thus requiring only a weak form of the molecular-clock hypothesis.

The endpoints of the similarity between genomic DNA sequences are often ambiguous. The probability of evolution at each evolutionary distance can be estimated over the entire set of alignments by choosing the best alignment at each distance and the corresponding probability of duplication at that evolutionary distance. A central value of this distribution provides a robust evolutionary distance estimate. We provide an efficient algorithm for computing the parametric alignment, considering evolutionary distance as the only parameter.

These techniques and estimates are used to infer the duplication history of the genomic sequence in C. elegans and in S. cerevisae. Our results indicate that repeats discovered using a single scoring matrix show a considerable bias in subsequent evolutionary distance estimates.

Model based sequence scoring metrics: PAM based DNA comparison metric has been extended to incorporate biases in nucleotide composition and mutation rates, extending earlier work (States, Gish and Altschul, 1993). A codon based scoring system has been developed that incorporates the effects biased codon utilization frequencies.

A dynamic programming algorithm has been developed that will optimally align sequences using a choice of comparison measures (non-coding vs. coding, etc.). We are in the process of evaluating this approach as a means for identifying likely coding regions in cDNA sequences.

Efficient sequence similarity search tools: Most sequence search tools have been designed for use with protein sequence queries a few hundred residues long. The analysis of genomic DNA sequence necessitates the use of queries hundreds of kilobases or even megabases in length. A memory and computationally efficient search tool has been developed for the identification of repeats and sequence similarity in very large segments of nucleic acid sequence. The tool implements optimal encoding of the word table, repeat filters, flexible scoring systems, and analytically parameterized search sensitivity. Output formats are designed for the presentation of genomic sequence searches.

Federated databases: A sybase server and mirror for GSDB are being developed to facilitate the annotation of repeat sequence elements in public data repositories.

DOE Grant No. DE-FG02-94ER61910.

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Ying Xu, Manesh Shah, J. Ralph Einstein, Sherri Matis, Xiaojun Guan, Sergey Petrov, Loren Hauser,[1] Richard J. Mural,[1] and **Edward C. Uberbacher**
Computer Science and Mathematics and [1]Biology Divisions; Oak Ridge National Laboratory; Oak Ridge, TN 37831
Uberbacher: 423/574-6134, Fax: -7860, *ube@ornl.gov*
http://compbio.ornl.gov

GRAIL is a modular expert system for the analysis and characterization of DNA sequences which facilitates the recognition of gene features and gene modeling. A new version of the system has been created with greater sensitivity for exon prediction (especially in AT rich regions), more accurate splice site prediction, and robust indel error detection capability. GRAIL 1.3 is available to the user in a Motif graphical client-server system (XGRAIL), through WWW-Netscape, by e-mail server, or callable from other analysis programs using Unix sockets.

In addition to the positions of protein coding regions and gene models, the user can view the positions of a number of other features including poly-A addition sites, potential Pol II promoters, CpG islands and both complex and simple repetitive DNA elements using algorithms developed at ORNL. XGRAIL also has a direct link to 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 special algorithms which consider potential frameshifts during sequence comparison.

Following an analysis session, the user can use an annotation tool which is part of the XGRAIL 1.3 system to generate a "feature table" report describing the current sequence and its properties. Links to the GSDB sequence database have been established to record computer-based analysis of sequences during submission to the database or as third party annotation.

Gene Modeling and Client-Server GRAIL: In addition to the current coding region recognition capabilities based on a multiple sensor-neural network and rule base, modules for the recognition of features such as splice junctions, transcription and translation start and stop, and other control regions have been constructed and incorporated into an expert system (GAP III) for reliable computer-based modeling of genes. Heuristic methods and dynamic programming are used to construct first pass gene models which include the potential for modification of initially predicted exons. These actions result in a net improvement in gene characterization, particularly in the rec-