Oasis: A Flexible Framework for High Throughput Sequence Analysis

Craig Tulig, Erik Sassaman, Jim Brown

Genome Therapeutics Corporation

craig.tulig@genomecorp.com, erik.sassaman@genomecorp.com, jim.brown@genomecorp.com

Oasis was designed and implemented as a robust system for generating, storing, and presenting sequence analysis data. Development of Oasis began in 1996 and implementation was started in 1997. The Oasis system integrates data from many different projects at GTC including Osteoporosis, High Bone Mass, Asthma, Tangiers, and Yeast 2 Hybrid. These projects produce large amounts of novel DNA sequence that must be presented to biologists in an intelligible fashion. Each project has specific needs, however, the same framework can be used across projects. To accomplish this, Oasis was engineered with an emphasis on flexibility and configuration.

Tonka is the Oasis back end module responsible for executing various sequence analysis protocols. This Workflow Engine defines a protocol language, allowing for complete flexibility in the way sequence analysis tools are applied to different DNA sequence types. Tonka currently has commands for Masking (cross_match), BLAST, MZEF, and GENSCAN, other tools are under evaluation for future incorporation. With so many different types of sequence analysis tools available, it is crucial to have facilities that ease the addition of a new tool. Tonka allows for sub protocol instructions that are nested and reusable. Protocols are stored in Sybase, then retrieved and viewed via a Perl/CGI web interface, facilitating verification by the scientists.

DNA data at GTC is collected from a variety of informatics sources, including our ABI Sequencing Pipeline, Sequencing Finishing Workbench, and submission of sequences of interest from public sources. The Workflow Engine utilizes APIs to interface with data sources, integrating, and formatting DNA sequences from these disparate sources. Providing data interaction with EST clustering tools is a challenging task due to complex formatting, and data inclusion issues. Once the Workflow Engine converts input DNA sequence to Oasis format, the appropriate sequence analysis protocols must be executed.

The Oasis web based front end employs Java based bioWidgets to present the sequence analysis results in a graphical representation. The Genomic and cDNA views show the user exons and BLAST hits in a color-coded aligned fashion. The Group and Annotation facilities allow multiple users to create groups of related sequence and annotate individual sequences and groups of sequences. User specific filtering preferences allow scientists to view blast hits based on Percent Identity, Alignment Length and other criteria.

The Oasis system provides a scalable architecture for efficiently analyzing gigabytes of data, storing these data in Sybase, retrieving and graphically displaying selected sequence analysis information based on user preferences.

S. Altschul, G. Warren, M. Webb, M. Eugene, and D. Lipman (1990). Basic local alignment search tool. J. Mol. Biol., 215:403-10.

C. Burge, S. Karlin, (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268:78-94.

J. Burke, H. Wang, W. Hide, D.B. Davidson. (1998) Alternative Gene Form Discovery and Candidate Gene Selection from Gene Indexing Projects. Genomic Research, 8:276-290.

CBIL bioWidgets for Java, http://www.cbil.upenn.edu/bioWidgets.

P. Green, cross_match, Copyright (C) 1993-1996.

M.Q.Zhang (1997). Identification of Protein Coding Regions in the Human Genome Based on Quadratic Discriminant Analysis. PNAS, 94:565-568.