

TITLE: **Gene Prediction**

PAGE: 1 of 2

SOP #: PA01

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:  
Sean Daugherty

PRIMARY REVIEWER:  
Bill Nelson

## 1 OVERVIEW

The analysis and annotation of a closed bacterial genome requires the location of coding and non-coding genes on the genome sequence.

### 1.1 Scope

These analyses are performed on all genomic sequences going through the TIGR annotation pipeline.

### 1.2 Related Documents

- Altschul S., et al. Basic local alignment search tool. *J. Mol. Biol.*, 215: 403-410 (1990).
- Delcher A.L., et al. Improved Microbial Gene Identification with Glimmer. *Nucleic Acids Res.*, 27(23): 4636-4641 (1999).
- Lowe T.M., et al. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25(5):955-64 (1997).
- Salzberg S., et al. Microbial Gene Identification using Interpolated Markov Models. *Nucleic Acids Res.*, 26(2): 544-548 (1998).
- Sam Griffiths-Jones, et al. Rfam: an RNA family database. *Nucleic Acids Res.* 2003 33(1):439-441.
- Sam Griffiths-Jones, et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005 33:D121-D124.

SOP PA03 – Start Site Curation

SOP PA04 – Overlap Analysis

### 1.3 Revision History

Author	Date	Change
Bill Nelson	22 Aug 2006	Basic edits and addition of SOP#

## 2 REQUIREMENTS

Genome sequence data.

TITLE: **Gene Prediction**

SOP #: PA01

REVISION LEVEL: .2

PAGE: 2 of 2

### 3 PROCEDURE

The genomic sequence data is first loaded into the database, either as a closed and edited molecule or as a pseudomolecule. A pseudomolecule is created when an unclosed molecule is entered into the TIGR annotation pipeline, due to the fact that Glimmer (Salzberg et al., 1998; Delcher et al., 1999) handles whole genome sequence better than shorter contigs. In the locations the pseudomolecule is stitched together the sequence “NNNNNCACACTTAATTAATTAAGTGTGTGNNNNN” is used, which places start and stop codons in all 6 reading frames. After the sequence is loaded into the database all RNAs are found by the use of RFAM (Sam Griffiths-Jones, et al, 2003,2005) BLAST (Altschul S., et al, 1990) and tRNA\_scan-SE (Lowe T.M., et al, 1997). All coding regions are identified using version 3.01 of the Glimmer gene finding System. Glimmer identifies the potential coding regions by using all of the long open reading frames found in the genome to build a training set. Then it uses the results from the initial Glimmer run as a training set for the second run. The identified coding sequences are then loaded into the database.

### 4 DATA MANAGEMENT

#### 4.1 *Quality Control*

- The number of sequences identified and loaded into the database should match approximately one gene per 900 nucleotides of genome sequence.
- An intergenic region analysis is run to see if any potential coding regions were missed by glimmer3. All regions that do not have an identified coding sequence or other genomic feature are run against a non-redundant amino acid database using BLAST. Then all regions are manually reviewed. This is done because the algorithm used by Glimmer3 has two limitations: it does a poor job of identifying short genes and genes in regions of an atypical nucleotide composition.
- After all coding sequences are identified, all start sites are curated (SOP PA03) and an overlap analysis (SOP PA04) is run.