Prokaryotic Genome Annotation Pipeline

Washington University Genome Center (WUGC)
(http://genome.wustl.edu)

Author: Kym Pepin

Version: 1.01

Effective Date: 06/15/2009

1. Abstract

None.

2. Introduction

This is the Washington University Genome Center (WUGC) (http://genome.wustl.edu) automated annotation pipeline standard operating procedure (SOP) detailing the processing of microbial genomic sequences through the automated annotation pipeline, composed of three parts: 1. gene prediction, 2. gene merging and 3. protein annotation (see section 4 for details).

3. Requirements

Software Requirements:

Gene Prediction: A combination of ab initio gene predictors, GeneMark (1) and Glimmer3 (2), are used to define gene coding regions followed by the generation of homology based gene predictions using blastx in regions not covered by the ab initio gene predictors.

Non-coding predictions: tRNAscan (3), Rfam (4) and RNAmmer (5) are used to predict transfer RNA's (tRNA), ribosomal RNA's (rRNA) and non-coding RNA's (ncRNA) respectively.

Protein Annotation: All final gene predictions are processed through an automated protein annotation pipeline which includes psort-b (6), pfam (7) KEGG (8) and BlastP against NR bacteria.

More details on use and run parameters can be found in the procedures section of this SOP.

Data Requirements:

Prokaryotic Genome Annotation Pipeline

Washington University Genome Center (WUGC)
(http://genome.wustl.edu)

Author: Kym Pepin

Version: 1.01

Effective Date: 06/15/2009

Input files to the pipeline are in fasta format. All processing is run on contigs. A bacterial NR database is required for BlastX gene creation, evidence collection and functional assignment.

Compute requirements:

The pipeline is designed in a workflow system and requires a compute cluster for processing. Data storage is in Oracle.

4. Procedure

1. Gene set generation

   Phase I gene set generation: Genomic DNA sequence in fasta format is processed through the following programs Rfam, RNAmmer, tRNAscan, Glimmer3 and GeneMark. Rfam is used for detecting ncrna's and RNAmmer for the detection of rRNA's, with tRNAscan predicting tRNA's in the sequence. All programs are used with the default settings for detection in prokaryotic sequences. Glimmer3 and GeneMark detect coding sequences. Glimmer3 is processed using overlap=200, minLength=90, codonTable=11 linear, long-orfs and genome specific icm file (build icm) for genome specific model file generation. For GeneMark we use a genome specific model file when available or the GC based heuristic model file which is chosen based on the genomes GC content.

2. Gene merging

   A. Once the raw ab initio gene predictions are generated the process clusters predictions with the same reading frame and stop codon and chooses the best gene for that locus based on gene prediction hierarchy of GeneMark then Glimmer3. If two gene predictions share the same loci, but have a different stop codon or are on opposite strands, no matter what the overlap, the gene is kept at this stage. Final selection is based on evidence in the order of Pfam, blast vs bacterial nr or longest open reading frame (ORF).

B. Phase II gene set generation: When two protein-coding genes with different reading frames overlap by 200 bases or 30% of their ORF length or are < 120 bases in length (no overlap required), these genes are processed through evidence collection for final gene selection. Cutoffs for evidence include

1. blastp: 130 bit score or E=10-6
2. pfam: default

Rules for resolving overlapping gene predictions is as follows:

1. Keep genes >= 60 bases and have evidence, discard remaining
2. If both are predicted only (no evidence), keep the longest ORF
3. If both genes have hits to pfam, keep both
4. If both genes have no hit to pfam and/or blastp, keep both
5. If one has pFAM and the other has blast or no evidence, keep the pfam gene
6. If only one has a hit to pfam and/or blastp, keep the gene with evidence
7. If one gene has blast and the other no evidence, keep the blast evidence gene

C. Intergenic region masking and Blastx gene creation: Mask the genomic sequence using the genes which remain after overlap evidence and length checks (Section B), retracting 300 bases on each end of the gene. Run blastx vs the remaining sequence using the top non-hypothetical blast hit at 30% PID/30% Cov to generate gene models missed by ab initio predictors.

D. Choose the best genes from the population of ab initio and evidence-based gene models by repeating the overlap and evidence checks between the phase 2 gene set and new blastx gene set following the same criteria listed above (Section B).

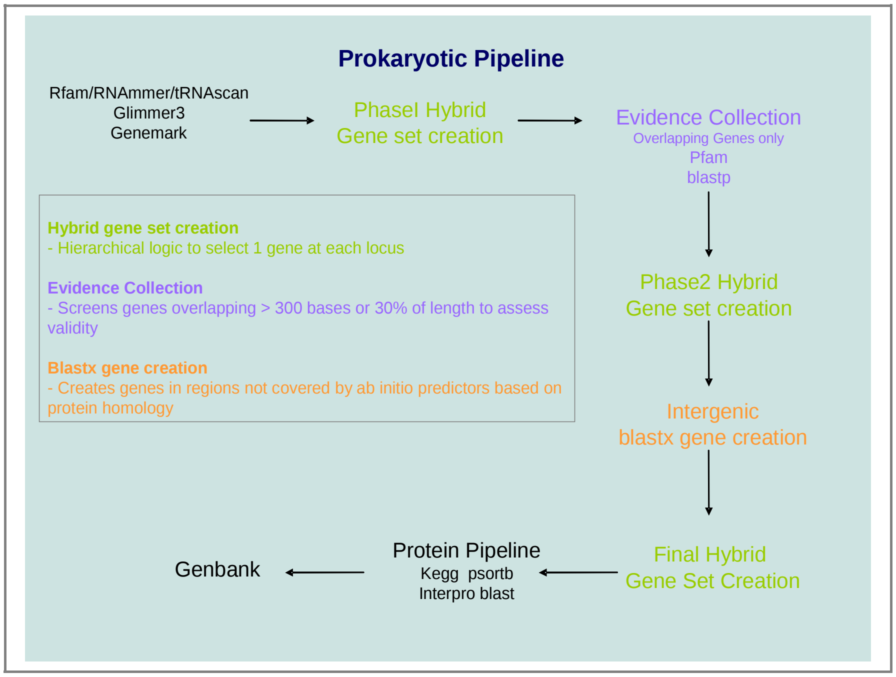
E. Resolving Overlaps between RNA and coding predictions: Criteria for resolving overlaps between adjacent coding genes and non-coding features such as tRNAs and rRNAs are listed below. All genes overlapping rRNA genes are discarded

1. No overlap allowed on the same strand with tRNA
2. Overlap allowed on opposite strand with tRNA if CDS has evidence (blast or Pfam)
3. If riboswitches or leaders on the same strand, we allow 50% overlap with a CDS gene
4. All other rna genes can overlap by up to 10% of length or CDS is discarded
5. Currently RyeA, RyeB, glms, Intron_gpII and DnaX can be ignored (overlaps are allowed)
6. All genes overlapping rRNA genes are discarded

3. Protein annotation: All genes in the final gene set are processed through a suite of protein annotation tools. Annotations include assignment of enzymes in KEGG pathways, similarity to protein domains from Pfam and TigrFAM, psort-b for cellular localization and blastp vs bacterial nr (E= 10^{-10}) for gene function. This step is currently in the process of transitioning to the gene naming pipeline from the JCVI (http://ber.sourceforge.net).

5. PIPELINE SUMMARY

Below is a graphical representation of our annotation pipeline workflow, details of which have been described in the above text.



Prokaryotic Pipeline

Rfam/RNAmmer/tRNAscan
Glimmer3
Genemark

PhaseI Hybrid
Gene set creation

Evidence Collection
Overlapping Genes only
Pfam
blastp

Hybrid gene set creation
- Hierarchical logic to select 1 gene at each locus

Evidence Collection
- Screens genes overlapping > 300 bases or 30% of length to assess validity

Blastx gene creation
- Creates genes in regions not covered by ab initio predictors based on protein homology

Phase2 Hybrid
Gene set creation

Intergenic
blastx gene creation

Final Hybrid
Gene Set Creation

Protein Pipeline
Kegg psortb
Interpro blast

Genbank

6. REFERENCES

1. Lukashin A. and Borodovsky M., GeneMark.hmm: new solutions for gene finding, NAR, 1998, Vol. 26, No. 4, pp. 1107-1115.

2. A.L. Delcher, K.A. Bratke, E.C. Powers, and S.L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 2007, 23(6):673-679

3. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence, Nucl. Acids Res. 1997, 25, 955-964.

4. An RNA family database. Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna and Sean R. Eddy. Nucleic Acids Research 2003, 31, 1, 439-441.

5. Lagesen K, Hallin PF,Rodland EA, Staerfeldt HH, Rognes T, and Ussery DW. Consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Research 2007;35(9):3100-8.

6. Jennifer L. Gardy, Cory Spencer, Ke Wang, Martin Ester, Gabor E. Tusnady, Istvan Simon, Sujun Hua, Katalin deFays, Christophe Lambert, Kenta Nakai and Fiona S.L. Brinkman (2003). PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria, Nucleic Acids Research 31(13):3613-17

7. R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, J.S. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer and A. Bateman. The Pfam protein families database, Nucleic Acids Research 2008, Database Issue 36:D281-D288

8. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000, 28, 27-30.

7. Revision History

Version	Author/Reviewer	Date	Change Made
1.01	Kym Pepin	6/15/09	Establish SOP