A Software Environment for Large-Scale Sequencing

Mark Graves
Department of Cell Biology; Baylor College of Medicine; Houston, TX 77030
713/798-8271, Fax: -3759; *mgraves@bcm.tmc.edu*
http://www.bcm.tmc.edu
http://stork.bcm.tmc.edu/gfp

Our approach is to implement software systems which manage primary laboratory sequence data and explore and annotate functional information in genome sequence and gene products.

Three software systems have been developed and are being used: two sequence data managers which use different sequence assembly packages, FAK and Phrap, and a series of analysis and annotation tools which are available via the Internet. In addition, we have developed a prototype application for data mining of sequence data as it is related to metabolic pathways.

Products of this project are the following:

1. GRM -a sequence reconstruction manager using the FAQ assembly engine (available since October 1995).

2. GFP -a sequence finishing support tool using the Phrap assembly engine (available since March 1996).

3. A series of gene recognition tools (available since early 1996).

4. A tool for visualizing metabolic pathways data and exploring sequence data related to metabolic pathways (prototype available since August 1996).

DOE Grant No. DE-FG03-94ER61618.

Generalized Hidden Markov Models for Genomic Sequence Analysis

David Haussler, Kevin Karplus,[1] and Richard Hughey[1]
Computer Science Department and [1]Computer Engineering Department; University of California; Santa Cruz, CA 95064
408/459 2105, Fax: -4829, *haussler@cse.ucsc.edu*
http://www.cse.ucsc.edu/research/compbio
http://www-hgc.lbl.gov/projects/genie.html

We have developed an integrated probabilistic method for locating genes in human DNA based on a generalized hidden Markov model (HMM). Each state of a generalized HMM represents a particular kind of region in DNA, such as an initial exon for a gene. The states are connected by transitions that model sites in DNA between adjacent regions, e.g. splice sites. In the full HMM, parametric statistical models are estimated for each of the states and transitions. Generalized HMMs allow a variety of choices for these models, such as neural networks, high order Markov models, etc. All that is required is that each model return a likelihood for the kind of region or transition it is supposed to model. These likelihoods are then combined by a dynamic programming method to compute the most likely annotation for a given DNA contig. Here the annotation simply consists of the locations of the transitions identified in the DNA, and the labeling of the regions between transitions with their corresponding states.

This method has been implemented in the genefinding program Genie, in collaboration with Frank Eeckman, Martin Reese and Nomi Harris at Lawrence Berkeley Labs. David Kulp, at UCSC, has been responsible for the core implementation. Martin Reese developed the splice site models, promoter models, and datasets. You can access Genie at the second www address given above, submit sequences, and have them annotated. Nomi Harris has written a display tool called Genotater that displays Genie's annotation along with the annotation of other genefinders, as well as the location of repetitive DNA, BLAST hits to the protein database, and other useful information. Papers and further information about Genie can be found at the first www address above. Since the ISMB '96 paper, Genie's exon models have been extended to explicitly incorporate BLAST and BLOCKS database hits into their probabilistic framework. This results in a substantial increase in gene predicting accuracy. Experimental results in tests using a standard set of annotated genes showed that Genie identified 95% of coding nucleotides correctly with a specificity of 88%, and 76% of exons were identified exactly.

DOE Grant No. DE-FG03-95ER62112.

Identification, Organization, and Analysis of Mammalian Repetitive DNA Information

Jerzy Jurka
Genetic Information Research Institute; Palo Alto, CA 94306
415/326-5588 Fax: -2001, *jurka@gnomic.stanford.edu*
http://charon.lpi.org

There are three major objectives in this project: organization of databases of mammalian repetitive sequences, development of specialized software for analysis of repetitive DNA, and sequence studies of new mammalian repeats.

Our approach is based on extensive usage of computer tools to investigate and organize publicly available sequence information. We also pursue collaborative research