

G-language Genome Analysis Environment Version 2: Integrated workbench for computational genome sequence analysis

Kazuharu Arakawa¹
gaou@sfc.keio.ac.jp

Masaru Tomita¹
mt@sfc.keio.ac.jp

¹ Institute for Advanced Biosciences, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan

Abstract

Although useful software libraries are developed and freely made available mainly by the collaborative efforts of Bio* projects, analysis environment for genome informatics that brings higher efficiency for the trial-and-error processes which comprise the majority of research routines is still less explored. Here we introduce the G-language Genome Analysis Environment (GAE) Version 2, a workbench based on Perl programming language and BioPerl that provide efficient interfaces for this purpose.

Keywords: bioinformatics, software, workbench, G-language GAE

1 Introduction

Since the term “bioinformatics” was first coined in the mid-1980s, the new discipline in molecular biology to utilize computers to analyze the biological data, as opposed to using laboratory equipments to analyze the living cells and organisms, rapidly became an invaluable approach in the research of life science. However, despite of its progress and the availability of numerous databases and software tools, the actual research procedures are nonetheless comprised of the perfunctory associations of such tools and inter-conversions of data and software output results. Although useful software libraries are developed and freely made available mainly by the collaborative efforts of Bio* projects [1], analysis environment for genome informatics that brings higher efficiency for the trial-and-error processes which comprise the majority of research routines is still less explored. In order to expedite these routine works for bioinformatics researches, we have been developing a software workbench designated the G-language Genome Analysis Environment (G-language GAE) since 2001 [2,3]. The software is based on Perl and is highly compatible with BioPerl. Here we present a major upgrade version of the software, that contains over 200 analysis programs and enhanced support for databases in several areas of molecular biology, encompassing genomics, transcriptomics, proteomics, and systems biology. Key features added in this release include remote sequence retrieval, distribution in a live-CD Linux requiring no installation, support for grid computing environment, large-scale caching with relational database, interactive shell, comprehensive documentation, and the unified interface development feature achieving a development cycle from a program to a GUI application.

2 Software Architecture

2.1 User Interface

G-language GAE is equipped with three interchangeable interfaces: Application Programming Interface (API) that is compatible with BioPerl [4], graphical user interface, and an interactive shell. The G-language Shell is newly introduced in version 2, which provides interactive command-line access to the entire G-language API, BioPerl, and Perl programming language, with basic shell functions such as line editing, tab-completion for filenames and function names, command history, EMACS key-bindings, data and

workspace persistence, online help, data searching and retrieval from public database, and use of UNIX commands.

2.2 Analysis Programs

This system provides over 200 analysis methods for genome informatics and systems biology, each implementing a specific tool or algorithm. Areas especially focused in genome informatics analysis includes methods for the identification of sequences with significant information content using information theory, observation of nucleotide composition and genomic compositional asymmetry, calculation of codon bias measures and prediction of gene expression levels, and statistical analysis of short oligomers such as short tandem repeats and palindromes [5]. These methods can easily be combined with several other applications and algorithms to produce a workflow in genome informatics research for studying specific biological questions.

2.2 Data OR-mapper and Database Interface.

In order to allow rapid access and manipulation of relational database, v.2 software includes an OR-mapper that ties multi-dimensional Perl data of any depth to any relational database. This function quickly achieves data persistence, efficient data access, and the creation of database interface for remote online data.

3 Conclusions

G-language GAE v.2 provides highly efficient interface for genome informatics analyses, equipped with more than 200 analysis programs and interactive shell. G-language GAE v.2 software and documentations, including API references, interactive help viewer, manuals, and tutorials, are freely available under GNU General Public License at our website [6]. Installation package is available as standard Perl module, and as a complete installer for MacOS X. Moreover, G-language GAE is also distributed in the CD/DVD-bootable Linux for bioinformatics, Knoppix for Bio (KNOB), which allows installation-free testing simply by restarting a Windows/Linux/Macintosh computer with the CD/DVD.

References

- [1] Mangalam, H., The Bio* toolkits--a brief overview, *Briefings in Bioinformatics*, 3: 296-302, 2002.
- [2] Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y., and Tomita, M., G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining, *Bioinformatics*, 19(2): 305-306, 2003.
- [3] Arakawa, K., and Tomita, M., G-language System as a platform for large-scale analysis of high-throughput omics data, *Journal of Pesticide Science*, 31(3):282-288, 2006.
- [4] Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E., The Bioperl Toolkit: Perl modules for the life sciences, *Genome Research*, 12(10):1611-1618, 2002.
- [5] Arakawa, K., Suzuki, H., and Tomita M., Computational Genome Analysis Using The G-language System", *Genes Genomes and Genomics*, 2(1):1-13, 2008.
- [6] <http://www.g-language.org/>