# PLATCOM: a Platform for Computational Comparative Genomics on the Web

Kwangmin Choi, Jeong-Hyeon Choi, Amit Saple, Zhiping Wang, Jason Lee, and Sun Kim
School of Informatics, Center for Genomics and Bioinformatics
Indiana University, IN 47408, USA
{kwchoi,jeochoi,asaple,zhipwang,jhl2,sunkim2}@indiana.edu

## 1 Introduction

The exponential accumulation of genomic sequence data demands systematic analysis of genetic information and requires use of various computational approaches to handle such huge sets of genomic data. Comparative genomics, with such organized data and diverse computational techniques, has become useful not only for finding common features in different genomes, but also for understanding evolutionary process and mechanism among multiple genomes.

Comparison of multiple genomes is a challenging task partially because combining multiple tools for sequence analysis requires a significant amount of programming work and knowledge on each tool and partially because it handles a huge amount of data. Another problem is the subjectivity of how to select multiple genomes. For example, there are 1,313,400 (= $\binom{200}{3}$ ) possible selections of three genomes out of 200 completely sequenced genomes. The inconsistency of input data from existing sources and the effective presentation of analysis result also raise problems.

Considering all these issues, it is not possible to perform multiple genome comparison on the web by simply using sequence analysis tools in an *ad hoc* fashion.

## 2 PLATCOM

We have been developing a genome comparison system PLATCOM, which is available at `http://platcom.informatics.indiana.edu`. PLATCOM is designed to be a high performance genome analysis system on the web which is easy to use and easy to maintain and update. PLATCOM is designed to perform sequence analysis on selected genomes. More information on sequences can be obtained via URLs or connectivity tools to other information rich databases.

**System architecture:** PLATCOM consists of four main components; databases, sequence analysis tools, genome analysis modules, and user interface. The whole system is built on internal databases, which consist of GenBank, SwissProt, COG, KEGG, and Pairwise Comparison Database (PCDB).

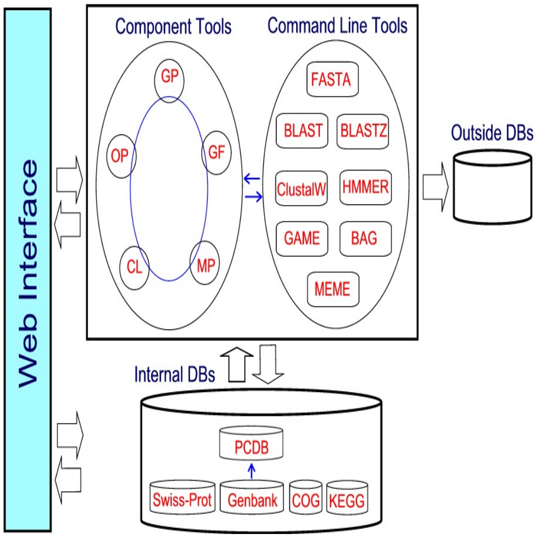

**Figure 1. Architecture of PLATCOM**

PCDB is designed to incorporate new genomes automatically so that PLATCOM can evolve as new genomes become available. All pairwise comparison for a new genome is performed using an AVIDD Linux cluster at Indiana and then the result is automatically transfered to the PLATCOM server [5]. 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 compute, but the pre-computed PCDB makes it possible to complete genome analysis very fast even on the web which allows users to select any subset of genomes to be compared freely.

Sequence analysis tools include widely used high performance sequence analysis tools such as FASTA, BLAST, BLASTZ, HMMER, GIBBS, and MEME. We have also developed high performance data mining tools of our own as described in the next section. With the databases and sequence analysis tools, genomes can be compared. There are currently six modules: genome plot, conserved gene neighborhood navigation, metabolic pathways, comparative sequence clustering analysis, putative gene fusion events detection, and multiple genome alignment. A set of genomes selected by users is submitted with parameter settings via web interface.

**Multi-step sequence analysis with scalable data mining tools:** Data mining techniques are useful in combining many sequence analysis tools and databases that can be utilized for genome annotation since data mining tools encapsulate multiple sequence analysis tasks in a single step. Thus well-defined data mining concepts and tools can make genome comparison much easier. It is also important that the data mining tools for genome comparison should be scalable. We have been developing such scalable tools: a sequence clustering algorithm BAG [3], a genome sequence alignment tool GAME [2], an algorithm for mining correlated gene sets [4], and a multiple genome sequence alignment algorithm by clustering local matches [1]. To summarize the analysis result, we have developed visualization tools for genome plot, multi-domain, gene-genome matching table, and genome alignment.

## 3 Plan

We have been using high performance sequence analysis tools to "simplify" sequence analysis tasks. For example, the BAG clustering tool can generate a set of sequence clusters in a single operation, rather than performing many sequence data searches using FASTA or BLAST and then combining the search results. However, our goal is to provide a web-based environment for genome comparison. To achieve this goal, many sequence analysis and data mining tools should be combined freely. Our approach is to introduce several data types for genome analysis so that sequence analysis and data mining tools can be combined using these data types. Almost all sequence analysis and data mining tools can be viewed as functions on the domain of "a set of sequences" and "a set of genomes". Thus we introduce two data types, $\mathcal{S}$ for a set of sequences and $\mathcal{G}$ for a set of genomes. To allow users to select sequences and genomes, we introduce to selection functions, $I_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{S}'$ to select a set $\mathcal{S}'$ of sequences from $\mathcal{S}$, and $I_{\mathcal{G}} : \mathcal{G} \rightarrow \mathcal{G}'$ to select a set $\mathcal{G}'$ of genomes from $\mathcal{G}$. $I_{\mathcal{S}}$ and $I_{\mathcal{G}}$ are implemented as a web page where a set of sequences or genomes are listed and then users select subsets by clicking checkboxes.

We briefly illustrate this concept using an example of combining two existing modules in PLATCOM. GenomePlot($G_i, G_j$) computes and plots gene matches in $G_i$ and $G_j$, and GeneClusterSearch($S, \mathcal{G}$) searches a set $\mathcal{G}$ of genomes for matches of a given set $S$ of sequences. These two modeules can be combined as follows:

1. GenomePlot( $I_{\mathcal{G}}$ (all genomes in PLATCOM)) generates a set of gene matches $MG$.
2. MCGS($MG$) [4] computes a set of gene clusters, $\{GC_1, \ldots, GC_k\}$.
3. For any $GC_i$, users can perform GeneClusterSearch($I_{\mathcal{S}}(GC_i), I_{\mathcal{G}}(all \quad genomes)$) where users select a set of genes from $GC_i$ via $I_{\mathcal{S}}(GC_i)$ and searches its occurrences in a set of selected genomes via $I_{\mathcal{G}}(all\ genomes)$).

We are currently working on a complete implementation of this concept in order to provide a flexible genome comparison environment on the web.

## Acknowledgments

This work partially supported by INGEN (Indiana Genomics Initiatives) and NSF CAREER Award DBI-0237901.

## References

[1] J. Choi, K. Choi, H.-G. Cho, and S. Kim. Multiple genome alignment by clustering pairwise matches. In J. Lagergren, editor, *Comparative Genomics, RECOMB 2004 International Workshop, Lecture Notes in Bioinformatics 3388*, pages 30–41. Springer, 2004.

[2] J.-H. Choi, H.-G. Cho, and S. Kim. GAME: A Simple and Efficient Whole Genome Alignment Method Using Maximal Exact Match Filtering. *Computational Biology and Chemistry*, To appear 2005.

[3] S. Kim. Graph theoretic sequence clustering algorithms and their applications to genome comparison. In J. T. L. Wang, C. H. Wu, and P. Wang, editors, *Computational Biology and Genome Informatics*, pages 81–116. World Scientific, 2003.

[4] S. Kim, J. Choi, and J. Yang. Gene Teams with Relaxed Proximity Constraint. *Proc. IEEE Computational Systems Bioinformatics Conference (CSB) 2005*, To appear 2005.

[5] Y. Ma, R. Bramley, and S. Kim. A Data Management Architecture for Computational Biology. *Proc. 3rd International Conference on Computing, Communications and Control Technologies (CCCT05)*, To appear 2005.