Genome analysis

PLATCOM: a Platform for Computational Comparative Genomics

Kwangmin Choi1, Yu Ma2, Jeong-Hyeon Choi1 and Sun Kim1,3,*

1School of Informatics, 2Department of Computer Science and 3Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47404, USA

Received on October 30 2004; revised on January 28, 2005; accepted on February 20, 2005

ABSTRACT

Motivation: As more whole genome sequences become available, comparing multiple genomes at the sequence level can provide insight into new biological discovery. However, there are significant challenges for genome comparison. The challenge includes requirement for computational resources owing to the large volume of genome data. More importantly, since the choice of genomes to be compared is entirely subjective, there are too many choices for genome comparison. For these reasons, there is pressing need for bioinformatics systems for comparing multiple genomes where users can choose genomes to be compared freely.

Results: PLATCOM (Platform for Computational Comparative Genomics) is an integrated system for the comparative analysis of multiple genomes. The system is built on several public databases and a suite of genome analysis applications are provided as exemplary genome data mining tools over these internal databases. Researchers are able to visually investigate genomic sequence similarities, conserved gene neighborhoods, conserved metabolic pathways and putative gene fusion events among a set of selected multiple genomes.

Availability: http://platcom.informatics.indiana.edu/platcom

Contact: sunkim2@indiana.edu; kwchoi@indiana.edu

INTRODUCTION

PLATCOM (Platform for Computational Comparative Genomics) is a computational environment where users can choose any combination of genomes from 312 replicons freely and compare them with a suite of computational tools. Our system is designed to evolve through three development stages. As of October 2004, the first stage has been completed and we have begun a public service to the community through its web interface, which is presented in this paper. It is designed in a modular way, so that the tools and databases can be freely integrated and biologists can easily design their own experimental protocol for comparative genome analysis. PLATCOM focuses rather on data mining for high-performance scalable systems, compared with similar genome annotation systems, such as euGenes (Gilbert, 2002), BioWorks (http://amdec-bioinfo.cu-genome.org/html/BioWorks.htm), SEALS (Walker and Koonin, 1997) and DAS (http://biodas.org/). Five component tools are functionally connected with other component tools as well as command-line tools (Fig. 1). Biologists can perform various comparative genomic analyses, such as, finding (1) conserved gene order, (2) conserved gene neighborhoods, (3) conserved metabolic pathways and (4) putative gene fusion events among a set of multiple genomes.

INTERNAL DATABASES

PLATCOM is built on internal databases, which consist of GenBank (ftp://ftp.ncbi.nih.gov/genomes), Swiss-Prot (http://www.ncbi.nlm.nih.gov/UniProt), COG (http://www.ncbi.nlm.nih.gov/COG), KEGG (http://www.genome.ad.jp/kegg) and Pairwise Comparison Database (PCDB). PCDB is designed to incorporate newer genomes automatically, so that PLATCOM can evolve as new genomes become available. FASTA and BLASTZ are used to compute all pairwise comparisons (97,034 entries) of protein sequence files (.faa) and whole-genome sequence files (.fna) of 312 replicons. Multiple genome comparisons usually take too much time to complete, but the precomputed PCDB makes it possible to complete genome analysis very fast even on the Web. In general, our system runs several hundred times faster than a system without PCDB when comparing several genomes.

GENOME ANALYSIS APPLICATIONS

Five sequence analysis tools are embedded in the system and each component tool is designed to be interconnected, using command-line tools, with each other and internal databases. A set of genomes selected by users is submitted with parameter settings via web interface.

2D-Plotting. GenomePlot is a visualization tool to generate a genome comparison diagonal plot between two selected genomes. It retrieves pairwise comparison data from pre-computed PCDB to generate two-dimensional (2D) plot and its image map. GenomePlot provides a strong intuition to understand the overall genome structure and phylogenetic distance between two given genomes. It is also an effective way to visually identify gene clusters that are conserved between two close genomes.

Operon analysis. OperonViz is a tool to generate graphical visualization of gene neighborhoods. Two versions of OperonViz are embedded in the system; OperonViz-COG uses COG database to identify homologs and OperonViz-BAG uses PCDB and the BAG clustering algorithm for the same purpose. If the distance is shorter than a given value (Default value is 200 bp), two genes are considered to belong to the same gene clusters (Rogozin et al., 2002).

*To whom correspondence should be addressed.
OperonViz is useful to identify horizontal gene transfers, functional coupling and functional hitchhiking.

Gene fusion event detection. FuzFinder uses PCDB to identify plausible gene fusion events among a set of submitted genomes. The definition of mutual best hit is as follows: (i) each of the two reference genes must match the same open reading frame (ORF) in the target genome with a higher Z-score than a given value; (ii) when split between the two hits, the two halves of the target ORF must match back to the original two reference genes with a higher Z-score than a given value; and (iii) the reference genes must not be homologous to each other (Suhre and Claverie, 2004). Although Z-score is a statistical score that depends on the database size, users can use the default value as genomes are fairly large. Of course, we provide an option to change the Z-score cut-off for pairwise matches.

Metabolic pathway analysis. MetaPath is a metabolic pathway analysis tool. It combines metabolic pathway information at KEGG and sequence information at GenBank to reconstruct metabolic pathways among the selected genomes. This tool aims to find missing genes in metabolic pathways by comparing reference genome with a set of genome selection. The result is represented as a table, but the directionality of metabolic pathway is not considered at this stage because of the lack of such information in KEGG database. MetaPath web service is limited only to prokaryotic genomes.

Gene clustering tools. Users can upload a set of protein sequences in the FASTA format using FASTA-BAG and
BLASTP-BAG or select genomes from the genome list using Genome-BAG service (Genome-BAG) for clustering analysis using BAG (Kim, 2003).

FUTURE WORK
The PLATCOM system is designed to evolve through three development stages. Only its first stage is complete: the underlying architecture and individual system modules. Although system modules in PLATCOM are designed to work in a cooperative manner at the system level, single integrated interfaces for specific tasks need to be developed to provide the integrated service on the Web. We plan to provide as many such interfaces as possible. However, the ultimate goal is to provide a flexible, reconfigurable system where users can combine different tools freely. This goal will be achieved through the second and third stages. The system modules will be integrated by gluing them together on the biological sequence level using high-performance data mining tools, e.g. BAG (Kim, 2003), and a genome analysis language of our own. In addition to sequence data, PLATCOM will include more data types such as gene expression data. As a result, a flexible, reconfigurable environment for comparative genomics will be provided.

ACKNOWLEDGEMENTS
We appreciate anonymous reviewers for their valuable comments. This work is partially supported by NSF CAREER DBI-0237901, Indiana Genomics Initiative and NSF 0116050.

REFERENCES