

TITLE: **Functional Automated Annotation**

PAGE: 1 of 2

SOP #: PA05

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:
Lauren Brinkac

PRIMARY REVIEWER:
Bill Nelson

1 OVERVIEW

To speed the annotation process, an initial automated parsing of the data collected for each gene model is performed. Initial annotation is assigned and the genes are functionally classified, allowing curators to approach the dataset in a logical manner.

1.1 Scope

This analysis is performed on all gene models.

1.2 Related Documents

SOP PA01 - Gene Prediction

SOP PA02 - Homology Searches

[Naming Convention Guidelines](#)

1.3 Revision History

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

2 REQUIREMENTS

A genome sequence with gene model predictions and homology searches run.

3 PROCEDURE

3.1 AutoAnnotate

AutoAnnotate is a programmatic approach to assigning descriptive functional annotation to gene models following naming conventions 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 TIGR and Gene Ontology (GO) role categories as follows.

TITLE: **Functional Automated Annotation**

SOP #: PA05

REVISION LEVEL: .2

PAGE: 2 of 2

- 3.1.1 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 (common name, role category, gene symbol, GO terms and EC number if applicable) is assigned to the gene model.
- 3.1.2 If there is no match to an equivalog-level HMM scoring above the trusted cutoff score, the BER search results are evaluated. The criteria for evaluation includes a full-length match of at least 80% of the length of the subject, with at least 35% identity. If more than one match is found, the program looks first for annotation originating at TIGR (that will, therefore, follow our naming conventions), then for the entry that conveys the most information (ie name and gene symbol instead of just name) and assigns a TIGR role category.
- 3.1.3 If the chosen BER match is a hypothetical protein from another species or if no pair-wise matches meet the match criteria, AutoAnnotate evaluates the HMM results and looks for family isology HMMs. If any hits exist, the protein is assigned a family name based on the HMM name.
- 3.1.4 Proteins with a pair-wise match to a hypothetical protein from another species, but no HMM hit, are named conserved hypothetical protein.
- 3.1.5 Proteins with no HMM or BER matches remained named hypothetical protein.

4 DATA MANAGEMENT

4.1 Quality Control

This is not a rigorous process. It was designed with the assumption that manual curation would follow. All assignments should be considered speculative (although the genes with equivalog HMM hits will have what should be considered high-confidence annotation).