

Note

WeGAS: A Web-Based Microbial Genome Annotation System

Daesang LEE,^{1,2,3} Hwajung SEO,² Chankyu PARK,^{1,†} and Kiejung PARK^{2,†}

¹Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Guseong-dong 373-1, Yuseong-gu, Daejeon 305-701, Republic of Korea

²Information Technology Institute, SmallSoft Co., Ltd., Jangdong 59-5, Yuseong-gu, Daejeon 305-343, Republic of Korea

³Department of Bioinformatics, Korea Bio Polytechnic, Nonsan, Chungnam 320-905, Republic of Korea

Received August 13, 2008; Accepted September 24, 2008; Online Publication, January 7, 2009

[doi:10.1271/bbb.80567]

We have developed WeGAS, a Web based microbial Genome Annotation System, which provides features that include gene prediction, homology search, promoter/motif analysis, genome browsing, gene ontology analysis based on the COGs and GO, and metabolic pathway analysis with web-based interfaces. Most raw data and intermediate data from genome projects can be managed with the WeGAS database system, and analysis results, including information on each gene and final genome maps, are provided by its visualization modules. Especially, a pie-view browser displaying circular maps of contigs and a COG-GO combination browser are very helpful for an overview of projects. Major public microbial genome databases can be imported, searched, and browsed through the WeGAS modules. WeGAS is freely accessible via web site <http://ns.smallsoft.co.kr:8051>.

Key words: annotation; computational analysis; gene ontology; genome; homology

Genome projects have produced tremendous amounts of biological sequence data, making annotation systems essential tools to determine the value of each sequence. These annotation systems enable molecular biologists to deal with genome data easily, as well as to access, edit, and update all related information at all steps of a genome project.

Several genome annotation systems for microbial genome projects have been developed and reported during the last decade. The first generation of genome annotation systems consisted of GAMBLER,¹⁾ MAGPIE,²⁾ and Pedant³⁾ systems. GAMBLER was a semi-automated genome analysis system to support the *Bacillus halodurans* genome project. MAGPIE and Pedant systems were mainly focused on making tables and preliminary graphic interfaces. ERGO, Pedant-Pro, Phylosopher, BioScout, and GenDB have common characteristics in extensive visualization functions and are classified as the second generation of genome annotation systems.⁴⁾ Recently, MaGe has reported that it offers a set of graphical interfaces that show synteny information of microbial genomes.⁵⁾

Although these reported microbial annotation systems have merits in their specific fields, they appear to give little information on ongoing microbial genome projects. Molecular biologists want to forecast the progress of

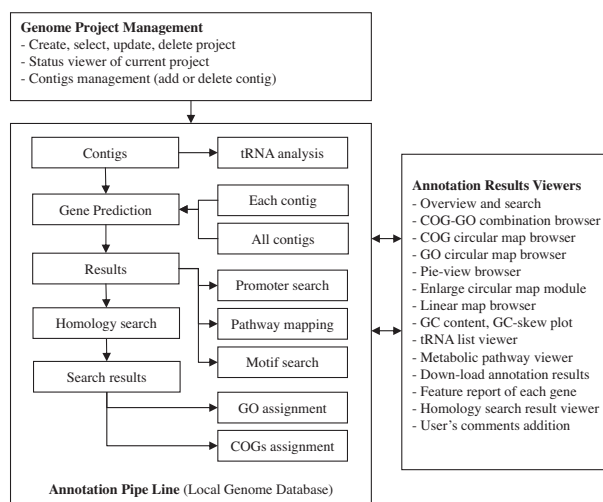


Fig. 1. An Overview of Workflow and Conceptual Scheme in the WeGAS.

The WeGAS consists of three major parts: Genome Project Management, Annotation Pipe Line, and Annotation Results Viewers. In Genome project management, users can start genome annotation by uploading their contigs and project information, such as project name, project description, etc. The main annotation processes are carried out in an annotation pipeline from gene prediction to the ontology analysis step by step. Final annotation results are accessible through various annotation results viewers. The feature report viewer for each gene provides a menu to add users' comments or notes to each gene. The local genome database stores and manages all the analysis results and data generated in the process of a microbial genome project.

genome work by exploring the circular maps with contigs generated in ongoing genome projects as well as finished genome sequences.

To address these needs, we developed WeGAS, a web-based microbial genome annotation system that provides web-based analysis interfaces from contigs uploading to genome annotation for ongoing as well as finished microbial genome projects. Among the features included in WeGAS are gene prediction, homology search, promoter/motif analysis, metabolic pathway analysis, and gene ontology analysis, with visualization modules that include genome browsers. The annotated information can be retrieved by WeGAS database searching and browsed with genome map browsers and a gene classification browser.

[†] To whom correspondence should be addressed. Chankyu PARK, Tel: +82-42-350-2629; Fax: +82-42-350-2610; E-mail: ckpark@kaist.ac.kr; Kiejung PARK, Tel: +82-17-535-5242; Fax: +82-42-385-9240; E-mail: kjpark@smallsoft.co.kr

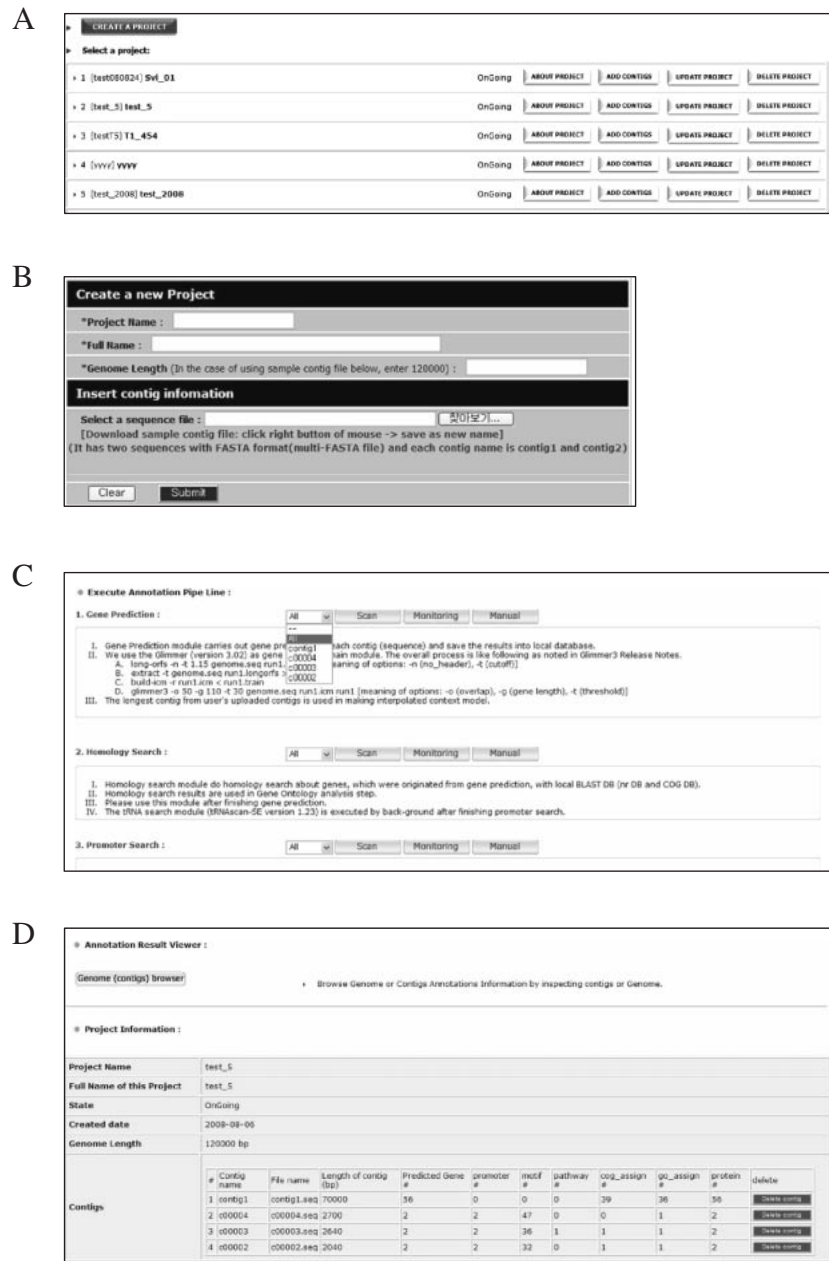


Fig. 2. Screen Shots of Project Management and Annotation Pipeline in the WeGAS.

A, The main menu, which shows a list of microbial genome projects. Users can create a new genome project by clicking the creation button or manage ongoing genome projects by selecting corresponding project. B, To start a new genome project, users upload contig sequences with project name, full description of project, and approximate genome size. For users' convenience, WeGAS automatically parses the multi-FastA format file into each FastA format. C, The annotation pipeline consists of seven major modules: gene prediction, homology search, promoter search, pathway mapping, motif search, COG assignment, and GO assignment. Through a monitoring button, users can check the processes of each pipeline as to whether an analysis module is finished or not. Users can use the annotation pipeline by selecting all contigs or a specific contig. D, Project information shows a brief summary of the annotation number of each contig, such as predicted gene, promoter, motif, pathway, and ontology analysis. The final, detailed annotation results are accessible by clicking the Genome browser button.

The WeGAS database contains data ranging from contigs to functional analysis of the final gene set for each genome project. The interface of each analysis tool is implemented not only to run each tool and to view the results but also to monitor progress. Analysis results are saved in the WeGAS local database system. The conceptual scheme and data management flows in the WeGAS are shown in Fig. 1, and Fig. 2 shows several screen shots of project management and annotation pipeline in the WeGAS.

The gene prediction tool used for WeGAS is Glimmer version 3.02. The longest contig from user's contigs is used to make interpolated context models.⁶⁾ WeGAS

uses PROSITE patterns (release version 20.0) for protein motif analysis.⁷⁾ Other forms of protein motif analysis, such as pFAM and SMART search, can be done by modifying a few interfaces. The progress/status of promoter and motif analyses can be monitored through appropriate monitoring modules. The major NCBI DNA databases (nt, est, sts, gss, and htg) are used for homology analysis of predicted genes using BLAST.⁸⁾ Both Clusters of Orthologous Groups of proteins (COGs)⁹⁾ and Gene Ontology databases¹⁰⁾ are used in assigning all translated proteins to their ontology categories. In the case of COG assignment, WeGAS does homology search annotated proteins with the COGs

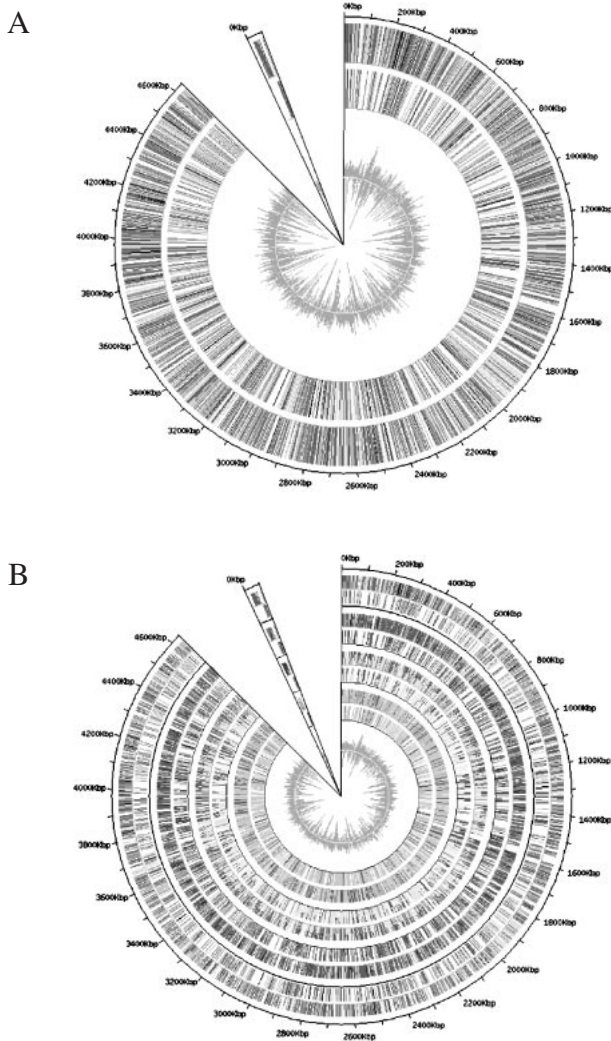


Fig. 3. Screen Shots of Annotation Results Viewer in the WeGAS.

A, COG pie-view browser. The first (plus strand) and second (minus strand) circles show gene distribution according to the functional category of the COG database. The third circle shows G + C content (higher values outward). B, COG-GO combination browser. The first and second circles are the same as those in A. Genes, assigned by the categories of molecular function, biological process, and cellular component in the GO database, are represented in circles 3 and 4, 5 and 6, and 7 and 8 respectively, each even-numbered circle representing the reverse strand and each odd numbered circle representing the forward strand. The ninth circle shows the G + C content of the contig sequences. The circular map viewer has a menu to enlarge the figure by 2 for users' convenience.

blast database (myva file from NCBI). Based on homology search results, WeGAS assigns proteins to the corresponding COG category with a cut-off value (e-value, 10^{-3}). For Gene Ontology analysis, WeGAS uses a gene-annotation dataset provided by CompuGen flat file (Release 0.5.1). From the homology search results of annotated proteins with nr DB, WeGAS filters the homology search results with a cut-off value (e-value, 10^{-5}), and then assigns the GO categories of the protein with the highest homology to the function of each protein.

A database searching module is implemented to query the annotation results of an ongoing or finished genome project. A linear map browser is used to visualize a whole genome map and to show detailed annotation information for each selected gene by further clicking.

A gene classification browser shows gene ontology analysis results with COG and GO for a whole genome.

A circular map is generated after retrieving gene ontology information for all genes in a genome and calculating the GC contents throughout the area of an entire genome. The circular map browser is designed to display a circular map for a genome. The map is displayed as a pie-view when there are more than two contigs for an ongoing genome project. The circular map and the linear map are hyperlinked to each other.

Ontology analysis results based on COG and GO can be displayed simultaneously (Fig. 3). Through these browsers, users can predict the final output of their genome projects. Several options and features were implemented to investigate selectively a specified ontology group category and region and to choose a drawing mode.

The tRNAscan-SE was used for tRNA prediction, and the results are shown in a tRNA analysis menu in WeGAS.¹¹⁾ A GC-skew plot for a microbial genome was integrated with a GC content viewer in our system.¹²⁾ Metabolic pathway analysis is performed using the KEGG sequence databases. After a homology search against the database, a pathway group with the best hit for each gene is assigned to that gene as the related pathway.¹³⁾ Our system also has parsing programs to import public microbial genome data, and result files are stored in local databases and accessed by browsers in WeGAS.

While WeGAS utilizes features similar to those of other annotation systems, such as Pedant, ERGO, and GenDB, we developed WeGAS to focus on stepwise and intuitive web interfaces familiar to biologists. A few intuitive and distinctive features are included, such as a COG-GO combination browser that shows the ontology analysis results simultaneously, and a pie-view browser, which forecasts the final blueprint for an ongoing genome project. WeGAS has been tested through practical genome projects: *Vibrio vulnificus* CMCP6,¹⁴⁾ *Streptomyces peucetius* ATCC 2795,¹⁵⁾ and *Thermotoga neapolitana* DSM 4359 (GenBank accession no. CP000916), including many practical industrial microbial genome projects, and was improved through intensive discussion with genome biologists. Additional features will be added and integrated into WeGAS, such as a genome alignment program, comparative genome analysis modules, chip data analysis, and protein-protein interaction analysis.

Acknowledgment

The authors would like especially to thank the bio-industrial companies GenoTech Corp., CJ Corp., Dae-sang Corp., and SolGent Co., Ltd. for helpful advice and suggestions. This research was financially supported by Ministry of Education, Science, and Technology and Korea Industrial Technology Foundation through the Human Resource Training Project, and by Ministry of Knowledge Economy grants from the Intelligence Bioinformatics and Application Center at the Korea Research Institute of Bioscience and Biotechnology.

References

- 1) Sakiyama, T., Takami, H., Ogasawara, N., Kuhara, S., Kozuki, T., Doga, K., Ohshima, A., and Horikoshi, K., An automated system for genome analysis to support microbial whole-genome shotgun sequencing. *Biosci. Biotechnol. Biochem.*, **64**, 670-673 (2000).
- 2) Gaasterland, T., and Sensen, C. W., MAGPIE: automated

- genome interpretation. *Trends Genet.*, **12**, 76–78 (1996).
- 3) Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., and Mewes, H. W., Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57 (2001).
 - 4) Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., and Puhler, A., GenDB: an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195 (2003).
 - 5) Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Medigue, C., MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65 (2006).
 - 6) Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L., Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679 (2007).
 - 7) Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A., The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238 (2002).
 - 8) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402 (1997).
 - 9) Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A., The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41 (2003).
 - 10) Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R., The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–261 (2004).
 - 11) Lowe, T. M., and Eddy, S. R., tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964 (1997).
 - 12) Lobry, J. R., Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665 (1996).
 - 13) Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34 (1999).
 - 14) Kim, Y. R., Lee, S. E., Kim, C. M., Kim, S. Y., Shin, E. K., Shin, D. H., Chung, S. S., Choy, H. E., Progulsk-Fox, A., Hillman, J. D., Handfield, M., and Rhee, J. H., Characterization and pathogenic significance of *Vibrio vulnificus* antigens preferentially expressed in septicemic patients. *Infect. Immun.*, **71**, 5461–5471 (2003).
 - 15) Parajuli, N., Basnet, D. B., Chan Lee, H., Sohng, J. K., and Liou, K., Genome analyses of *Streptomyces peucetius* ATCC 27952 for the identification and comparison of cytochrome P450 complement with other *Streptomyces*. *Arch. Biochem. Biophys.*, **425**, 233–241 (2004).