Prokaryotic Annotation Pipeline SOP
HGSC, Baylor College of Medicine

Author: Xiang Qin
Version: 1.01
Effective Date: July 2009

1 Abstract

A prokaryotic annotation pipeline was developed to automatically annotate draft and complete bacterial genomes. The protein coding genes in the genomes are predicted by the combination of Glimmer and GeneMark, while tRNA, rRNA and other non-coding RNAs are predicted by tRNAScan, RNAmmer and Rfam, respectively. Conflict of gene predictions at each locus is resolved by empirical rules described in this SOPs. Gene product names are assigned based on sequence homology to known genes in the public databases. Functions of protein coding genes are also assigned according Pfam and COG and KEGG categories.

2 Introduction

This SOP describes the prokaryotic annotation pipeline at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine. This pipeline identifies protein coding sequences and RNA genes in complete and draft bacterial genome sequences, and assigns gene product names and functional annotation based on sequence homology to known genes. The main user of this pipeline is the automatic annotation pipeline for HMP bacterial reference genomes at the HGSC at Baylor College of Medicine.

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 Compute requirements

This protocol requires a Linux or Unix operating system to execute. Linux or Unix compute clusters are also required.

4 Procedure

4.1 Fasta file formatting

The DNA 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 and a scaffold file in AGP format (see: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml) is provided, the FASTA sequences will be linearized according to the scaffold file. If the input file is a multifasta file but no scaffold file is provided, the FASTA sequences will be concatenated into one

pseudo FASTA file with 200 Ns between the FASTA sequences. Sequence with fewer than 500 bp will not be annotated in the current version v1.10.

4.2 RNA prediction

We apply the following three RNA prediction tools to the entire genome sequence: tRNAScan, RNAmmer, and RFAM/infernal. Together these tools provide highly accurate predictions for tRNAs, rRNAs, and other non-coding RNA genes.

The software version and parameters used for RNA predictions:

Software	Version	Parameter	Prediction
tRNAScan	1.11	default parameters, -P option for prokaryotic sequences	tRNA
RNAmmer	1.2	default parameters	rRNA
Rfam	infernal-0.7	default parameters	non-coding RNA

The programs can be downloaded from the following websites:
ftp://selab.janelia.org/pub/software/tRNAscan-SE/
http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?rnammer
http://rfam.sanger.ac.uk/help?tab=helpFtpBlock

4.3 Gene prediction programs

Glimmer and GeneMark are used for *ab initio* protein coding gene prediction. The software version and parameters used are as the following:

Software	Version	Parameter	Prediction
Glimmer	3.01	--gene_len=90 --max_olap=200 --codonTable=11 --linear	*ab initio* protein coding gene prediction
GeneMark	2.4	default parameters Use closest related genome as model or use heuristic model	*ab initio* protein coding gene prediction

The programs can be downloaded from the following websites:
http://www.cbcb.umd.edu/software/glimmer/
http://opal.biology.gatech.edu/GeneMark/

4.4 BLAST

BLAST is used for sequence similarity search of known genes using the NCBI NR and Pfam databases. The BLAST results are used as evidence in the small ORF filtering process and in resolving conflicts involving multiple gene predictions at one locus. BLAST is also used to find

missing genes in *ab initia* gene predictions, and to name gene products and assign functional annotations to protein coding genes.

BLAST parameters used in the pipeline:

Program	Parameter	Database
BLASTx	-F F -e 1e-10	NR
BLASTp	-F F -e 1e-5	Pfam

BLAST programs and the NCBI NR and Pfam databases can be downloaded from these websites:
ftp://ftp.ncbi.nih.gov/blast/executables/release/
ftp://ftp.ncbi.nih.gov/blast/db/
http://pfam.janelia.org/

4.5 Gene calling

The best gene call at each locus is determined by the following empirical rules:

4.51 Minimum length cutoff for protein coding ORFs

ORF with BLAST evidence	60 bp
ORF without BLAST evidence	120 bp

4.52 Protein coding ORFs in the same reading frame at the same locus

- ORFs in the same reading frame with different length at the same locus are resolved based on the best BLAST evidence.
- When there is not BLAST evidence for both ORFs, the longest ORF will be chosen.
- If all the BLAST evidence has the same best score, the longest ORF will be kept.

4.53 Overlaps between protein coding ORFs and functional RNAs

- ORFs within functional RNAs are excluded.
- ORFs without evidence that contain functional RNAs are excluded.
- ORFs with evidence that overlap functional RNAs by less than 30% of the ORF length are allowed.
- ORFs with evidence that overlap functional RNAs by more than 30% of the ORF length are excluded.

4.54 Overlaps between two protein coding ORFs

- If both ORFs do not have BLAST evidence, keep the longest ORF.
- If only one ORF has BLAST evidence, keep the ORF with evidence.
- If both ORFs have BLAST evidences and the overlap is less than 30% of the length of the shorter ORF and less than 200 bp, keep both ORFs. Otherwise, keep the longer ORF.
- An ORF within another ORF will be excluded.

4.55 Pseudogene and frameshift detection

This version does not have automatic frame shift or pseudogene detection. This function will be added in the future.

Author: Xiang Qin
Version: 1.01
Effective Date: July 2009

4.6 Gene product naming and other features

4.61 Gene product naming

Gene products are named based on BLAST matches to known genes in HGSC curated microbial database or NR database using a 30% identity and 60 % subject length cutoff. ORFs with matches to known genes below the cutoffs will be named with "possible". ORFs with matches to hypothetical proteins will be named as "conserved hypothetical protein".

4.62 EC number and gene name assignment

EC number and gene name assignments are based on the BLAST matches to NR database and Swissprot Enzyme nomenclature database (http://ca.expasy.org/enzyme/)

4.63 Controlled vocabulary

At the HGSC we currently use in-house programs to check gene product names in order to make the gene product names as consistent as possible.

4.7 Gene functional annotations

Program	Parameter cut-off	Database	Purpose
HMMER	-e 1e-5	Pfam database	Pfam/InterProScan
BLASTp	-e 1e-10 25% identity 50% subject length	COG database March 2003	COG number
BLASTp	-e 1e-5 30 % identity 50% subject length	KEGG database Feb. 2009	KEGG number
InterProScan	default parameters	Pfam database	InterPro number

The programs and databases can be downloaded from the following websites:
http://hmmer.janelia.org/
http://pfam.janelia.org/
http://www.genome.jp/kegg/
http://www.ebi.ac.uk/Tools/InterProScan/
ftp://ftp.ncbi.nih.gov/pub/COG/COG/

4 Implementation

This is a SIGS requirement; to be included in HMP SOPs when applicable.

5 Discussion

This is a SIGS requirement, but optional for HMP SOP submission.

6 Related Documents & References

This is a SIGS requirement; to be included in HMP SOPs when applicable

SIGS Guideline:

- Software tools and data sets should be referenced to a publication (or alternatively a project web site) whenever possible.
- Technical references that are internal to an institution and not public should be excluded. Examples include intranet URLs, file system paths, and computer server names that are not publicly accessible on the Internet.

7 Revision History

This is an HMP_specific requirement, not included in the SIGS submission. Please be sure to update this when any changes as made, to help the DACC organize SOPs.

Version	Author/Reviewer	Date	Change Made
1.01	Xiang Qin, Kim Worley, Lan Zhang	7/10/2009	Establish SOP