Poster A-14

A rapid genome re-annotation system for comparative study of bacterial genomes

Authors:
Satoshi Tamaki (*Institute for Advanced Biosciences, Keio University, Kanagawa 252-8520, Japan*)
Kazuharu Arakawa (*Japan Society for the Promotion of Science*)
Masaru Tomita (*Institute for Advanced Biosciences, Keio University, Kanagawa 252-8520, Japan*)

Short Abstract: Comparative genomics requires genome data annotated with uniform criteria and methods including computer friendly semantics, and a rapid and automatic means for this purpose. We have developed a very fast genome re-annotation system for the comparative study of bacterial genomes, available under GNU GPL at: http://www.g-language.org.

Long Abstract:
A rapid genome re-annotation system for comparative study of bacterial genomes

Satoshi Tamaki1,2, Kazuharu Arakawa1,3, Masaru Tomita1,2

1 Institute for Advanced Biosciences, Keio University, Kanagawa 252-8520, Japan
2 Department of Environmental Information
3 Japan Society for the Promotion of Science

With the advent of genome sequencing techniques, the number of complete genomes is growing at an increasingly rapid rate. Genomes OnLine Database (GOLD: http://www.genomesonline.org/) currently lists more than 2000 published and on-going genome projects. Having this wealth of information at hand, comprehensive comparative study of bacterial genomes is now possible. However, in order to conduct comparative genome informatics analyses with multiple genomes, these genome data are desirable to be annotated with computer friendly semantics under a uniform method and criteria. Currently the complete genomes published in GenBank/EMBL/DDBJ repositories are diverse in terms of annotation completeness and in annotation strategies and methods, and some genomes at very early stages of annotation have very little, or sometimes no functional information of the genes. Genome annotation has been one of the most focused and successful area of bioinformatics, and many powerful software systems already exist for this purpose, achieving high efficiency and accuracy. On the other hand, most tools are aimed for genome projects and thus are semi-automatic and premised to have final expert curation, equipped with rich interface for this purpose. While this is still powerful and necessary for genome projects, fully automatic and rapid means (therefore inevitably less accurate but covers a majority of information) of re-annotation making use of the wealth of sequence databases under a uniform standard is desirable for post-genome comparative study of hundreds of genomes.

In this work, we introduce an open source automatic genome re-annotation system

developed upon the generic bioinformatics workbench G-language Genome Analysis Environment (G-language GAE) under GNU General Public License. Based on the available complete genome flatfile, this system re-annotates all genes using similarity searches of the amino acid sequences against UniProt/Swiss-Prot, UniProt/TrEMBL, and NCBI NR databases in the order of priority, with computer friendly terms including the Gene Ontology (GO) terms. Credibility of the similarity searches are marked with four levels: level 1 are those scoring with E-value less than 1e-70 and with more than 98 percent identity, level 2 are those scoring with E-value less than 1e-50 and with more than 95 percent identity, level 3 are those scoring with E-value less than 1e-30 and with more than 90 percent identity, and level 4 are those scoring with E-value less than 1e-10 and with more than 80 percent identity. In this way, the users can select the genes to use for the comparative study according to the annotation credibility, removing erroneous annotations which is highly likely for a certain percentage of genes with this kind of automatic method. BLAST Like Alignment Tool (BLAT) is used for similarity search to achieve a high speed and accurate performance, finishing the entire annotation in several minutes for one genome. The genomes can be further annotated for orthologous clusters with NCBI COGs, protein domains using HMMPfam, and localization with PsortDB. The system uses the predicted coding regions annotated in the original genome since gene identification in bacteria is well-established, but the users may optionally choose to use Glimmer for gene identification. In all annotation steps, top five hits are recorded, so that the information may be used for further manual curation. All hits to the employed databases are recorded with database IDs as well as information in text for humans. The resulting re-annotated genome can be generated in common database formats supported by G-language GAE or BioPerl including GenBank, EMBL, and GFF. The system took only 209 seconds to annotate the whole genome of Escherichia coli and covered 100% of the 4239 genes known in the EMBL database. It achieves high accuracy for well-characterized bacterial genomes such as in this example with E.coli, and with generally more than 70% coverage for all genomes. The software is available at http://www.g-language.org/ under GNU General Public License.

Arakawa K, Mori K, Ikeda K, Matsuzaki T, Kobayashi Y, Tomita M, "G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining", Bioinformatics, 2003, 19(2):305-306

Arakawa K, Nakayama Y, Tomita M, "GPAC: Benchmarking the sensitivity of genome informatics analysis to genome annotation completeness", In Silico Biology, 2006, 6:0006