average load time of the Phylogenomic Dataset Browser under a variety of test computer architectures and input datasets.

Test case	Mac OSX 10.9, 2.2GHz core i7, 8Gb 1333MHz DDR3 RAM, 250 Gb SSD.	Ubuntu 14.04, CPU, RAM, Memory	CentOS cluster version CPU, RAM, Memory	Windows 7, CPU, RAM, Memory	Windows XP SP3, CPU, RAM, Memory
692 Nucleotide alignments, 7 taxa, XXX-XXX (mean XXX) nt					
2,326 Nucleotide alignments, 22 taxa, XXX-XXX (mean XXX) nt					
392 Nucleotide alignments, 7 taxa, XXX-XXX (mean XXX) nt					
10 phylogenies, XXX taxa					
1000 phylogenies, XXX taxa					

Table 2

Figure 1: Phylogenomic Dataset Browser schematic. The schematic logic flow of the phylogenomic dataset browser is shown with descriptions of key analysis steps, in flow diagram format.

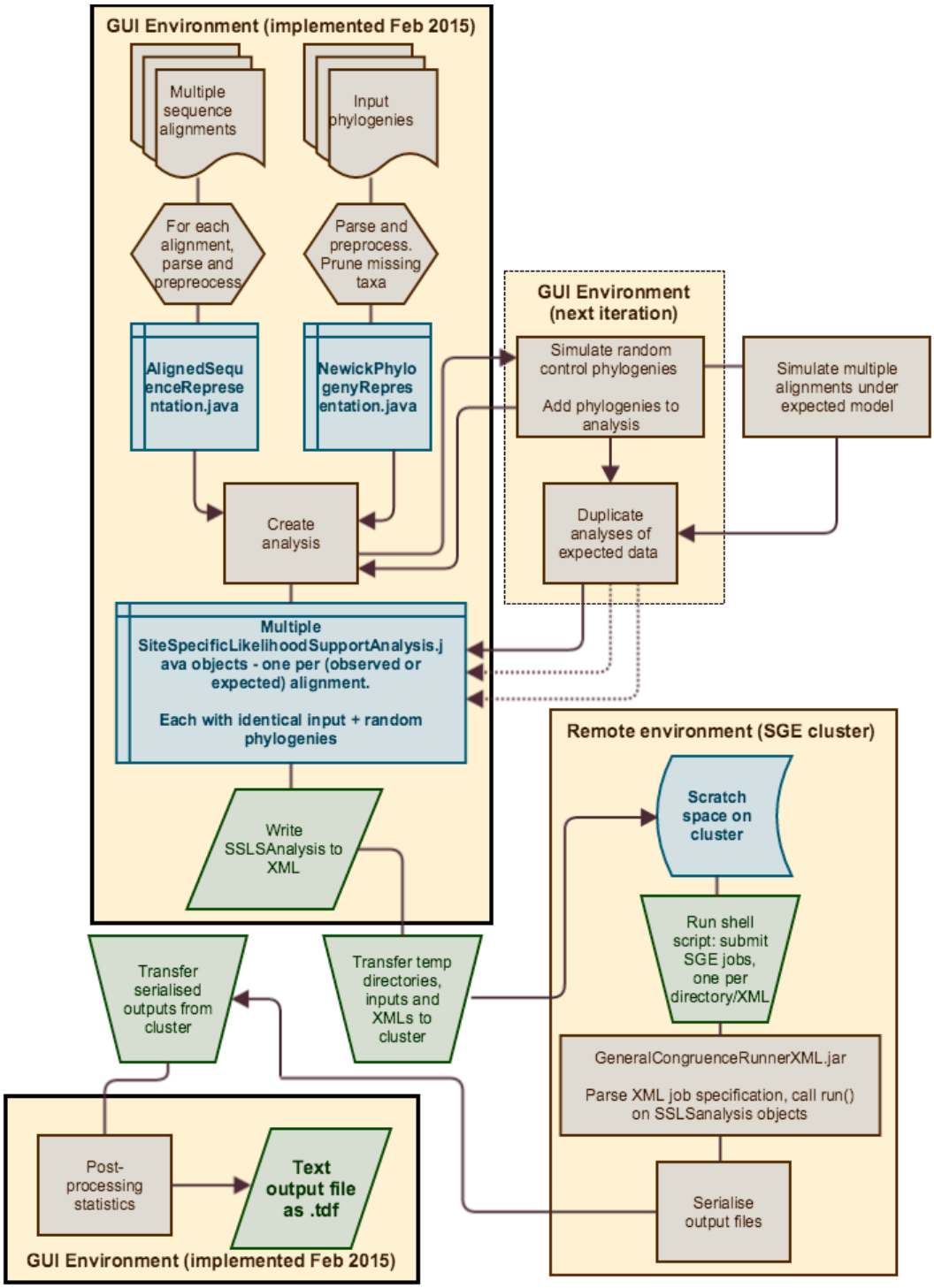
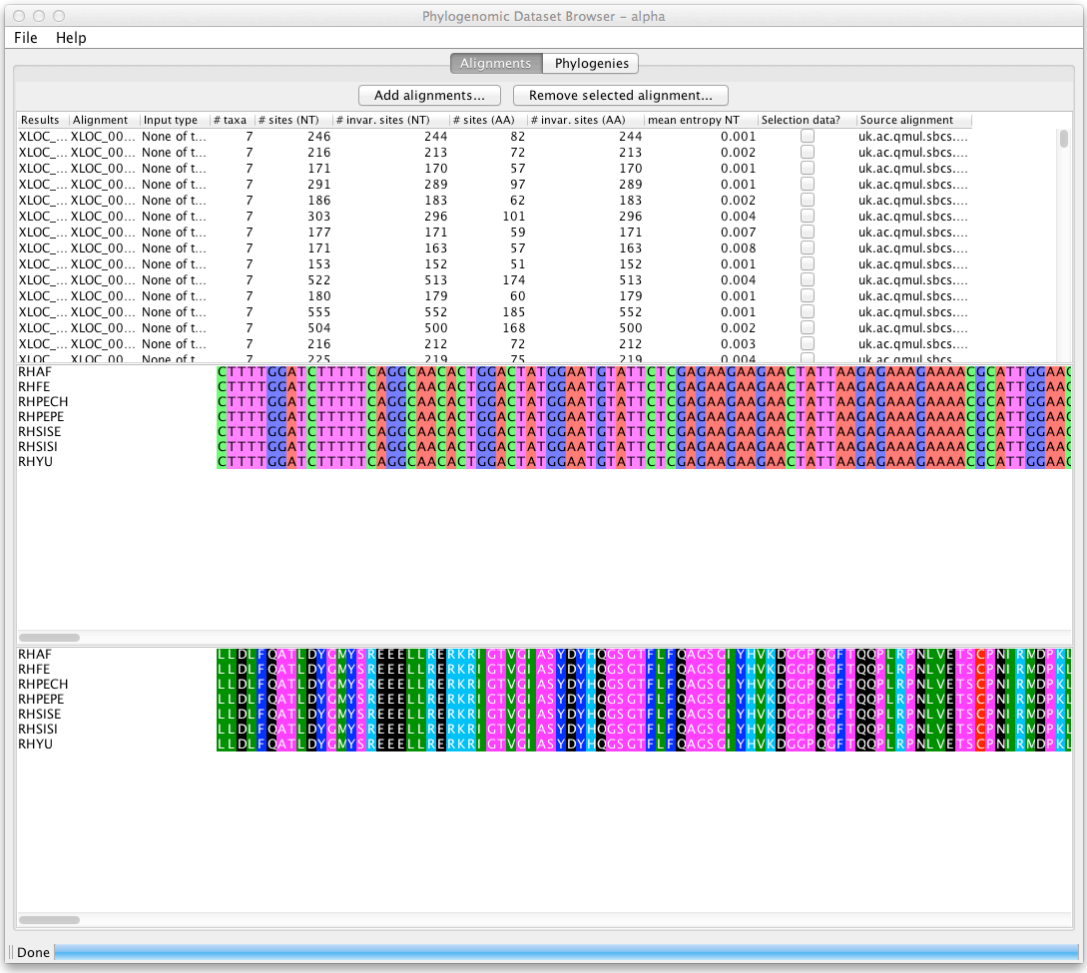
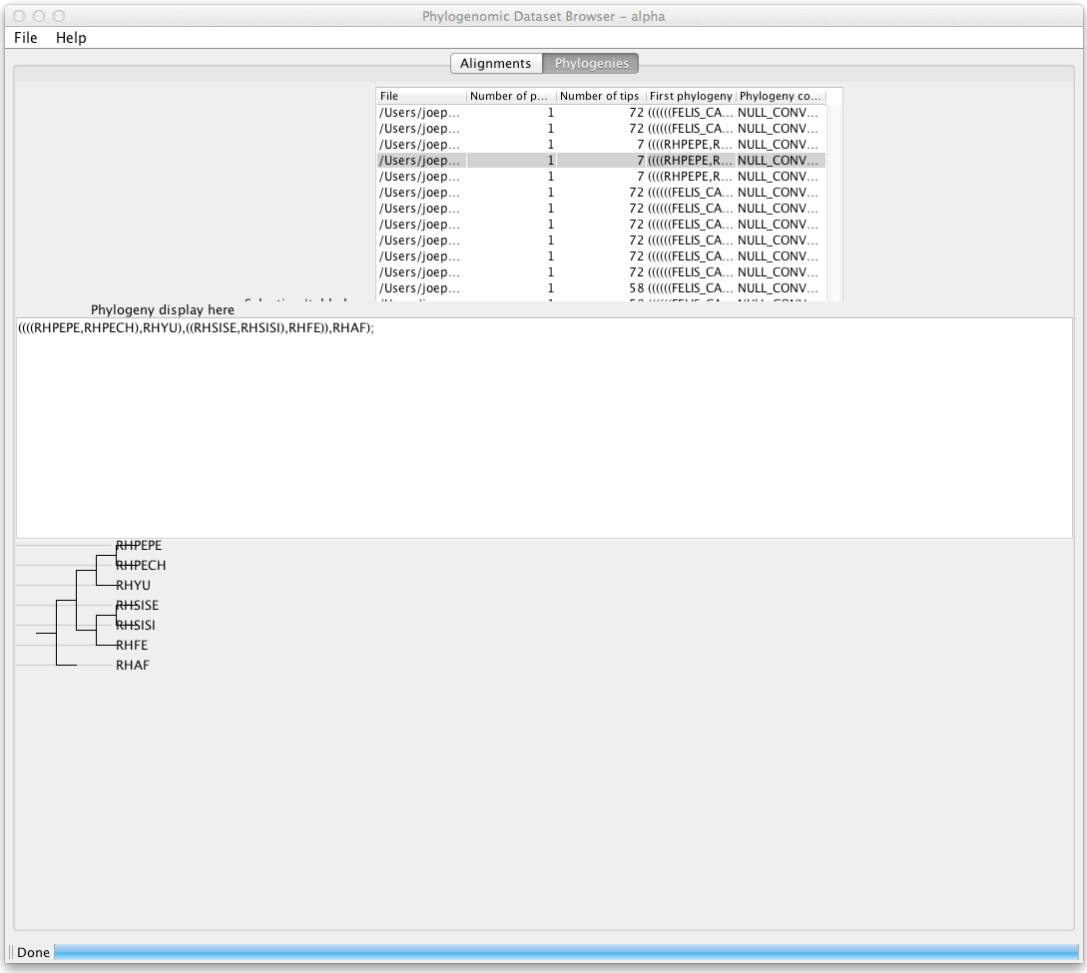


Figure 2: Screenshots showing visualisation of example datasets: (a) The alignment input screen, showing 692 multiple sequence alignments together with statistics; (b) The phylogeny input screen, showing phylogenies with graphical phylogeny display.





Footnotes

References