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 sciences disciplines where researchers are even less accustomed to high-throughput approaches. Agua [http://www.aguadev.org](http://www.aguadev.org/) lowers these technology barriers by providing a flexible, end-to-end cloud bioinformatics workflow solution, from data provenance and management to genomic visualization.

Keywords

NGS, cloud, bioinformatics, workflows, visualization, version control, JSON.

Article subdivisions

**Introduction**

Bioscience is increasingly 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[[1]](#footnote-2). These sources are mostly simple websites with limited data format compatibility between them even though 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[[2]](#footnote-3). As the $1,000 genome approaches[[3]](#footnote-4), genomic data volumes will increase and the bottleneck will shift to data integration, e.g., clinical with genomic, and responding to the question: How can we understand this data? Agua helps address this problem by providing a user-centric open-source platform that facilitates sharing of workflows and interpretation of results.

The availability of common tools to enable the comparison of results from different data sets[[4]](#footnote-5) paradoxically underscores the difficulty faced by scientists without programming experience. Software libraries such as Bioperl, Biopython and Bioconductor, and bioinformatics toolkits such as EMBOSS are valuable contributions in terms of content and approach. They do not provide a more generalized solution to allow scientists with little or no informatics skills to run well-defined workflows or generate novel workflows with new computational tools.

**Workflow Standards**

BioExtract[[5]](#footnote-6), KNIME[[6]](#footnote-7), Galaxy[[7]](#footnote-8), GenePattern[[8]](#footnote-9), GeneProf, Mobyle[[9]](#footnote-10), and Pegasus[[10]](#footnote-11) are among the most 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. However, only ten out of 299 papers published in 2011 that explicitly cite the project actually used the tools recommended by the consortium for mapping and variant discovery[[11]](#footnote-12). This suggests a more pressing need for establishing standards for reporting workflows than for establishing 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 scratch to finish while generating a series of snapshots of the workflow in the form of a JSON-format workflow file. These workflow creation steps are synchronized to a repository on Github, BitBucket or a similar Git repository service provider. This version history can be stored as desired to a private repository or published to a public repository which can be referenced in a scientific paper. This simplifies the process of sharing and reusing workflows, maximizes transparency and allows scientists to easily evaluate and reproduce each other's research.

As a further response to the tendency among scientists towards a heterogeneity of computational approaches, Agua provides a smooth, user-centric interface that enables scientists to easily add new applications and use these applications to create workflows and share their own custom workflows.

**The User-Centric Experience**

Previous workflow GUIs had drawbacks such as difficulty navigating and searching, having to navigate back and forth between forms and results, user difficulty keeping track of past actions, insufficient documentation of application options and ambiguities concerning data storage[[12]](#footnote-13). Agua is designed to minimize these drawbacks and maximize 'traction' in the user experience by providing a highly intuitive and transparent interface.

Agua's design focuses on empowering the domain-specific view of scientists rather than enforcing canonical workflows or optimizing job scheduling. This paradigm focuses on allowing the scientist easy access to import the data types, tools and distributed resources that are most suitable for their particular research domain.

List of abbreviations used (if any)

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

“AJAX Command Definition (ACD Files).” European Molecular Biology Laboratory (European Bioinformatics Institute). Accessed August 5, 2013. http://emboss.sourceforge.net/developers/acd/.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. “KNIME: The Konstanz Information Miner.” In *Data Analysis, Machine Learning and Applications*, edited by Christine Preisach, Professor Dr Hans Burkhardt, Professor Dr Lars Schmidt-Thieme, and Professor Dr Reinhold Decker, 319–326. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, 2008. http://link.springer.com/chapter/10.1007/978-3-540-78246-9\_38.

Bolchini, Davide, Anthony Finkelstein, Vito Perrone, and Sylvia Nagl. “Better Bioinformatics through Usability Analysis.” *Bioinformatics* 25, no. 3 (February 1, 2009): 406–412. doi:10.1093/bioinformatics/btn633.

Deelman, Ewa, James Blythe, Yolanda Gil, and Carl Kesselman. *Pegasus: Planning for Execution in Grids*, 2002.

“FASTSTATS - Deaths and Mortality.” Accessed August 5, 2013. http://www.cdc.gov/nchs/fastats/deaths.htm.

Galperin, Michael Y., and Xose M. Fernandez-Suarez. “The 2012 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection.” *Nucleic Acids Research* 40, no. D1 (January 2012): D1–D8. doi:10.1093/nar/gkr1196.

Goecks, Jeremy, Anton Nekrutenko, and James Taylor. “Galaxy: a Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences.” *Genome Biology* 11, no. 8 (2010): R86. doi:10.1186/gb-2010-11-8-r86.

Huang, Yunda, and Raphael Gottardo. “Comparability and Reproducibility of Biomedical Data.” *Briefings in Bioinformatics* 14, no. 4 (July 1, 2013): 391–401. doi:10.1093/bib/bbs078.

Kedes, Larry, and Grant Campany. “The New Date, New Format, New Goals and New Sponsor of the Archon Genomics X PRIZE Competition.” *Nature Genetics* 43, no. 11 (November 2011): 1055–1058. doi:10.1038/ng.988.

Lushbough, Carol, Michael K. Bergman, Carolyn J. Lawrence, Doug Jennewein, and Volker Brendel. “BioExtract Server&#x2014;An Integrated Workflow-Enabling System to Access and Analyze Heterogeneous, Distributed Biomolecular Data.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7, no. 1 (2010): 12–24. doi:10.1109/TCBB.2008.98.

Mesirov, Jill P. “Accessible Reproducible Research.” *Science* 327, no. 5964 (January 22, 2010): 415–416. doi:10.1126/science.1179653.

Nekrutenko, Anton, and James Taylor. “Next-generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility.” *Nature Reviews Genetics* 13, no. 9 (September 2012): 667–672. doi:10.1038/nrg3305.

Néron, Bertrand, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. “Mobyle: a New Full Web Bioinformatics Framework.” *Bioinformatics* 25, no. 22 (November 15, 2009): 3005–3011. doi:10.1093/bioinformatics/btp493.

Reich, Michael, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P. Mesirov. “GenePattern 2.0.” *Nature Genetics* 38, no. 5 (May 2006): 500–501. doi:10.1038/ng0506-500.

Taylor, James, Ian Schenck, Dan Blankenberg, and Anton Nekrutenko. “Using Galaxy to Perform Large-Scale Interactive Data Analyses.” In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002. http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1005s19/abstract.

Illustrations and figures (if any)

Tables and captions (if any)



Figure 1. Genomic viewer and bioinformatics workflow tools - functionality coverage

1. Galperin and Fernandez-Suarez, “The 2012 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection.” [↑](#footnote-ref-2)
2. Bolchini et al., “Better Bioinformatics through Usability Analysis.” [↑](#footnote-ref-3)
3. Kedes and Campany, “The New Date, New Format, New Goals and New Sponsor of the Archon Genomics X PRIZE Competition.” [↑](#footnote-ref-4)
4. Mesirov, “Accessible Reproducible Research”; Goecks, Nekrutenko, and Taylor, “Galaxy”; Huang and Gottardo, “Comparability and Reproducibility of Biomedical Data.” [↑](#footnote-ref-5)
5. Lushbough et al., “BioExtract Server&#x2014;An Integrated Workflow-Enabling System to Access and Analyze Heterogeneous, Distributed Biomolecular Data.” [↑](#footnote-ref-6)
6. Berthold et al., “KNIME.” [↑](#footnote-ref-7)
7. Taylor et al., “Using Galaxy to Perform Large-Scale Interactive Data Analyses.” [↑](#footnote-ref-8)
8. Reich et al., “GenePattern 2.0.” [↑](#footnote-ref-9)
9. Néron et al., “Mobyle.” [↑](#footnote-ref-10)
10. Deelman et al., *Pegasus*. [↑](#footnote-ref-11)
11. Nekrutenko and Taylor, “Next-generation Sequencing Data Interpretation.” [↑](#footnote-ref-12)
12. Néron et al., “Mobyle”; Bolchini et al., “Better Bioinformatics through Usability Analysis.” [↑](#footnote-ref-13)