BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 14 2010, pages 1779–1780
doi:10.1093/bioinformatics/btq282

Genome analysis

Advance Access publication May 28, 2010

geWorkbench: an open source platform for integrative genomics

Aris Floratos[1,*], Kenneth Smith[1], Zhou Ji[1], John Watkinson[2] and Andrea Califano[1]
[1]Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032 and [2]Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, NY 10027, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: geWorkbench (genomics Workbench) is an open source Java desktop application that provides access to an integrated suite of tools for the analysis and visualization of data from a wide range of genomics domains (gene expression, sequence, protein structure and systems biology). More than 70 distinct plug-in modules are currently available implementing both classical analyses (several variants of clustering, classification, homology detection, etc.) as well as state of the art algorithms for the reverse engineering of regulatory networks and for protein structure prediction, among many others. geWorkbench leverages standards-based middleware technologies to provide seamless access to remote data, annotation and computational servers, thus, enabling researchers with limited local resources to benefit from available public infrastructure.

Availability: The project site (http://www.geworkbench.org) includes links to self-extracting installers for most operating system (OS) platforms as well as instructions for building the application from scratch using the source code [which is freely available from the project's SVN (subversion) repository]. geWorkbench support is available through the end-user and developer forums of the caBIG® Molecular Analysis Tools Knowledge Center, https://cabig-kc.nci.nih.gov/Molecular/forums/

Contact: geworkbench@c2b2.columbia.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 20, 2010; revised on May 20, 2010; accepted on May 24, 2010

1 INTRODUCTION

A large number of bioinformatics techniques have been developed in recent years to serve the needs of biomedical research. The field is moving rapidly, with new and improved approaches appearing frequently. The fast pace of change and the technical sophistication of these approaches creates a barrier of adoption for ordinary biologists. The problem is exacerbated by the integrative nature of biomedical research, which requires combining data from multiple genomic/biomedical databases and using an array of advanced analyses, often available only in the form of command line programs (Kumar and Dudley, 2007, Reich *et al.*, 2006). Additionally, due to their sheer size and dimensionality, analysis of genomic data sets can be computationally very demanding. It is unlikely that every biomedical researcher that would like to utilize such analyses will have access to local/institutional hardware resources capable of supporting their execution. It is then important to facilitate sharing of public infrastructure through the use of technologies such as grid computing (Kurc *et al.*, 2009, Stevens *et al.*, 2003).

geWorkbench integrates many computational resources and makes them available through a unified user interface. For biomedical researchers with little or no computational training, this approach facilitates adoption by eliminating many steps that require programming skills (e.g. transformations from one file format to another; staging of databases and programs; dealing with operating system (OS) shells in order to execute command line programs). For software developers, geWorkbench provides an open source, component-based architecture that enables the addition of new functionality in the form of plug-in modules (which can further leverage existing tooling that streamlines access to server side components). Extensive documentation is available (manuals, online help and tutorials) to guide users in the proper use of the application. An innovative logging framework collects and mines data about how various modules are being utilized, offering the possibility for novice users to learn from their more advanced peers.

2 DESIGN, IMPLEMENTATION AND SUPPORT

geWorkbench comprises at present more than 70 distinct modules, supporting the integrated analysis and visualization of many types of genomic data. Users of geWorkbench can:

- Load data from disk or from remote data sources (such as the caArray microarray data repository at the National Cancer Institute, https://array.nci.nih.gov/; the PDB protein structure database; and sequence databases at the University of Santa Cruz and the European Molecular Biology Laboratory).
- Visualize gene expression, molecular interaction networks, protein sequence and protein structure data in a variety of ways.
- Access client- and server-side computational analysis tools such as *t*-test analysis, hierarchical clustering, self-organizing maps, analysis of variance (ANOVA), regulatory and signaling network reconstruction, basic local alignment search tool (BLAST) searches, pattern/motif discovery, etc.
- Validate computational hypothesis through the integration of gene and pathway annotation information from curated sources as well as through enrichment analyses.

Several modules have been developed in collaboration with investigators from the Center for the Multi-scale Analysis of Genomic and Cellular Networks (MAGNet, http://magnet.c2b2.columbia.edu/), one of seven National Centers for Biomedical Computing (NCBCs, http://www.ncbcs.org);

[*]To whom correspondence should be addressed.

Table 1. geWorkbench plug in modules based on MAGNet tools

Plugin Name	Description
ARACNe	Prediction of transcriptional interactions from gene expression data.
MINDy	Identification of modulators of transcriptional regulation.
Matrix REDUCE	Physics-based prediction of DNA-binding sites.
MEDUSA	Machine learning-based prediction of DNA-binding sites.
MarkUs	Functional annotation of protein structures.
Pudge	Protein structure prediction from sequence
SkyBase	Database of predicted protein structure models.

See also Supplementary Material.

The mission of the MAGNet Center is to provide the research community with novel, structural and systems biology methods and tools for the dissection of molecular interactions in the cell and for the interaction-based elucidation of cellular phenotypes (Table 1 lists some of the MAGNet plug ins; Supplementary Figs 1 and 2 provide related screenshots). Other geWorkbench modules are wrapped versions of pre-existing third party software tools such as Cytoscape (Shannon *et al.*, 2003), Ontologizer (Bauer *et al.*, 2008), and several analysis modules from the Microarray Experiment Viewer (Saeed *et al.*, 2003) and GenePattern (Reich *et al.*, 2006). A complete module listing (with screenshots) is available at the project web site, http://www.geworkbench.org/, under the 'Plug-ins' page.

2.1 Software architecture

geWorkbench is designed around a component-based architecture whose main purpose is to facilitate the expeditious integration of algorithms and data sources as plug in modules. Its key elements are described in the Supplementary Material.

2.2 Server side computing

The computational, administration and storage requirements of many bioinformatics resources developed by MAGNet (such as large genomic databases and CPU/memory-intensive algorithms) make deployment on a desktop computer infeasible. To make them available to geWorkbench (as well as other clients), such resources are deployed as programmatically accessible grid services using caGrid (Saltz *et al.*, 2006), the grid middleware layer of the caBIG® initiative. An attractive feature of caGrid is that if offers tooling (Introduce Toolkit; Hastings *et al.*, 2007), which streamlines the process of defining, deploying and registering caGrid-aware services. caGrid services have been developed by other participants of the caBIG® program as well and geWorkbench provides access to many of them (caArray, the Cancer Gene Index, the NCI Pathway Interaction Database, etc.).

2.3 Documentation and user support

geWorkbench is accompanied by detailed end-user documentation, demonstrating how to perform common analysis tasks and explaining the theoretical underpinnings of many analysis modules. Community support is provided by the caBIG® Molecular Analysis Tools Knowledge Center (MAT-KC, https://cabig-kc.nci.nih.gov/Molecular/KC). Additionally, geWorkbench utilizes novel data mining approaches (Murphy *et al.*, 2008) to create communities of practice through activity awareness. See Supplementary Material for details.

3 DISCUSSION

geWorkbench provides biomedical researchers with an integrated suite of well-documented analysis and visualization tools. By removing the need for programmatic manipulations, it facilitates utilization of multi-module analysis pipelines. It leverages state of the art middleware software to allow seamless access to remote data sources and computational infrastructure. It offers access to many advanced systems and structure biology methods developed by investigators at the MAGNet Center. Moreover, by adopting an open source development model and by using an extensible component-based architecture, it encourages community contributions and addition of new modules so that available functionality can keep up with the needs of the user base. Finally, the MAT-KC makes extensive resources available to assist users with learning how to use the application and its various plug in modules.

ACKNOWLEDGEMENTS

We thank the many people who have made significant contributions to geWorkbench, from past and present members of the core development team (Kiran Keshav, Min You, Thomas Garben, Oleg Shteynbuk, Udo Tobben, Meng Wang, Mark Chiang, Bernd Jagla, Matt Hall, Manjunath Kustagi, Xiaoxing Zhang) to our many scientific collaborators at the MAGNet Center (Barry Honig, Diana Murray, Harmen Bussemaker, Chris Wiggins, Christina Leslie, Gail Kaiser, Andrey Rzhetsky).

Funding: National Institutes of Health (U54CA121852); caBIG® Initiative of the National Cancer Institute (29XS192).

Conflict of Interest: none declared.

REFERENCES

Bauer,S. *et al*. (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.

Hastings,S. *et al*. (2007) Introduce: an open source toolkit for rapid development of strongly typed grid services. *J. Grid Comput.*, **5**, 407–427.

Kumar,S. and Dudley,J. (2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics*, **23**, 1713–1717.

Kurc,T. *et al*. (2009) HPC and grid computing for integrative biomedical research. *Int. J. High Perform. Comput. Appl.*, **23**, 252–264.

Murphy,C. *et al*. (2008) genSpace: exploring social networking metaphors for knowledge sharing and scientific collaborative work. In *23rd IEEE/ACM International Conference on Automated Software Engineering—Workshop Proceedings (ASE Workshops 2008)*. IEEE, L'Aquila, Italy.

Reich,M. *et al*. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.

Saeed,A.I. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Saltz,J. *et al*. (2006) caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics*, **22**, 1910–1916.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Stevens,R.D. *et al*. (2003) myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, **19** (Suppl. 1), i302–304.