

Genome Analysis Toolkit and Server

M. Shah, I. Vokler, D. Hyatt, M. Land, F. Larimer, P. Locascio, G. Chen, V. Olman, J. Snoddy, E. Uberbacher, Oak Ridge National Laboratory, Oak Ridge, TN 37831

In the past year we have continued to enhance and extend the capabilities of the Genome Analysis Toolkit (**GAT**), made improvements to the individual tools incorporated in the toolkit, redesigned the client server system that provides access to the toolkit services, and made refinements to the Web interfaces for access to the toolkit. This system is used for the analysis of all sequence data in ORNL Genome Channel and Genome Catalog, consisting of finished and draft human and mouse genomes sequences, as well as finished and draft microbial genome sequences.

The toolkit now includes a wider variety of genome analysis tools. Gene finding systems include the new version of GrailEXP (version 3.2) and Genscan for eukaryotic genomes, and Generation and Glimmer for microbial gene prediction. Grail suite of feature-finding tools provides analysis for CpG islands, PolyA sites and simple and complex repetitive elements, and BAC End identification. Also included are NCBI STS E-PCR, Pfam, RepeatMasker and tRNAscan-se systems. Similarity search tools include NCBI BLAST, and Beauty post-processing.

The client server system has been redesigned to facilitate handling of increased load on the system. The client server protocol now incorporates issuance of a ticket (request ID), and an estimate of processing time for the submitted request, by the server to the client. This mechanism eliminates the need for maintaining the client – server connection for a long duration, since the client can query the server for results at a later time. The server distributes the incoming requests intelligently on the available pool of compute server machines, based on a set of predefined criteria. Highly compute-intensive service requests (like BLAST, Pfam and RepeatMasker) are redirected to the **GIST** (Genomic Integrated Supercomputing Toolkit) server running on ORNL supercomputers.

Refinements to the GrailEXP gene finding system have resulted in significant improvement in its gene prediction accuracy. The database search module of GrailEXP provides very accurate EST/mRNA alignments with the genomic sequence being analyzed. GrailEXP utilizes this evidence to identify alternatively spliced gene models. We are currently benchmarking this version of GrailEXP and plan to present the results at the meeting.

Web access to the analysis tools is provided at <http://compbio.ornl.gov/tools>

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-00OR22725 with UT-Battelle, LLC)