SOP for Bacterial Genome Annotation at BioHealthBase Version 1

Authors: Shubhada Godbole, Zuoming Deng

Submission Date: January 15th, 2007

Introduction

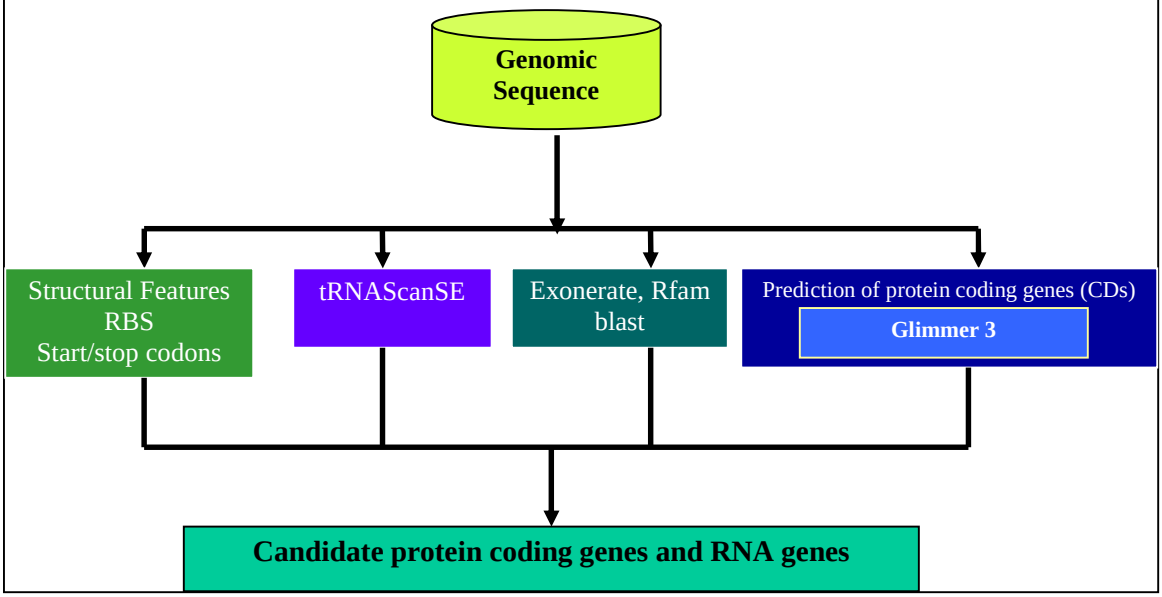
The main goal of a genome sequencing project is to gain knowledge about the cellular processes in an organism. In the context of pathogenic bacteria, genome sequencing projects are focused on understanding the specific mechanisms underlying bacterial survival under extreme stress environments, virulence and pathogenicity factors, resistance to antimicrobial drugs, etc., and exploiting this knowledge in designing vaccines and better drugs for controlling bacterial infection. An essential step in gaining this valuable insight is genome annotation, which primarily involves identifying the genes and predicting their function. Current automated gene prediction tools are based on identification of structural features such as translational start and stop sites, ribosome binding sites as well as sequence homology with known genes. The functional predictions are derived from integrated information from several resources relating to functional domains, experimental evidence, homology based annotation tools, genome features such as operons, analysis of genome structure and evolution, etc.

BioHealthBase Annotation Pipeline for Bacterial Pathogens

Genome annotation includes the prediction of the location of protein coding and non-protein coding genes, predicting the functional identity of the genes and the manual curation of these predictions. TIGR's pipeline for gene prediction (Glimmer3[1]), automated annotation pipeline (AutoAnnotate[2]) and manual curation (Manatee[2]) was used.

Section describing gene location prediction is shown in Figure 1. First step after receiving the genomic sequence is to re-format the sequence so that dnaA is the first gene and then sequence coordinates are reassigned before submitting the sequence for gene predictions. Glimmer3, the gene prediction algorithm is based on Interpolated Markov Model (IMM) and can be trained from the raw sequence alone or with a related, annotated sequence. Glimmer uses the training sequence it to build an IMM which is then used to scan the genome and predict all protein coding genes, with criteria imposed for presence of an initiation codon and length of ORF. The program tRNAscanSE[3] is used to find tRNAs where as a sequence similarity search using Exonerate[4] is used to identify 16S and 23S ribosomal RNA genes. A sequence similarity search against RFam[5], a comprehensive database of non-coding RNA (ncRNA) families is used to identify genes coding for other non-coding RNAs, (such as 5S ribosomal RNAs). Prediction of ribosome binding sites (RBS) is done using RBSfinder[6] algorithm developed by TIGR.

Figure 1: Gene prediction pipeline



Section describing gene function prediction is shown in Figure 2. Each predicted protein is searched against a non-redundant amino acid database (nraa) made up of all proteins available from GenBank, PIR and SWISS-PROT. The search algorithm employed for these searches is BLAST-Extend-Repraze (BER). This program first does a BLAST[7] search of each protein against nraa and stores all significant matches in a mini-database. Then a modified Smith-Waterman alignment[8] is performed on the protein against the mini-database of BLAST hits. In order to identify potential frameshifts or point mutations in the sequence, the gene is extended 300 nucleotides upstream and downstream of the predicted coding region. If significant homology to a match protein exists and extends into a different frame from that predicted, or extends through a stop codon, the program will continue the alignment past the boundaries of the predicted coding region. All of the proteins from the genome sequences are also searched against the Pfam[9] HMMs and TIGRFAMs[10] built from highly curated multiple alignments of proteins thought to share the same function or to be members of the same family. Each HMM has an associated cutoff score above which hits are known to be significant. Additional searches run on the sequence include prediction of transmembrane helices using TMHMM[11], prediction of signal peptide with signalP[12], lipoprotein motif and COG[13] (Clusters of Orthologous Groups of proteins based on phylogenetic classification of proteins encoded in complete genomes) relationships. These data are used by AutoAnnotate to make functional predictions for proteins and are made available in the Manatee interface for manual evaluation of the AutoAnnotate predicted function.

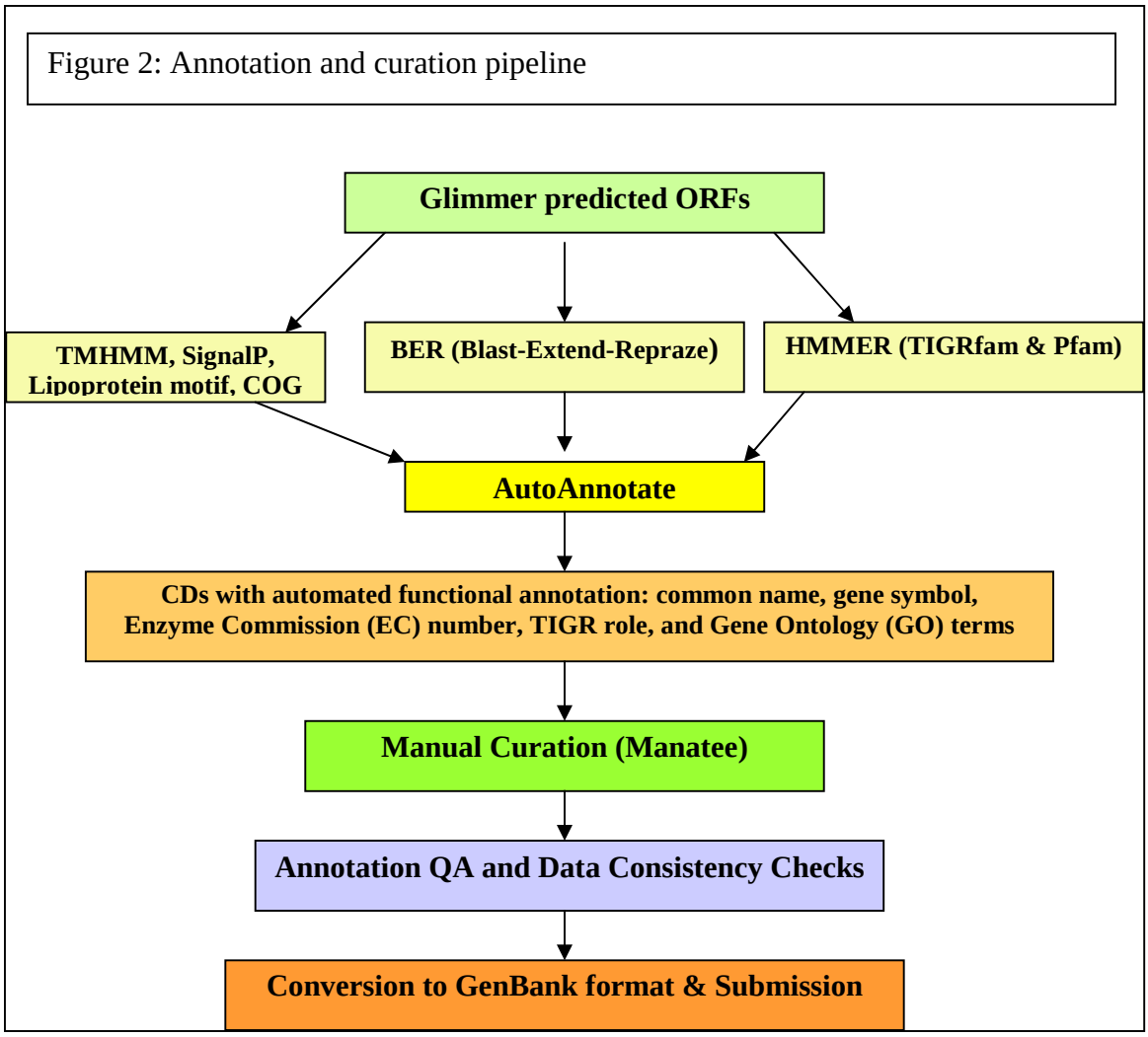
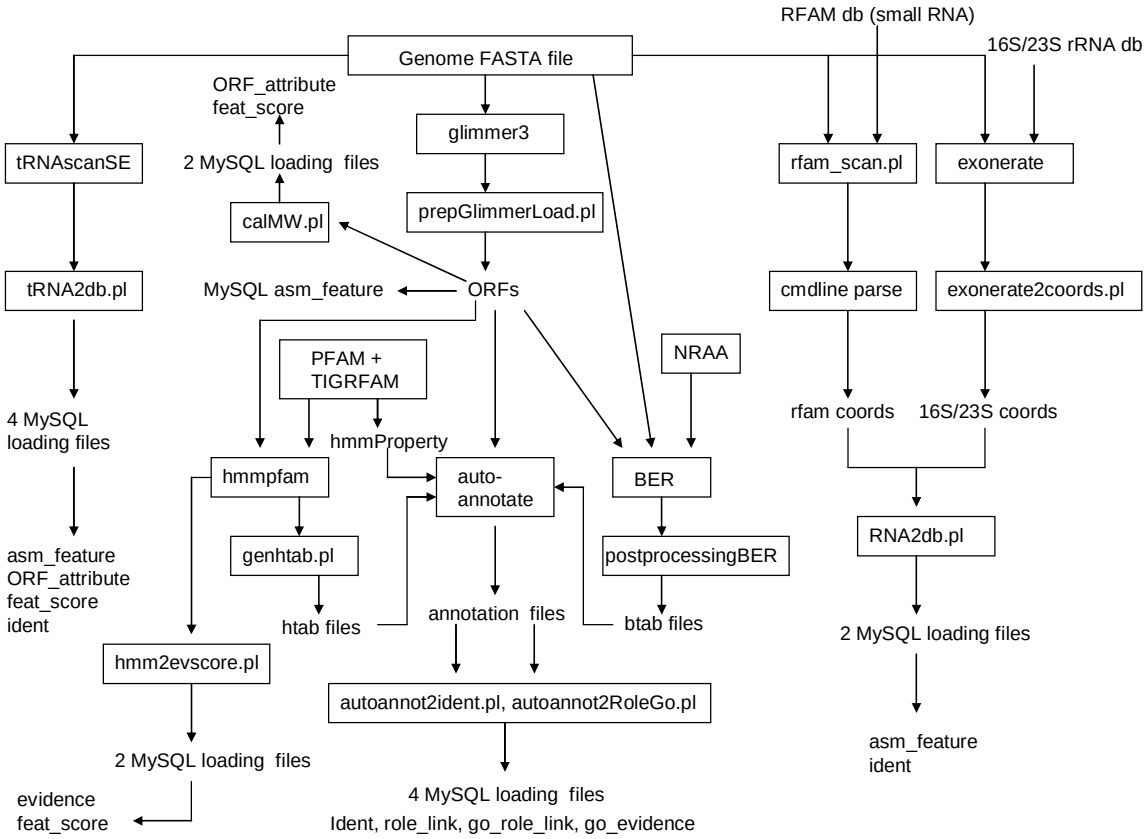


Figure 3 shows the technical details of software tools integration and data transformation and the software versions as wellas the default parameters are listed below:

Glimmer version 3.02: Default parameters were used.
PFAM: release 20.0
TIGRFAM: release 6.0
hmmpfam HMMER: version 2.3.2 parameters: -E 0.1 --cut_ga
tRNAscanSE: version 1.23 parameters: -B
blastp against NRAA (2/8/2006): e-value cutoff is 0.1
rfam_scan.pl version 0.1 parameters: --bt 0.1 --nobig
RFAM release: 7.0
Exonerate: version 1.0.0 parameters: -m NER

.

Figure 3: Software integration and workflow



Manual curation of annotations

TIGR's guidelines for functional assignments, including gene symbols, EC number and GO annotations are followed closely for manual curation with a few modifications to the gene-naming practices. TIGRfam and Pfam HMM alignments along with blast-extend-repraze (BER) alignments are used as primary evidence for functional annotations, with TMHMM, SignalP, COG relationships and literature data used as supporting evidence.

The term 'pseudogene' commonly used for potentially disrupted ORFs can be misleading in the absence of experimental evidence indicating that the gene is indeed non-functional. Recent proteomics data have shown that some ORFs that had been assigned as 'pseudogenes' are expressed and translated. We chose not to use the term 'pseudogene' and instead labeled the genes 'transposon disrupted ORF', 'frameshift disrupted ORF' etc, making it clear that further experimental validation is required to correctly assess expression and function.

Quality checks on annotation and data consistency that are performed prior to submission of the annotated genome to GenBank are listed below:

Annotation Quality Checks

1. Overlapping genes report for both protein coding and RNA genes: TIGR protocol allows overlap of about 10-20 amino acids, for others with larger overlaps, the genes calls were edited using TIGR's guidelines.

   Some overlaps can be resolved by start edits, if the wrong start has been chosen for one or both of the overlaps.

   For example:

   ---------->
   ----------> (perhaps this protein needs a start edit)

   OR

   <----------
   ----------> (perhaps one or both of these need a start edit)

   However, in these case:

   ---------->
   <----------

   ---------->
   <----------

   start edits will not help. So, the other ways we resolve overlaps are:

   a. Look for one of the pair matching a known protein, domain or motif (HMM, another protein, SignalP, TmHMM, Prosite, etc.) while the other does not - then delete the one that does not.
   b. If both have no matches - see which one has a better start and RBS, or see if one fits nicely into an operon and the other does not.
   c. Third position GC skew for GC rich genomes - see which of the overlapping frames has the highest percentage of GC in the third position, this will likely be the real gene. (Due to the constraints of the genetic code the first two positions are less flexible while the third can reflect the high GC nature of the organism.)

   2. ORFs without translation: ORFs other than Transposons with programmed frameshifts/stops were checked for correct translation and fixed by editing the start sites.

   3. ORFs not marked complete: manual curation and 'completion' of all genes was confirmed.

   4. ORFs with *gene_sym* or EC but having 'putative' in gene name: gene_symbols

were removed for ORFs having putative in the gene name, partial EC numbers were allowed if supporting evidence for the functionality is present.

5. For ORFs with *gene_sym* but having 'family protein' in gene name, the gene symbol was removed unless literature evidence supported gene-symbol.

6. ORFs with 'frameshift' or 'stop' or 'hypothetical' in gene name were checked for *gene_sym* or EC or GO assignments, any such assignments were removed but noted down in private comments to be reassigned after verification of frameshift if required.

7. Capitalization was checked for assigned *gene_sym*: standard gene nomenclature protocols were observed where the first three letters are lower case with the last letter, if any, capitalized e.g. dnaA.

Data consistency and integrity checks

1. ORFs that are in evidence but not *asm_feature*.
2. Features that are in *ident* but not *asm_feature*
3. Looked for and edited features that are in *ORF_attribute* but not *asm_feature*.
4. ORFs that are in *ORF_attribute* without corresponding info in *feat_score*.
5. ORFs that are in *evidence* without corresponding info in *feat_score*.
6. Rows in role_link without corresponding info in ident.
7. Features that are in *asm_feature* but not *ident*.
8. ORFs in *asm_feature* without corresponding info in *ORF_attribute*.
9. Rows in *feat_score* without corresponding info in *ORF_attribute* or *evidence*.
10. Intergenic regions were blasted against NCBI database. Any regions having blast hits with significant scores to a full length, functionally annotated protein were manually added to the genome.

References:

1. Delcher *et al*. *Nucleic Acids Research,* **27**(23)**:**4636-4641 (1999).
2. http://manatee.sourceforge.net/
3. Lowe, T. M. & Eddy, S. R. *Nucleic Acids Research*, 25:955-964 (1997).
4. Slater, G. S. & Birney, E. *BMC Bioinformatics*. 15: 6-31 (2005).
5. Griffiths-Jones, S., *et al Nucleic Acids Research*, 33:D121-D141 (2005).
6. http://www.tigr.org/software/genefinding.shtml
7. Altschul S., *et al*. *J. Mol. Biol.*, 215: 403-410 (1990)
8. Smith T.F., *et al*. *J. Mol. Biol*.147(1): 195-197 (1981).
9. Bateman A., *et al*. *Nucleic Acids Res*. 28(1): 263-266 (2000).

10. Haft D., *et al*. *Nucleic Acids Res*. 29(1): 41-3 (2001).
11. Krogh, A., *et al*. *Journal of Molecular Biology*, 305(3):567-580