Microbial Annotation and Analysis: Manatee and the Comprehensive Microbial Resource

Lauren Brinkac, Tanja M. Davidsen, and Owen White

The Institute for Genomic Research, Rockville, MD, U.S.A.

The ever-increasing volume of genomic data and wide variety of query tools available for microbial annotation and analysis has been a challenge for the efficient utilization of bioinformatic resources and effective display of consistent, uniform information to the scientific community. Manatee (Manual Annotation Tool Etc, Etc) (http://manatee.sourceforge.net/) is a web-based gene evaluation and genome annotation tool designed for prokaryotic and eukaryotic genomic analysis. Its unique interface combines a diverse selection of analysis tools, enabling the efficient identification and high quality functional assignment of genes. Such tools include homology searches, Hidden Markov Models, Prosite motifs, paralogous families, GO classification, and suggestions from automated functional assignments generated by the Annotation Engine. Currently, Manatee and the Annotation Engine protocol are integrated into production at TIGR and have been successful for the manual annotation of several prokaryotic and eukaryotic genomes. Completed microbial genomes are displayed on the Comprehensive Microbial Resource (CMR) (http://www.tigr.rg/tigr-scripts/CMR2/CMRHomePage.spl) data management system, which contains an assortment of data mining and retrieval resources used for single gene and whole comparative genomic analyses. This suite of bioinformatic tools facilitates multi-genomic applications, analyses, and searches of all completed microbial genomic sequences available in one location.

Introduction

Annotation Engine is an automated annotation service run by The Institute for Genomic Research (TIGR). It is a service provided at no cost to sequencing centers for preparation, processing, and automated curation of genomic sequence data. Genomic sequence submitted to the Annotation Engine is processed according to TIGR's annotation pipeline, with resulting data returned to the sequencing center for analysis as a MySQL dataset. Data obtained from the Annotation Engine and supplied to the sequencing center include: all predicted open reading frames and RNAs, their underlying homology search results and protein properties, and all associated functional assignments such as common name and gene symbol, enzyme commission (EC) number, TIGR role category, and GO classification. Results can be viewed and modified using a freely available web browser annotation tool, Manatee, for the manual curation of genomic sequence data. Completed data can then be displayed on the Comprehensive Microbial Resource for additional analyses.

Gene Prediction

The Annotation Engine first identifies all coding regions of a genome using the Glimmer system (Salzberg et al., 1998; Delcher et al., 1999). Glimmer uses an interpolated markov model to score new sequences for representative genes based a set of known genes. This training set of known genes is composed of non overlapping open reading frames 500-1000bp in length depending on GC content of the organism, as well as open reading frames with significant homology to other organisms using BLAST searches (Altschul, et al.,

1990). Glimmer then uses this training set to build an interpolated markov model and predict putative genes, resolving any gene overlap. Glimmer has been shown to reliably identify 99% of the genes of bacterial, archaeal, or viral genomes in a fully automated fashion. The Glimmer system is freely available to nonprofit research institutions, and has been distributed to hundreds of sites worldwide (http://www.tigr.org/software).

Homology Searches

BLAST-extend-repraze (BER)

After Glimmer identifies candidate genes, the translated proteins are queried against an internal non-redundant amino acid database (NRAA) using BLAST searches (Altschul, et al., 1990). NRAA is composed of all proteins maintained at GenBank (http://www.ncbi.nlm.nih.gov), SwissProt (http://www expasy.ch/sprot), PIR (http://pir.georgetown.edu), and TIGR's CMR database, the Omniome (http://www. igr.org/tigr-scripts/CMR2/CMRHomePage.spl). Significant hits are stored in a mini-database. In order to identify potential frameshifts or point mutations in the sequence, a modified Smith-Waterman alignment (Smith and Waterman, 1981) is performed on the protein against the mini-database. Genes are extended 300bp upstream and downstream of the predicted coding region. If significant homology exists to a match protein in the mini-database and extends into a different frame from that predicted, or extends beyond a stop codon, the program signals either a frameshift or point mutation by continuing the alignment past the boundaries of the predicted coding region. Both full length pairwise and multiple sequence alignments of the top scoring matches can be viewed though Manatee.

Hidden Markov Models

Proteins are queried against two complementary sets of hidden markov models (HMMs), Pfam HMMs (Bateman, et al., 2000), and TIGRFAMs (Haft, et al., 2001), using the program hmmpfam (Eddy, 1999). Built from a multiple alignment of highly curated protein sequences thought to share the same function or belong to the same protein family. HMMs are designed to support the automated functional identification of proteins by sequence homology. Each HMM is assigned a specific cutoff score for which hits are known to be significant protein matches, i.e., membership in the family/function for which the model was built. Membership can be classified into several isology types.

Equivalog - all members share the same function.

Superfamily - members which have full length protein sequence similarity and similar domain architecture. May or may not have an associated function.

Subfamily - members which have full length homology, but which do not necessarily represent an entire superfamily.

Domain - members with regions of homology, varying in length. May or may not have an associated function.

Hypothetical_equivalog - all members share the same function, although the function is not known.

Models classified as equivalog strive to provide the best information suitable for applying functionally accurate gene assignments in an automated fashion. To date, >1500 TIGRFAMs have been built, with >900 classified as equivalogs. A graphic display of HMM results can be viewed through Manatee, and additional information can be found at http://www.tigr.org/TIGRFAMs/index.shtml.

Additional Searches

In addition to HMM results, a variety of other searches are run for each protein sequence and are graphically displayed on Manatee. These searches include PROSITE motifs (http://us.expasy.org/prosite/) and InterPro data (http://www.ebi.ac.uk/interpro/). Protein properties such as molecular weight and pI, signal peptide prediction, third position GC skew, GC content, membrane spanning regions, and lipoprotein lipid attachment sites are also determined and displayed. Proteins are classified into TIGR role categories along with suggested GO classifications where appropriate.

AutoAnnotate

Autoannotate is an automated process for assigning common name, gene symbol, enzyme commission number, role category, and GO term for each gene. It uses an assignment hierarchy whereby the associated information is assigned based on the significance and type of evidence derived by homology search results. Autoannotate first accesses HMM results. If a protein scores above the trusted cutoff for an equivalog level HMM, then the associated identifying information (common name, role category, gene symbol and EC number if applicable) is assigned to the predicted coding region. If there is no equivalog level HMM hit, then Autoannotate evaluates the BER results for the best full length match homologous >80% similarity and >35% identity. When BER results are only "hypothetical proteins" and/or do not fulfill BER match criteria, then non-equivalog level HMMs are used. If a hit exists, then the protein will be assigned to a family membership based on HMM name. Proteins with no HMM hit and full length BER matches to "hypothetical proteins" from another species, are named "conserved hypothetical protein". Proteins with no HMM or BER matches remain "hypothetical protein".

Manatee

Once genes are identified, searches are run, and Autoannotate determines a gene assignment, the resulting data is made available for manual curation using the web based annotation tool, Manatee. Manatee is an open source Perl program run under CGI of a web server, such as Apache. Manatee requires at least one MySQL (or Sybase) project database with associated search files. These project databases use a data model and schema developed at and used by TIGR for representing genomic data. Additional information, software requirements, and installation instructions can be found at http://manatee.sourceforge.net/.

Comprehensive Microbial Resource

The Comprehensive Microbial Resource (CMR) houses sequence data from every completed bacterial genome to date across all sequencing centers, as well as a suite of analytical tools designed for public access and in-depth analyses on all or a subset of genomic sequences in one location (Peterson, et al., 2001). Currently, the CMR contains 74 organisms, across 35 different sequencing centers. For genomes not sequenced at TIGR, the CMR displays two types of annotation stored in a database called the Omniome. In addition to the primary annotation generated by each sequencing center, automated assignments are generated following TIGR's annotation engine pipeline and included for each genome sequence. In cases where the identical gene is identified, both annotations are linked, ensuring an entirely non-redundant dataset Additionally, the display of uniform datatypes across genomes enables efficient cross-genome analyses and sophisticated data retrievals. The tools available for such analysis can be classified into three sections: individual genome, individual gene, and cross genome analyses.

Individual Genome

Each genome loaded into the CMR has a specific genome information page listing general properties of the genome such as sequencing center, publication, and taxonomic classification, as well as schematic representations of the genome in either circular or linear displays. The displays are interactive, enabling the user to enlarge specific sections for additional information and detailed analysis. Detailed information for each DNA molecule found in the organism is provided: topology, length, %A, T, G, C and number of genes. In order to determine the closest homolog in the completed set of genomes, prerun blast searches of all sequences against themselves in the Omniome are displayed. Programs on the CMR enable users to display the results in various graphical formats, either by total protein hits or best protein hits. Searches are routinely rerun as new genomic sequences are integrated into the Omniome. All TIGR genome sequence data is available for download (ftp://ftp.tigr.org/pub/data/).

Individual Gene

Every gene in the CMR has an associated information page that is linked throughout the CMR. This page enables sequence retrieval and display of primary and TIGR annotation. Hydrophobicity, GC content, secondary structure, and third position GC skew are just some of the few analytical tools available for each sequence. Additionally, prerun pairwise alignments of each gene's protein to all other proteins stored in the CMR can be viewed. Alignments are rerun routinely when new genomic data is loaded into the CMR.

Cross Genome

Complex queries across all genomes can be based on any one property(s) of an organism such as pI, molecular weight, and GC-content, and/or queries incorporating attributes such as taxon, paralogy, functional role and GO classification, and protein similarity criteria. Genomes can be queried according to locus, common name, enzyme commission number, or gene symbol. Users can perform blast and hidden markov model homology searches of a particular sequence against all sequences represented in the CMR, and whole genome alignments can be calculated and graphically displayed using the MUMmer algorithm (Delcher, et al., 1999).

Future Directions

TIGR offers a two day course in Prokaryotic Annotation and Analysis to familiarize users with TIGR's prokaryotic annotation tools and the analyses available on the Comprehensive Microbial Resource (CMR). Scientists interested in utilizing TIGR's Annotation Engine service are strongly encouraged to attend. Detailed information and class schedule can be found at http://webtest/edutrain/training/prokaryotic.shtml.

References

1. Altschul S., et al. Basic local alignment search tool. J. Mol. Biol., 215: 403-410 (1990).
2. Bateman A., et al. The Pfam protein families database. Nucleic Acids Res. 28(1): 263-266 (2000).
3. Delcher A.L., et al. Alignment of whole genomes. Nucleic Acids Res., 27(11):2369-2376 (1999).
4. Delcher A.L., et al. Improved Microbial Gene Identification with Glimmer. Nucleic Acids Res., 27(23): 4636-4641 (1999).
5. Eddy S. Profile hidden Markov models. Bioinformatics, 14(9):755-763 (1998).
6. Haft D., et al. TIGRFAMs: A protein family resource for the functional identification of proteins. Nucleic Acids Res. 29(1): 41-3 (2001).
7. Peterson J.D., et al. The Comprehensive Microbial Resource. Nucleic Acids Res., 29(1): 123-125 (2001).
8. Salzberg S., et al. Microbial Gene Identification using Interpolated Markov Models. Nucleic Acids Res., 26(2): 544-548 (1998).
9. Smith T.F. and M. Waterman. Identification of common molecular subsequences. J. Mol. Biol. 147(1): 195-197 (1981).