Development of Bacteria Analysis System v.2 on the G-Language Genome Analysis Environment

Daisuke Kyuma[1,2] q@g-language.org
Kazuharu Arakawa[1,3] gaou@g-language.org
Koya Mori[1,3] mory@sfc.keio.ac.jp
Masaru Tomita[1,2] mt@sfc.keio.ac.jp

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0017, Japan
[2] Faculty of Environmental Information
[3] Bioinformatics Program, Graduate School of Media and Governance

Keywords: G-language Genome Analysis Environment, analysis software, computational analysis

1 Introduction

We have been developing a generic analysis environment called the G-language Genome Analysis Environment (G-language GAE) [1] to construct an integrated environment for the development of analysis software, to systematically accumulate existing analysis software and their results, and to construct generic analysis packages that allow users to avoid redundancy in the process of analysis. We are distributing the software system at our web site, `http://www.g-language.org/`.

G-language GAE has over 200 methods installed; however, using them requires perl programming skills. Therefore we developed Bacteria Analysis System (BAS) in order to analyze easily via the Graphical User Interface. On G-language GAE, set of functions forming a protocol of analyses is called a "System". While BAS is one of such "Systems", version 2 of the BAS is made more generic in terms of analysis purposes, and is greatly enhanced in its extensibility and interconnectivity with other "Systems", in order to enable more complex analysis required in the post-genome era.

2 System Architecture

Presently several systems regarding the analysis of bacterial genomes exist upon the G-language GAE, such as Comparative Genome Analysis System (COMGA), Experiment System (XP), Chi Sequence Analysis System, Comparative Analysis System for Overlapping gene(CAOS), and Readthrough Candidate Extraction System(RCES). To extend the BAS functionalities, the above independent analysis systems are thus integrated in version 2, interconnecting the different analysis engines that the specific systems are based on. Along with the addition of new methods of the G-language GAE, BAS version 2 can access the external systems as well as the methods within the G-language GAE, utilizing the function daemon at the core calculation engine.

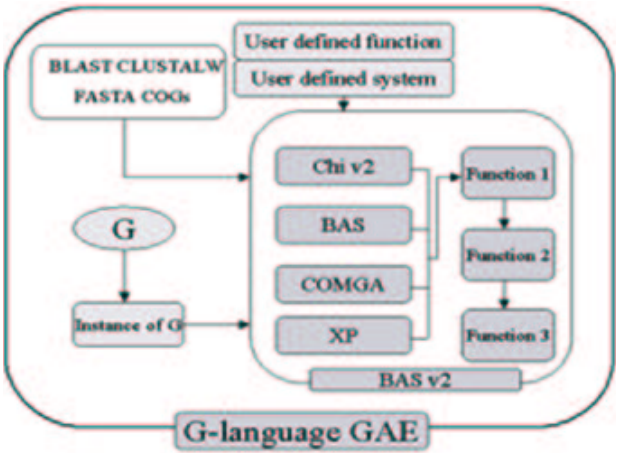

Figure 1: BAS v2 architecture.

The system core calculation engine can treat G-language GAE Configuration File (GCF) as a library and a configuration file, and specialized methods maintained in the GCF file operates with the native analysis methods of the G-language GAE. If users need external tools such as BLAST, FASTA, ClustalW [2] and database such as COGs [3], BAS v.2 can be scripted on the fly to include access through the dynamic loading structure.

3 Methods

G-language GAE includes many methods, but the text outputs from the methods are usually difficult to understand intuitively for users inexperienced in programming. Taking advantage of the new interface layer developed upon the G-language GAE v.2 which is based on HTML, BAS v.2 focuses on outputs to be dynamic and interactive. Powered by hyperlink of the Internet to connect unlimited resources, applications with interface developed using Flash and SVG multimedia content authoring tools are interactive, realizing datamining directly from the results obtained. New analysis functions incorporated in BAS v.2 as shown in Table 1 and output examples, graphical w and comga codon usage (Figure 2).

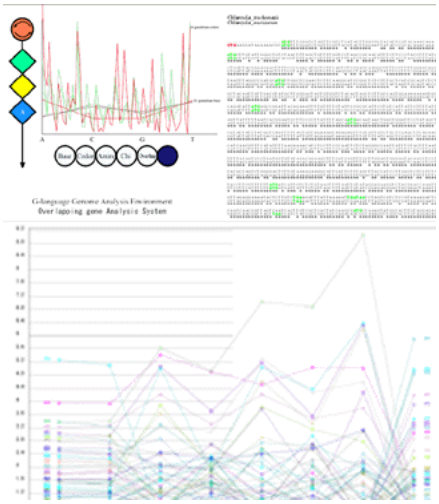

Figure 2: Output examples.

Table 1: Methods implemented in BAS.

Method	Function
codon table v2	Counts the number of codons and output tables using SVG
comga cogs grapher	Dispose orthologous genes and output images using SVG
comga correlation	Calculates correlation and output tables using SVG
genome map v2	Output genome map using SVG
graphical w	Output aln file using SVG
h gene seeker	Seek homologous genes
plasmid map	Output plasmid map using SVG

References

[1] 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*, 19:305–306, 2003.

[2] Tatusov, R.L., The COG database : new developments in phyogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 29(1):22–28, 2001.

[3] Thompson, J.D., Higgins, D.G., and Gibson, T.J., CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22:4673–4680, 1994.