

TITLE: **Start Site Curation**

PAGE: 1 of 2

SOP #: PA03

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:
Ramana Madupu

PRIMARY REVIEWER:
Bill Nelson

1 OVERVIEW

Gene prediction packages frequently have trouble correctly predicting the initiator codon of the genes found. Genes called too short means data is missing that could potentially help identify the protein; genes called to long may lead to spurious, misleading alignments. We review and correct the predicted gene models to improve the quality of the dataset.

1.1 Scope

All predicted coding genes are analyzed.

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

3.1.1 Homology

The best data to examine when trying to identify the start codon is a multiple alignment of homologs. It is important to have a broad phylogenetic representation in the multiple alignment, as a good alignment between homologs in closely related species is likely even in extra-genic regions. If the gene model appears too long or too short in the multiple alignment, the open reading frame model within is examined for alternate start codons that better match the homologs.

TITLE: **Start Site Curation**

SOP #: PA03

REVISION LEVEL: .2

PAGE: 2 of 2

3.1.2 Gene Context

A gene model's position relative to surrounding genes is examined to see if overlaps exist. Resolving such collisions (PA04) may be critical in identifying the proper start.

3.1.3 3rd Position GC Skew

In GC-rich organisms, the %GC of the 3rd position of the codons is a strong indicator of coding sequence, and can be useful in evaluating gene models.

3.1.4 Ribosomal Binding Sites

Upstream sequence is examined for potential ribosome binding sites, and the start codon itself is taken into account (ATG>>GTG>>TTG). Occasionally, the data suggests an alternative start codon is being used (eg, CTT or ATT).

3.1.5 Other

In cases where it is not possible to be highly confident that a specific start codon is correct, the most upstream start codon supported by the data is chosen.

4 DATA MANAGEMENT

4.1 *Quality Control*

Underlying chromatograms are checked for those gene models with alternative start codons to ensure a sequencing error has not occurred. Pair-wise and multiple sequence alignments are regenerated in cases where the start site has been edited, and the updated alignments are reviewed. Overlap analysis (PA04) is run.