Title Page

Agua: a cloud bioinformatics workflow platform

Abstract

The increasing availability of huge volumes of genomic data has given rise to a crisis of reliance upon computational approaches that are scarcely reproducible and barely accessible. A lack of bioinformatics tools, personnel, workflow standards and computational power hinders the execution and replication of experiments by genomics researchers. This problem is exacerbated as genomic analysis spreads to life science disciplines where researchers are even less accustomed to high-throughput approaches. Agua http://www.aguadev.org lowers these technology barriers by providing a flexible, end-to-end cloud bioinformatics workflow solution from sequencing to genomic visualization.

Keywords

NGS, cloud, bioinformatics, workflows, visualization, JBrowse, StarCluster, Git, JSON

Article subdivisions

Introduction

Genomics plays a part in nine of the ten leading causes of death in the United States[1]. In the effort to use and understand genomic data, bioscience has become a quantitative analysis activity requiring the computational processing and interpretation of massive data sets. Biological data sets have multiplied to over 1,380 major Internet data sources in 2012[2]. These data sets generally have limited data compatibility but researchers often need to combine data from multiple sites to generate bioinformatics analyses. Workflow-based tools solve this problem by generating large and complex systems from collections of programs, data sources and structured data services[3]. As the $1,000 genome approaches[4], genomic data volumes will increase and the bottleneck will shift to the integration of clinical data with genomic data and responding to the question: How can we understand this data?
Reproducibility is a core principle of the scientific method. For an experiment to be scientific, it must be able to be accurately reproduced by others working independently from the original experimenter. Bioinformatics has suffered from a lack of reproducibility which continues to block scientific progress[5–7]. Agua addresses these issues by providing a user-centric open-source platform that facilitates sharing of workflows, reproducibility and interpretation of results.

Workflow Standards

The current availability of common tools to enable the comparison of results from different data sets[6,8,9] paradoxically underscores the difficulty faced by scientists without programming experience. Software libraries such as Bioperl, Biopython and Bioconductor, and bioinformatics toolkits such as EMBOSS[10,11] and NCBITools[12] are valuable contributions in terms of content and approach. However, they do not provide a generalized solution to allow scientists with minimal informatics skills to run existing workflows or generate novel workflows.

BioExtract[13], KNIME[14], Galaxy[15], GenePattern[16], GeneProf, Mobyle[17], and Pegasus[18] are among many

noteworthy attempts to provide a generalized solution for automated bioinformatics workflows. They are standalone or web applications that provide interactive tools for scientists to execute workflows using predefined tools and view their results in real-time. Galaxy and BioExtract also allow the user to save workflows and share data by providing web access to application-specific data objects. These data object formats are not widely adopted nor are they interchangeable with workflow data from other applications.

It is also important to note that the need for 'industry standard' workflows may have been overestimated. Genome scientists appear to be resisting the imposition of canonical data analysis methods. For example, the 1000 Genomes Project is a multinational effort to map genetic variation in major population groups around the world and is arguably the most celebrated large-scale NGS project to date. The project provides well-documented procedures for data analysis. Paradoxically, itnavigation might be expected that the adoption of a consensus on best practices would be essential to maximizing comparability of results. However, only ten out of 299 papers published in 2011 that explicitly cite the 1000 Genomes Project actually used the tools recommended by the consortium for mapping and variant discovery[5]. This suggests a more pressing need for standards for describing workflows rather than standard workflows.

Agua proposes a standardized workflow description format based on JSON and leveraging the Git version control system to provide workflow provenance. Using Agua, scientists can develop a workflow from start to finish while generating a series of snapshots of a JSON file describing the workflow. These workflow creation steps can be synchronized to a remote Git service such as Github or BitBucket. The user can store the workflow version history as a private repository or as a public repository which is visible to the world and can be referenced in a scientific paper.

Publication to a Git repository simplifies the process of sharing and reusing workflows, maximizes transparency and allows scientists to easily evaluate and reproduce each other's research. Version control also provides disaster recovery; in the event of a catastrophic local data loss, workflows can be restored by simply cloning from the remote repository.

The User-Centric Experience

User interface design is critical to the efficient use of available data and application resources. Many biological data websites and workflow GUIs suffered from drawbacks that caused unsatisfactory user experiences and impeded their adoption. Users had difficulty navigating and searching, had to navigate back and forth between forms and results, had difficulty keeping track of past actions, often could not obtain sufficient documentation of application options and were confused by ambiguities concerning data storage[1]. Agua is designed to minimize these drawbacks and maximize 'traction' in the user experience by providing a web application interface that is highly intuitive and transparent.

Agua's web interface uses the Dojo javascript toolkit (http://www.dojotoolkit.org) which contains a rich assortment of web application components and utilities. Agua's design focuses on empowering the domain-specific view of scientists by enabling easy access to the data types, tools and distributed resources that are specific to their particular research domain. Agua uses the AJAX paradigm[19] to communicate between browser and server, which has been shown to significantly enhance user satisfaction and efficiency of use[20,21].

Agua's tabular, AJAX-based design is relatively unaffected by the issues of navigation and 'data scent' associated with websites[22]. However, while websites have issues of navigation, web applications have issues of state updating and user notification of changes[23]. Agua uses a publish/subscribe model to maintain state across its various displays. The user can open multiple tabs which communicate with

each other; a change in the data in one tab is automatically propagated to all other tabs. Agua also uses visual 'standby' notifications to inform the user when processing has completed and server push to avoid server timeout errors when AJAX responses are delayed due to long-running processes on the server.

In response to the tendency among scientists towards a heterogeneity of computational approaches, Agua provides a easy-to-use way to create custom workflows. Agua enables users with basic web page interaction skills to create and run complex workflows. A simple drag and drop adds an application to a workflow. For each application, the user can use the default parameters or manually set parameter values. For custom pipelines, Agua provides unparalleled ease of implementation. Any application that can be invoked from the Linux command line can be added. New applications are deployed in two simple steps: 1) A user adds an application to the server filesystem, 2) the administrative user inputs the application details into a simple form in the 'Apps' tab. Required details are the application's name, location and the name of the package containing the application.

The administrative user can sync a package to a public Git repository to share all its applications with the global research community. GitHub and BitBucket are free for open-source projects and have many features which facilitate teams working on software or documents. Publishing workflows using distributed repositories avoids dependence on accurate and uniform descriptions of biological data types and relations for which there is no commonly-accepted ontology [28–30] or language, despite some developments[31]. Agua also implements a package installation and version control method based on a proposed Open Package Schema which can also be published to distributed repositories.

Data Transparency

Researchers, funders and journals are in broad agreement that data must be accessible to support the conclusions of scientific publications and for the research to have impact. What is lacking is agreement on timing, formatting and attribution[24].

The central metaphor of the user interface is symmetry between the logical structure and the file structure. Each project represents a folder of the same name containing one or more subdirectories, each representing a workflow of the same name. Each workflow is composed of one or more applications whose input data resides in the workflow folder. Each application has inputs, resources and outputs. This simple and intuitive symmetry is designed to increase transparency and facilitate data navigation. Agua's 'Folders' tab, which is similar in appearance to a desktop file explorer, enables the user to browse, delete, upload, download and view inside files.

Data provenance is explicitly defined in the data input stages of the workflow, such as downloading files by FTP, accessing S3 buckets or creating EC2 volumes from snapshots of data archives. Agua also retains workflow provenance information such as the owner, workflow application run dates and content type.

Sharing

The group is central to Agua's data access model. Users can create groups and add other users to their groups. Group access permissions mirror the three progressive levels of Linux file permissions: the owner's access, group members' access and public access. Users can add a project to a group to permit the group users to access the project's files and workflow. For example, to access a project folder owned by another user, the user has to be added to the group to which the project belongs. Similarly, genomic views can be shared with other users by adding them to the group to which the project containing the views belongs.

Scalable Computing

Current NGS technologies require considerable computing resources for sequence processing and analysis. Aside from completing jobs in less time, using more nodes also reduces storage costs. However, using too many nodes during less computationally demanding steps in a workflow can result needless waste of resources. Agua integrates the StarCluster (http://star.mit.edu/cluster/) cluster computing toolkit, which is scalable to 10,000 nodes[25]. Once a cluster is started, StarCluster begins to auto-scale, adding new nodes when the number and average duration of jobs is high and removing nodes when demand slacks off. The user can preconfigure the minimum and maximum number of nodes for each StarCluster tailored to each workflow. The user can also change the minimum and maximum number of nodes while the workflow is in progress to best suit their budget and time constraints.

Genomic Viewer

Agua's genome viewer functionality incorporates the AJAX-enabled JBrowse genome viewer (http://jbrowse.org) to provide a fast, fluid and responsive genome browser interface with drag and drop feature track selection. Users can zoom and pan in real time from nucleotide to chromosome level views. JBrowse displays pre-loaded feature tracks alongside feature tracks generated from workflow outputs. Right-click context menus allow the user to display detailed feature information and open linked web pages.

Conclusion

The rate of genomic data generation far surpasses our capacity to analyse the data. Agua proposes a user-centric approach to facilitating workflow sharing, reproducibility and interpretation to accelerate genomic analysis. Agua aims to maximize reproducibility and transparency by enabling the sharing of workflow configurations using distributed version control and a proposed JSON-based workflow descriptor format that is human-friendly and readily usable in web and server applications. Integration with distributed version control systems enhances reproducibility and transparency and is particularly suited to a federated model of workflow standards.

Agua's modular, plugin-based architecture, provides a flexible, end-to-end cloud platform with rich features and easy extensibility. By enabling advanced management and interrogation of sequencing results by informatics non-experts, Agua lowers the technology barriers to using and understanding genomic data.

Although intended primarily as a bioinformatics platform, Agua's feature set and underlying technologies make it suitable for use as a workflow tool in any scientific or commercial domain that uses Linux command-line applications.

As the bioinformatics bottleneck shifts to the integration of clinical and phenotypic data with genomic data, reproducibility remains critical to the advancement of scientific knowledge. The open and user-centric design philosophy of Agua and its implemented features are a valuable alternative and complement to other workflow systems.

List of abbreviations

GUI: Graphical User Interface; JSON: Javascript Object Notation; NGS: next-generation sequencing.

Authors' contributions

SY and JG designed the approach, acquired the data, interpreted the results and wrote the manuscript.

SY implemented the Agua framework and maintains its website, repository and cloud presence.

Authors' information

Acknowledgements

References

1. FASTSTATS - Deaths and Mortality. at <http://www.cdc.gov/nchs/fastats/deaths.htm>
2. Galperin, M. Y. & Fernandez-Suarez, X. M. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.* **40,** D1–D8 (2012).
3. Bolchini, D., Finkelstein, A., Perrone, V. & Nagl, S. Better bioinformatics through usability analysis. *Bioinformatics* **25,** 406–412 (2009).
4. Kedes, L. & Campany, G. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE Competition. *Nat. Genet.* **43,** 1055–1058 (2011).
5. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **13,** 667–672 (2012).
6. Huang, Y. & Gottardo, R. Comparability and reproducibility of biomedical data. *Brief. Bioinform.* **14,** 391–401 (2013).
7. Baggerly, K. A. & Coombes, K. R. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *arXiv:1010.1092* (2010). doi:10.1214/09-AOAS291
8. Mesirov, J. P. Accessible Reproducible Research. *Science* **327,** 415–416 (2010).
9. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
10. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* **16,** 276–277 (2000).
11. EMBOSS. at <http://emboss.sourceforge.net/>
12. NCBITools (NCBI Software Collaboration). at <https://github.com/NCBITools/>
13. Lushbough, C., Bergman, M. K., Lawrence, C. J., Jennewein, D. & Brendel, V. BioExtract Server—An Integrated Workflow-Enabling System to Access and Analyze Heterogeneous, Distributed Biomolecular Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **7,** 12–24 (2010).
14. Berthold, M. R. *et al.* in *Data Anal. Mach. Learn. Appl.* (Preisach, C., Burkhardt, P. D. H., Schmidt-Thieme, P. D. L. & Decker, P. D. R.) 319–326 (Springer Berlin Heidelberg, 2008). at <http://link.springer.com/chapter/10.1007/978-3-540-78246-9_38>
15. Taylor, J., Schenck, I., Blankenberg, D. & Nekrutenko, A. in *Curr. Protoc. Bioinforma.* (John Wiley & Sons, Inc., 2002). at <http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1005s19/abstract>
16. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38,** 500–501 (2006).
17. Néron, B. *et al.* Mobyle: a new full web bioinformatics framework. *Bioinformatics* **25,** 3005–3011 (2009).
18. Deelman, E., Blythe, J., Gil, Y. & Kesselman, C. *Pegasus: Planning for Execution in Grids*. (2002).
19. Garret, J. Ajax: A New Approach to Web Applications. (2005). at <www.adaptivepath.com/publications/essays/archives/0 00385.php>
20. Kasemvilas, S. & Firpo, D. Effects of AJAX Technology on the Usability of Blogs. in *Proc. Symp. Hum. Interface 2009 Hum. Interface Manag. Inf. Inf. Interact. Part II Held Part HCI Int. 2009*

45–54 (Springer-Verlag, 2009). doi:10.1007/978-3-642-02559-4_6
21. Kluge, J., Kargl, F. & Weber, M. The Effects of the AJAX Technology on Web Application Usability. in *WEBIST 2007 Int. Conf. Web Inf. Syst. Technol.* 289–294 (2007).
22. Designing for the Scent of Information. The Essentials Every Designer Needs to Know About How Users Navigate Through Large Web Sites. at <http://www.uie.com/reports/scent_of_information/>
23. Pilgrim, C. J. An investigation of usability issues in AJAX based web sites. (2013). at <http://researchbank.swinburne.edu.au/vital/access/manager/Repository/swin:33109>
24. It's not about the data. *Nat. Genet.* **44,** 111–111 (2012).
25. Ho, R. Running a 10,000-node Grid Engine Cluster in Amazon EC2. at <http://blogs.scalablelogic.com/2012/11/running-10000-node-grid-engine-cluster.html>

Illustrations and figures (if any)

Figure 1. Workflow tab enables drag and drop of applications to create pipelines

Figure 2. Folder tab provides file manager view of remote filesystem

Figure 3. View tab displays genomic features with drag and drop add/remove of feature tracks

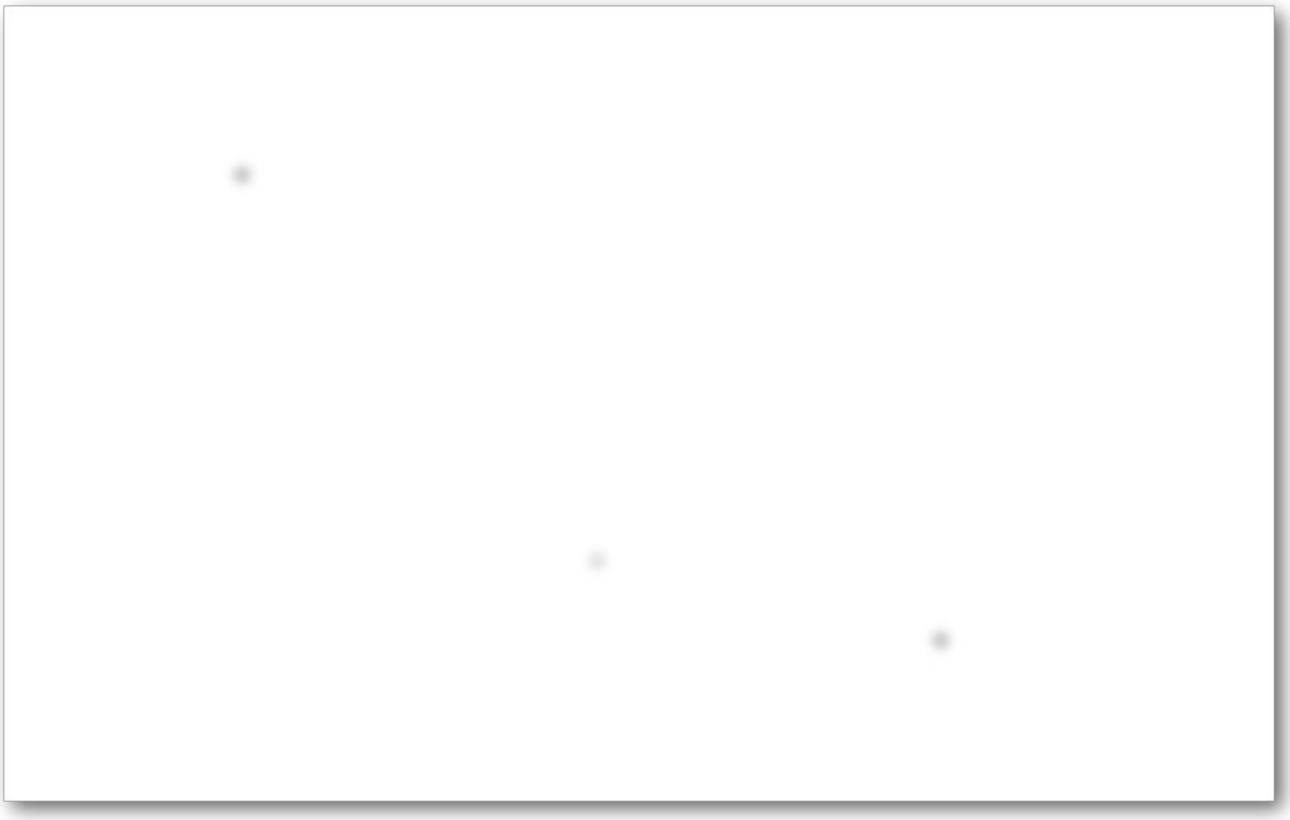

Figure 4. Admin tab enables quick addition of new applications and packages

Figure 5. Cloud tab provides easy management of cloud authentication and settings

Tables and captions (if any)

Table 1. Functionality coverage of genomic viewers and bioinformatics workflow tools