Prokaryotic Annotation Pipeline
Broad Institute

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

---

1. Abstract

N/A

2. Introduction

This SOP describes the Prokaryotic Annotation Pipeline at the **Broad Institute**. This pipeline identifies protein coding sequences and RNA genes in complete and draft bacterial genome sequences, and assigns gene product names based on homology to known sequences.

Broad Institute Public Website:

http://www.broadinstitute.org/seq/msc

3. Requirements

3.1. Data requirements

A draft or complete bacterial genome sequence file in FASTA file format. The FASTA file can contain more than one FASTA sequence.

3.2. Software requirements

Numerous software packages and data sets are required as referenced in the Procedure section.

3.3. Computer requirements

This protocol requires use of a machine capable of running linux or another unix-compatible

Operating system.

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

4. Procedure

4.1. Fasta file formatting

The sequence file used for data generation from the pipeline must first be formatted correctly, by running though a program that confirms or corrects it to the FASTA specifications. If the input file is a multifasta file, it is then split into individual single fasta files. This will result in separate annotated molecules, necessary when more than one molecule comprises the overall genome of the organism.

4.2. RNA prediction

We apply the following RNA prediction tools to the entire genome sequence:

- tRNAscan
- RNAmmer
- RFAM

Together these tools provide highly accurate predictions for tRNAs, rRNAs, and other non-coding RNA genes.

4.2.1. tRNAscan

The tRNAscan software is used for detecting tRNAs on the genome assembly. Prokaryotic model of tRNA scan is run on the entire genome sequence using the default parameters.

4.2.2. Rfam and RNAmmer

Rfam and RNAmmer predict common RNA features such as ribosomal RNA, small regulatory non-coding RNA, non coding RNA. Besides representing these useful biological entities on the genome, the presence and organization of the features provide contextual information between RNAs and protein coding genes and further aid in the removal of spurious protein coding predictions. rRNA operons or clusters, if and when represented fully, are good indicators of the completeness of the genome assembly.

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

Infernal is used to search the RFAM library against the genome sequence (expect Value Cutoff of 0.01). RNAmmer is run using the default options for bacterial species

4.3. Blast

Blast homology search against the Genbank's NR database produces a set of protein-genome alignments. Individual blastx alignments are then clustered into single blast clusters by linking the blast alignments derived from the same blast hit. Several such overlapping blast clusters on the genomic axis represents what we call as blast loci on the genome assembly.

Current Blast parameters used by the different centers are below:

	Details	**Gene product naming**
Blast Database	NR (bacteria)	NR (bacteria)
Max E value	10^{-10}	10^{-10}
min identity	30%	30%
min query coverage	30%	30%

30% query coverage cut-off represents the minimum cut-off for coverage; however, majority of the evidence is 70% or more. Considering the draft nature of the assemblies used for annotation, using blast evidence with at least 30% or less query coverage is important for identifying genes that span contigs and/or touch gaps.

4.4. Hmmer

Broad's annotation pipeline runs Hmmer searches using Pfam library and TIGRfam library to find domain homologies on six-frame translations of the genomic sequence (expect Value Cutoff of 0.1)

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

4.5. Gene Prediction Programs

Potential protein coding genes on draft genome assemblies were predicted with the following programs:

- GENEMARK
- GLIMMER
- METAGENE

Gene finding programs use slightly different algorithmic and heuristic approaches for finding potential coding genes. This is obvious in terms of the observed differences in the gene count and their predicted structure by different programs.

Glimmer3	--gene_len 90 (min gene length) --trans_table 11 (codon table for bacteria) --linear (no gene wrap around at sequence ends)
GeneMark	Genome Specific parameter file
MetaGene	Default

4.6. Gene Calling

4.6.1. Selection of Consensus Gene Models

Broad's automated gene calling process uses a rule-based selection process to evaluate the evidence and build consensus gene models.

Ab initio predictions and blast are clustered into potential gene loci. We select the most likely non-conflicting gene models based on the best evidence available at each locus. Our method uses heuristics such as relative overlap with the BLAST hits to choose the prediction most in accord with the evidence. It does not have an internal model of gene structure and thus runs on a wide variety of eukaryotic and prokaryotic organisms without training.

Author: Qiandong Zeng

Version: 1
Effective Date: July 2009

4.6.2. Length cutoff of genes and gene-gene overlaps

Min Gene Length Without Evidence	**120 bases**
Min Gene Length With Evidence	**60 bases**

Maximum overlap allowed	**200 bases**

4.6.3. Exclusion of open reading frames with overlaps to non-coding features

Our gene selection process offers several useful options to exclude calling genes at certain loci. For example, we can exclude genes in regions with tRNA and rRNA using a conservative overlap criterion.

4.6.4. Open reading frames with problems

Despite all the progress in the field of gene finding, accurate gene finding on draft genomes is still a challenge. We make an effort to track easily identifiable problematic gene models and tag them with appropriate curation flags to alert the users of the nature of the problems. These tags are also used by manual annotators to specifically target manual editing and fine-tuning of poor gene models.

4.6.5. Conflict resolution in gene calls

In case of overlapping predictions with different ORF lengths, blast evidence serves as a reference data point for picking the best gene models. In addition, blast evidence is also used for retaining overlaps between two adjacent genes. Blast features are also used to create Blast extended features which are very useful for finding genes missed by commonly used ab initio predictors.

The following table lists some of the rules used to resolve overlapping predictions:

1. If both are predicted only, keep the longest ORF
2. If both contain pFAM, keep both
3. If one has pFAM & blast and other doesn't, keep the one with pFAM hit
4. If one has pFAM and other low confidence blast, keep the one with the pFAm domain

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

5. If both have blast and pFAM, keep both (lesser overlapping in silico prediction is chosen)
6. ORFs *within* ORFs -same or different strand and ORFs with the same reading frames are never allowed, even if both have evidence- choose the longest ORF in this case.

4.6.6. Detection and tagging of ORFs with frame shift

ORFs with one or more frame shifts are referred to as disrupted ORFs (dORFs). Disruption in an ORF may be caused by sequence errors or degeneration of the coding sequence leading to creation of pseudogenes. On finished genomes, these dORFs are referred to as pseudogenes. However on the draft genomes, one can not easily characterize them as pseudogenes as they could be a result of common sequence problems and gaps in the assembled sequence.

Disrupted ORFs can only be detected when they have blast evidence with indications of frame shift. At such loci, ab initio predictions will either split genes into two or more reading frames or predict a single ORF that is significantly shorter in length as compared to the evidence. Despite the fact that dORFs are common among bacteria, what makes the detection of these dORFs on draft genomes even more difficult is that not all split genes with blast evidence are dORFs: Some of them represent *real* splits. According to the Rosetta-stone hypothesis two or more functionally related genes can occur as either single ORFs or two or more smaller ORFs, each representing a functional ORF.

The consensus among the genome centers is that we should tag the easily identifiable defective ORFs with the curation flag ‘contains frame shift’ to indicate the presence of the frame shift. Blast extended ORFs or genewise predictions are the common predictions capable of identifying dORFs.

In general, the following issues in the blast extended ORFs indicate the presence of dORFs:

- A single blast loci with two *ab initio* predictions in which each prediction corresponds to a part of a single blast alignment.
- Two or more blast alignments in different reading frames.
- Only a fraction of the ORF is recognizable as compared to the blast query sequence.

Author: Qiandong Zeng
Version: 1
Effective Date: July 2009

5. Implementation

N/A

6. Related Documents & References

N/A

7. Revision History

Version	Author/Reviewer	Date	Change Made
1.01	Teena/Brian/Qian	7/10/2009	Establish SOP