

Section II - Functional Annotation

The translated sequence of each gene model (identified by Glimmer or other methods) is searched against a variety of public and private databases. These search results are either stored in genome project specific databases, or maintained as protein specific search files retrievable for analysis.

II.1 Homology Searches

II.1.1 Non-identical amino acid database (NIAA)

Each protein is searched against an internal non-identical amino acid database (NIAA) comprised of all proteins available from GenBank ([HYPERLINK "http://www.ncbi.nlm.nih.gov"](http://www.ncbi.nlm.nih.gov) <http://www.ncbi.nlm.nih.gov>), PDB ([HYPERLINK "http://www.rcsb.org/pdb/Welcome.do"](http://www.rcsb.org/pdb/Welcome.do) <http://www.rcsb.org/pdb/Welcome.do>), UniProt (<http://www.pir2.uniprot.org/>), and the Comprehensive Microbial Resource database (<http://www.tigr.org/CMR>).

II.1.2 Blast-Extend-Repraze (BER)

The BLAST-Extend-Repraze (BER) search algorithm ([HYPERLINK "http://ber.sourceforge.net"](http://ber.sourceforge.net) <http://ber.sourceforge.net>) initially runs a BLAST search (Altschul, et al., 1990) for each protein against NIAA and stores all significant matches in a mini-database. The nucleotide sequence of each gene is then extended 300nt upstream and downstream, and a modified Smith Waterman alignment (Smith et al., 1981) is performed against the mini-database. The extension of the sequence allows the resulting alignments to be evaluated for frameshift mutation or point mutations that introduce in-frame stop codons. If significant homology to a match protein exists and extends into a different frame from that predicted, or extends through a stop codon, the program continues the alignment past the boundaries of the predicted coding region.

II.2 Protein Families

II.2.1 Hidden Markov Models (HMMs)

Protein translations of each gene model are searched against Hidden Markov models (HMMs) using the HMMer package (Eddy, 1998) Two libraries of HMMs are used: the Pfam HMMs (Bateman, et al., 2000), and TIGRFAMs (Haft, et al., 2001).

II.3 Sequence signatures

Other amino acid sequence signatures, domains, or functional sites are predicted by searching all proteins against the PROSITE database (Falquet et al., 2002). The SignalP (Bendtsen et al., 2004) and TMHMM (Krogh et al., 2001) algorithms are

used to predict putative signal sequences and membrane spanning domains respectively.

II.4 Gene Naming

AutoAnnotate is a programmatic approach to assigning descriptive functional annotation to gene models following JCVI naming convention guidelines in an automated fashion. It uses a heuristic approach to evaluate results of homology searches. By analyzing the BER and HMM search results, AutoAnnotate assigns a common name, gene symbol, Enzyme Commission (EC) number, and JCVI functional role categories and Gene Ontology terms (GO) as follows.

II.4.1

AutoAnnotate first evaluates the isology and threshold score of each HMM match. If there is a hit to an equivalog-level HMM with a threshold score above the trusted cutoff score, the identifying information attached to that HMM (protein name, role category, gene symbol, GO terms and EC number if applicable) is assigned to the gene model.

II.4.2

If there are no matches to an equivalog-level HMM with above the trusted cutoff score, the BER search results are evaluated. The program follows a specified ranking system to evaluate matches starting with the criteria to include a full-length match of at least 80% of the length of the subject and 80% identity. If more than one match is found, the program assigns highest rank to a Characterized protein from CHAR (**CHAR**acterized Protein Database). BER matches with varying percent identities over the length of the protein are ranked and prioritized in the following order: accessions from CHAR, Uniprot accessions and OMNI accessions - for annotation originating at JCVI. Interleaved into this ranking system are non-equivalog TIGRFAM and PFAM HMM hits to include subfamily, superfamily, domain and repeat HMM models. The specificity of final protein name is fine-tuned to reflect the evidence data type used to assign the protein name.

II.4.3

Proteins with a pair-wise match to a hypothetical protein from another species, but no HMM hit, are named conserved hypothetical protein.

II.4.4

Proteins with no HMM or BER matches remain named hypothetical protein.

II.4.5

Hypothetical proteins with predicted lipoprotein signatures are names 'putative, lipoprotein'. Hypothetical proteins with five or more predicted membrane spanning

regions are names 'putative membrane protein'.