

TITLE: **Curation of Functional Annotation**

PAGE: 1 of 2

SOP #: PA06

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:
Lauren Brinkac

PRIMARY REVIEWER:
Bill Nelson

1 OVERVIEW

The most accurate way to assign descriptive functional information to predicted gene models is through manual curation of each gene model. Although AutoAnnotate is an efficient programmatic approach for predicting functional annotation, manual evaluation is ultimately required to assess those evidence types that cannot be derived computationally, thereby facilitating more precise and consistent prediction of gene function.

1.1 Scope

This analysis is performed on all gene predictions whose protein translations display sequence similarity to other proteins or domains.

1.2 Related Documents

SOP PA01 - Gene Prediction

SOP PA02 - Homology Searches

SOP PA05 - Automated Functional Annotation

[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, homology searches run, and automated annotation.

3 PROCEDURE

Manual curation involves a layered, conservative, evidence-based system. All available evidence for each protein is evaluated manually using the web-based annotation tool Manatee (<http://manatee.sourceforge.net>), to satisfy evidence criteria described in naming. This evidence includes homology search results, genome properties analysis, and genomic context.

TITLE: **Curation of Functional Annotation**

SOP #: PA06

REVISION LEVEL: .2

PAGE: 2 of 2

3.1 Datatypes

3.1.1 Complex datatypes

The more complex datatypes (HMMs, multiple alignments, phylogenetic trees, Genome Properties) are examined first. Highly specific annotation can be obtained from hits to equivalog HMMs and Genome Properties. Protein family relationships can be determined from phylogenetic trees, non-equivalog isology HMMs, and multiple alignments.

3.1.1.1 Evaluation Criteria

Matches to HMMs must score above the trusted score threshold to be considered a positive assertion. Matches with scores between the trusted and noise score thresholds are considered putative assertions.

3.1.2 Supporting datatypes

The simpler datatypes (BER, PROSITE, positional context, transmembrane helix predictions, etc.) are reviewed to verify that they support conclusions drawn from the complex datatypes.

3.2 Pairwise alignments

If specific annotation cannot be determined from the complex data types, it can frequently be obtained from the BER results.

3.2.1 Experimental characterizations

In order to reduce transitive annotation errors, any BER evidence used for functional curation must be from an experimentally characterized protein, whose quality of characterization is determined by manual review of the relevant literature.

3.2.1.1 Evaluation Criteria

Matches to experimentally characterized proteins must be at least 80% of the length of the subject, with at least 35% identity for specific annotation to be assigned.

TITLE: **Curation of Functional Annotation**

SOP #: PA06

REVISION LEVEL: .2

PAGE: 3 of 2

3.3 Additional searches

When deemed appropriate by the curator, additional searches are run against relevant external public databases (eg MEROPS) in order to extract additional layers of functional evidence.

3.4 Specificity uncertain

If specific annotation is not assignable, the gene model will be named based on its protein family membership or domain structure. The family or domain from which the name comes must be published in either a literature publication or public database

3.5 Structural motifs

A gene model that is not a member of a published protein family may be named after a structural motif ('putative membrane protein', 'putative lipoprotein').

3.6 Limited sequence similarity

If there is no other evidence than BER matches to hypothetical proteins from other organisms, the name "conserved hypothetical protein" is applied. In a complete absence of homology data, the model is named "hypothetical protein".

4 DATA MANAGEMENT

Gene models are marked as complete once functional annotation has been assigned. Functional annotation is reviewed whenever homology searches (PA02) are updated.

4.1 Quality Control

A variety of database queries are run to check annotation against the TIGR Gene Naming Conventions. The Genome Properties system is also capable of identifying both overly aggressive annotation and suboptimal annotation.