

An Integrated System for Microbial Genome Annotation and Comparative Genomic Analysis

Pan-Gyu Kim^{1,2}, Hwajung Seo¹, Daesang Lee¹, Hongseok Tae¹, Kiejung Park¹

Keywords: annotation, genome, ontology, COG, viewer, browser, comparative genomics

1 Introduction.

As genome projects have produced tremendous biological sequence data, annotation is considered as an essential part of genome sequencing projects to elucidate the value of the sequence [1]. Through annotation systems, molecular biologists could deal with genome data easily and all related information could be accessed, edited, or updated without additional efforts. We have developed an integrated microbial genome annotation system for microbial genome annotation and comparative genomic analysis, which provides web-based analysis interface for gene prediction, homology search, promoter analysis, motif analysis, genome browser, gene ontology analysis, and genome comparison.

The annotated information can be retrieved with database searching and browsed with a genome map browser and a gene classification viewer, where we have added various visualization features and functions. Public microbial genome databases are imported and can be searched and browsed through the same interface.

2 Methods

The database on our annotation system contains from contig data to functional analysis data of the final gene set. The interface of each annotation tool was implemented not only for running each tool and viewing the result but also for monitoring the progress. Analysis results are saved in the database with primary keys indicating the relationship between data .

As a first step in genome analysis, general methods for gene prediction are applied in this system and a few analysis options are provided. For promoter analysis, we implemented a general promoter pattern search and a two-component analysis search against all the predicted genes. For motif analysis, the Prosite DB patterns are searched against all protein sequences which are translated from the gene prediction. Fast algorithms were developed to accomplish fast searching for motif patterns of regular expression. The progress/status of promoter and motif analysis can be monitored through the web interface. For homology analysis of all the predicted genes, we implemented the interface for NCBI BLAST and both COGs(Clusters of Orthologous Groups) and GO(Gene Ontology) databases were used to classify the homology search result . Other features such as a GC-skew plot, tRNA analysis are implemented. Figure 1 shows the miscellaneous features of our system.

A database searching module was implemented to query for the annotation results of an in-progress or finished genome project and a linear map browser was implemented to visualize the whole genome map and detailed annotation information for each selected gene by further clicking. A gene classification viewer was implemented to show gene ontology analysis result with COG, GO, and COG/GO for a whole genome. A circular map is generated after retrieving gene ontology information of all the genes of a genome and calculating a few features for the whole genome area. A few options and features were implemented to select a specified category, a region and a drawing mode.

For public microbial genome data, input programs were implemented to parse the genome data of GenBank format and import into local databases. Each imported genome can be searched and browsed as an annotated genome can be.

Comparative genomic information can be shown as a genome alignment, that is, genomic DNA sequence comparison, using our genome comparison and visualization program, a Windows program connected to our main web interface via Active-X. Coding sequence comparison by genome scale is a kind of a comparative genomic analysis to show common genes and different genes between two organisms. More detailed alignment information between a pair of genes from two organisms respectively are shown on an alignment window which is generated from a pairwise alignment after pre-selection by BLAST. Our system also supports the result of comparison between COGs, promoters and protein motifs from different genomes as a comparative genome analysis method. This feature is very useful to research the differences of wild type and mutant because we can easily find out the insertion, deletion, substitution of mutant in comparison with wild type.

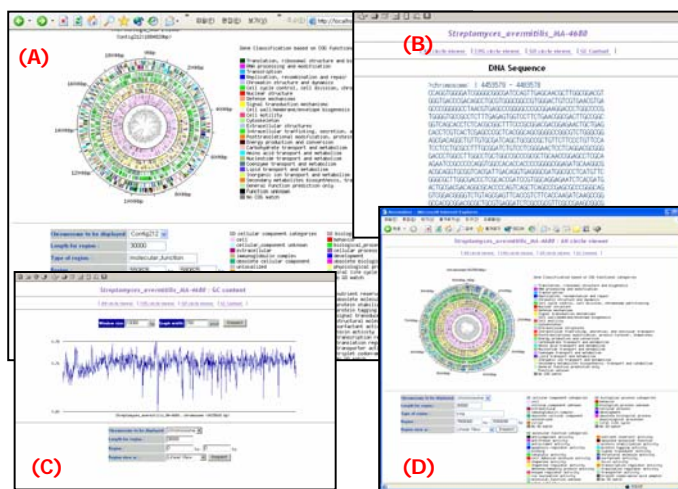


Figure 1: miscellaneous features: an overview & search viewer(A), a COG/GO viewer(B), a GC content viewer(C) and a linear map viewer(D).

3 Results and Discussion.

We have developed an integrated microbial genome annotation system, which provides web-based interfaces for gene prediction, homology search, promoter analysis, motif analysis, gene ontology analysis and genome comparison. A database searching module, linear map browsers and gene classification viewers based on both COG and GO were implemented. We have also developed a few comparative genomic analysis programs. The system has been used as a practical tool for a few microbial genome projects and has been evolved through additional requirement analysis against genome researchers.

We will add more features in the near future. These include metabolic pathway analysis and PKS (polyketide synthase) analysis system. This system will be very helpful not only for analysis of public microbial genome annotation but also as a practical analysis system of genomics/comparative genomics and metabolomics in practical microbial projects of genome scale for both academic and industrial purpose.

References

[1] Seo, H., Tae, H., Nam, H., and Park, K. 2003. Development of a Web-based Genome Annotation System. *International Conference on Research in Computational Molecular Biology (RECOMB 2003)*, Berlin, Germany.