

TITLE: **Homology Searches**

PAGE: 1 of 2

SOP #: PA02

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:
Lauren Brinkac

PRIMARY REVIEWER:
Bill Nelson

1 OVERVIEW

Many different types of data are collected for each gene model to facilitate functional annotation.

1.1 Scope

This analysis is performed on all protein translations of every gene model identified by the Glimmer gene finding system prior to the functional annotation process.

1.2 Related Documents

SOP PA01 - Gene Prediction

- Altschul S., et al. Basic local alignment search tool. *J. Mol. Biol.*, 215: 403-410 (1990).
- Bateman A., et al. The Pfam protein families database. *Nucleic Acids Res.* 28(1): 263-266 (2000).
- Bendtsen, J.D., et al. Improved prediction of signal peptides: SignalP 3.0 *J. Mol. Biol.*, 340:783-795, (2004).
- Eddy S.R. Profile hidden Markov models. *Bioinformatics*, 14(9):755-763 (1998).
- Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research* 32(5), 1792-97 (2004).
- Falquet L., et al. The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30(1):235-8 (2002).
- Haft D., et al. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29(1): 41-3 (2001).
- Haft D., et al. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics.* 21(3):293-306 (2005).
- Krogh A., et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305(3):567-80 (2001).
- Smith T.F., et al. Identification of common molecular subsequences. *J Mol Biol.* 147(1):195-7 (1981).
- Tatusov R.L., et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 4:41 (2003).

1.3 Revision History

Author	Date	Change
Bill Nelson	14 July 2006	Basic edits and addition of SOP#

TITLE: **Homology Searches**

SOP #: PA02

REVISION LEVEL: .2

PAGE: 2 of 2

2 REQUIREMENTS

A genome sequence with gene model predictions.

3 PROCEDURE

The translation of each gene model identified by Glimmer 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.

3.1 Pairwise alignments

3.1.1 Non-identical amino acid database (NIAA)

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

3.1.2 Blast-Extend-Repraze (BER)

The search algorithm used for these searches is BLAST-Extend-Repraze (BER) (<http://ber.sourceforge.net>). This program first runs a BLAST search (Altschul, et al., 1990) of 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.

3.2 Multiple alignments

Multiple alignments of the top twenty-five scoring praze hits are calculated using MUSCLE (Edgar, 2004).

3.3 Family prediction

3.3.1 Hidden Markov Models (HMMs)

TITLE: **Homology Searches**

SOP #: PA02

REVISION LEVEL: .2

PAGE: 3 of 2

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).

3.3.2 NCBI clusters of orthologous genes (COG)

Family prediction is done by searching (BLAST) all predicted proteins against version 2 of the NCBI clusters of orthologous genes (COG) database (Tatusov et al., 2003). Domain-based paralogous families are built on the basis of related HMM hits and homologous regions (detected using BLAST) not covered by the HMM models.

3.4 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.

3.5 Genome Properties

To predict metabolic pathways, systems, and protein complexes, all proteins are searched against the Genome Properties system (Haft et al., 2005), and the predicted attributes generated are stored for evaluation.

4 DATA MANAGEMENT

4.1 Quality Control

All proteins are submitted to nightly updates. Searches are regenerated for any gene models that were edited, or when NIAA, HMM libraries, and/or the Genome Properties system is updated.