

TITLE: **Analysis and Curation of Short Gene Models**

PAGE: 1 of 2

SOP #: PA08

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:
Bill Nelson

PRIMARY REVIEWER:
Bill Nelson

1 OVERVIEW

The accuracy of statistically based gene model prediction programs like glimmer is inversely proportional to the length of the gene called. This leads to false positives in the resulting gene list which can complicate further analyses. By examining other data types that support or conflict with short gene predictions, false positives can be minimized.

1.1 Scope

This analysis is performed on all gene predictions shorter than 250 nucleotides having no homology to any experimentally identified gene.

1.2 Related Documents

SOP PA01 – Gene Prediction

SOP PA02 – Homology Searches

1.3 Revision History

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

2 REQUIREMENTS

A genome sequence.

3 PROCEDURE

3.1 Evaluation Criteria

Short gene models are evaluated for:

- data supporting maintenance of the gene model prediction
- data supporting deletion of the gene model prediction
- insufficient supporting evidence

3.1.1 Data supporting maintenance of the gene model prediction

Data supporting maintenance includes any significant similarity to known proteins, HMM hits that have scores above the noise threshold or significant fragment HMM hits, the presence of motifs or structural domains, presence of a cononical ribosome binding site, and/or membership in a paralogous family. In addition, if a prediction is

TITLE: **Analysis and Curation of Short Gene Models**

SOP #: PA08

REVISION LEVEL: .2

PAGE: 2 of 2

located within a putative operon and on the same strand as the other genes, that can be considered evidence for maintaining the prediction.

3.1.2 Data supporting deletion of the gene model prediction

Data supporting deletion includes significant overlap with a known gene (coding or non-coding), abnormal composition of the predicted protein sequence (eg high proline or arginine content in high-GC organisms), and/or location within an operon and on the opposite strand from all the other genes. Predictions meeting these criteria are deleted from the gene set

3.1.3 Insufficient supporting evidence

Many genes do not have enough evidence to make a decision one way or the other. Those predictions are maintained.

4 DATA MANAGEMENT

4.1 *Quality Control*

The average gene length is checked for the organism. In most prokaryotes, average gene length is ~900nt. An excess of short gene predictions can lower that value.