ORNL/CP-94810

CONF-971146--

GRAIL and GenQuest Sequence Annotation Tools*

Ying Xu, Manesh B. Shah, J. Ralph Einstein, Morey Parang,
Jay Snoddy, Sergey Petrov, Victor Olman, Ge Zhang,
Richard J. Mural, and Edward C. Uberbacher

Computational Biosciences Section
Life Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6480

To be presented at the *The Sixth Department of Energy Contractor and Grantee Workshop*, Santa Fe, NM, November 9-13, 1997.

*Research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

GRAIL and GenQuest Sequence Annotation Tools

Ying Xu, Manesh B. Shah, J. Ralph Einstein, Morey Parang, Jay Snoddy,
Sergey Petrov, Victor Olman, Ge Zhang, Richard J. Mural and Edward C. Uberbacher

Computational Biosciences Section, Life Sciences Division,
Oak Ridge National Laboratory, Oak Ridge, TN 37831-1060
Email: GRAILMAIL@*ornl.gov*

Our goal is to develop and implement an integrated intelligent system which can recognize biologically significant features in DNA sequence and provide insight into the organization and function of regions of genomic DNA. GRAIL is a modular expert system which facilitates the recognition of gene features and provides an environment for the construction of sequence annotation. The last several years have seen a rapid evolution of the technology for analyzing genomic DNA sequences. The current GRAIL systems (including the e-mail, XGRAIL, JAVA-GRAIL and genQuest systems) are perhaps the most widely used, comprehensive, and user friendly systems available for computational characterization of genomic DNA sequence. In the past 2 years of the project we have:

- Developed improved systems for recognition of exons, splice junctions, promoter elements and other features of biological importance, including greater sensitivity for exon prediction (especially in AT rich regions) and robust indel error detection capability.

- Developed improved and more efficient algorithms for constructing models of the spliced mRNA products of human genes.

- Developed and implemented methods for the analysis and visualization of sequence features including poly-A addition sites, potential Pol II promoters, CpG islands and repetitive DNA elements.

- Designed and implemented new methods for detecting potential sequence errors which can be used to "correct" frameshifts, add quality assurance to sequencing operations, and better detect coding regions in low pass sequences such as ESTs.

- Developed systems for a number of model organisms including mouse, *Escherichia coli, Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae* and a number of microbial genomes.

- Implemented methods for the incorporation of protein, EST, and mRNA sequence evidence in the multiple gene modeling process.

- Constructed a powerful and intuitive graphical user interface and client-server architecture which supports Unix workstations and JAVA Web-based access from many platforms.

- Improved algorithms and infrastructure in the genQuest server, allowing characterization of newly obtained sequences by homology-based methods using a number of protein, DNA, and motif databases and comparison methods such as FASTA, BLAST, parallel Smith-Waterman, and algorithms which consider potential frameshifts during sequence comparison.

- An improved "batch" GRAIL client allows users to analyze groups of short (300-400 bp) sequences for coding character (with frameshift compensation options) and automates database searches of translations of putative coding regions.

- Provided support for GRAIL use in more than a thousand laboratories and at a rate of over 4000 analysis requests per month.

The imminent wealth of genomic sequence data will present significant new challenges for sequence analysis systems. Our vision for the future entails incorporation of a more sophisticated view of biology into the GRAIL system. Computational systems for genome analysis have thus far focused on generic or textbook-like examples of single isolated genes which can be described fairly simply using the most usual assumptions, and fall far short of the intelligence necessary to interpret complex multiple gene domains. In its next phase the GRAIL project will involve the development of new pattern recognition methods and modeling algorithms for DNA sequence, expert systems for interpretation using experimental evidence and comparative genomics, and interoperation with other tools and databases. More specifically we will focus on several development areas:

(1) Improved accuracy of feature recognition and greatly increased tolerance to sequencing errors,

(2) development of technology to describe the structure and regulation of large, complex genomic regions containing multiple genes,

(3) automated and interactive methods for the incorporation of experimental evidence such as ESTs, mRNAs, and protein sequence homologs in multi-gene domains (GRAIL-EXP),

(4) more comprehensive feature recognition and increased biological sophistication in the areas of expression and regulation,

(5) capabilities for direct comparison of genomes,

(6) a comprehensive suite of microbial genome analysis systems,

(7) infrastructure for use of high-performance computing systems and specialized hardware to facilitate analysis and annotation of large volumes of sequence data,

(8) improved interoperation with other tools, databases and methods for integrating information from multiple sources, particularly within the *Genome Annotation Consortium* framework and *Genome Channel*, and

(9) continued community and user support, technology transfer, and educational outreach.

These developments will enable GRAIL to become more comprehensive and biologically sophisticated, and yet remain a user-friendly analysis environment which can be used interactively or in fully automated modes.

GRAIL and genQuest related tools are available as a Motif Graphical client (anonymous ftp from grail.lsd.ornl.gov (134.167.140.9)), through WWW interfaces (URL http://www.compbio.lsd.ornl.gov/), or by email server (*grail@ornl.gov*) and genQuest at (*Q@ornl.gov*). Communications with the GRAIL staff should be addressed to *GRAILMAIL@ornl.gov*. (Supported by the Office of Health and Environmental Research, United States Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.)

M98000582

Report Number (14) ORNL/CP-94810

CONF-971146--

Publ. Date (11) 199709

Sponsor Code (18) DOE/ER, XF

UC Category (19) UC-400, DOE/ER

DOE