Metagenome Annotation Standard Operating Procedure for IMG

Background

Metagenome sequences used for predicting genes can be broadly classified into two categories -- assembled sequences and unassembled sequences. Assembled sequences, irrespective of the underlying sequencing technology, are typically longer. Unassembled sequences generated by the 454 sequencing platform are longer and therefore easier to assemble and predict genes on, but suffer from lower quality base calls. In contrast, unassembled sequences generated by the Illumina sequencing platform are shorter but have higher quality base calls.

Preprocessing

Metagenome sequences are prepared for annotation by performing the following preprocessing steps:

1. **File cleanup**. This includes the removal of commonly occurring discrepancies in the input sequence files, such as duplicate sequence IDs. Ambiguous nucleotides are replaced by Ns and sequences with characters not occurring in {A,C,G,T,N} are not considered further.
2. **Trimming**. Sequences are trimmed in order to remove low-quality regions and trailing 'N's. While trimming may be applied to both assembled and unassembled sequences, currently, only unassembled sequences are trimmed. In the case of sequences generated using 454 sequencing, quality data from FastQ files is used with Lucy [1] with a threshold of Q13 for Illumina-generated sequences and Q20 for 454-generated sequences to identify regions of low-quality at the ends of sequences and trim them. In the case of sequences obtained from Illumina sequencing, the longest contiguous subsequence that passes the Q13 threshold for all residues is retained. In all cases, trailing 'N's are deleted. Sequences containing more than five occurrences of 'N's are not retained. The output from the trimming process is a Fasta formatted file with sequences that were accepted and trimmed. Each sequence
3. **Masking of low-complexity regions**. Low complexity noisy sequences are identified using the DUST [2] application and eliminated.
4. **Dereplication**. When two or more sequences are more than 95% identical, only one copy is retained. Additionally, the first 3 bps for 454-generated sequences and the first 5 bps for Illumina-generated sequences need o be identical for two sequences to be considered as replicates.

Gene prediction

Genes are predicted in the following order: CRISPRs, non-coding RNA genes, protein-coding genes.

CRISPR elements are identified using the programs CRT [5] and PILER-CR [6]. The predictions from both programs are concatenated and, in case of overlapping predictions, the longer prediction is retained.

The first category of non-coding RNAs, **tRNAs**, are predicted using tRNAscan SE-1.23 [3]. The domain of the organism (*Bacteria*, *Archaea, Eukaryota*) is a parameter that is required for the program. A metagenome is a potential mixture of the three domains of life, so the program is run three times, one for each domain, with custom parameters for each. The best scoring predictions are then selected. Since the program cannot detect fragmented tRNAs at the ends of sequences, sequences are compared to a database of nt sequences of tRNAs identified in all isolate genomes (For sequences longer than 300bps, only the first 150bps and the last 150 bps are matched). Hits with high similarity (at least 85% identity and a minimum alignment length of 40) are kept; all other parameters are set to default values.

Ribosomal RNA (rRNA) genes are another category of non-coding RNAs. Three types of rRNA genes exist -- 5s, 16s, and 23s. Internally developed rRNA models [4] are utilized for predicting each type of rRNA genes.

Protein-coding genes are identified using four different gene calling tools, GeneMark (v.2.6r) [7] or Metagene (v. Aug08) [8], Prodigal [9] and FragGeneScan [10], all *ab initio* gene prediction programs. A majority rule-based decision schema is followed to select gene calls. When there is no clear decision, selection is based on a preference order of gene callers determined by runs on simulated metagenomic datasets. (Genemark > Prodigal > Metagene > FragGeneScan).

CDS and other feature predictions are then consolidated in terms of resolution of overlaps. In the event of an overlap between a protein-coding gene and an RNA gene or CRISPR element, the RNA gene or CRISPR element is retained. Small 3'-3' overlaps between CDSs and RNA genes are allowed.

Every annotated gene is assigned a locus tag of the form PREFIX_#####. The # in the first position indicates the sequence type: 1=assembled, 2=unassembled 454 sequence, 3=unassembled Illumina sequence.) Each locus tag is guaranteed to identify a unique gene within a sequencing project. However, it is up to the user to submit a unique locus tag prefix that will distinguish this project from other genome projects. The number part of each locus tag is incremented by 1 per locus tag. Loci are simply identifiers and are not guaranteed to have any particular order or internal structure. The output of this stage consists of two files: a Fasta formatted file containing all CDS protein sequences and a GFF formatted file placing predicted features on the metagenome sequence.

Functional Annotation

Protein Families

Functional annotation associates proteins coding genes with Pfams, COGs, KO terms, EC numbers, and phylogeny.

Genes are associated with **Pfam-A** are made using hmmsearch [11] to generate Pfam-A HMMs. Model specific trusted cut-offs are used together with an e-value of 0.1 to gather hits. hmmsearch outputs hits that do not qualify the trusted cutoffs when there are multiple hits from the same sequence to different domains on the query HMM; hits that do not satisfy the trusted cutoffs are filtered out. If the overlap between two Pfam predictions is greater than half of the length of the shorter model, the hit having the largest bitscore, lowest e-value, or longer alignment length, in that order of preference is retained.

Genes are associated with **COG**s by comparing protein sequences with the database of PSSMs for COGs downloaded from NCBI. rpsblast v. 2.26 [12] is used with an e-value of 0.01 to find hits. Output filtering follows the same rules as in the case of Pfams.

Assignments of **KO** terms, **EC** numbers, and **phylogeny** are made using similarity searches. The reference database is constructed by starting with the set of all non-redundant sequences taken from public genomes in IMG. To this dataset are added sequences from the KEGG database that are not present in IMG. The reference database is then consolidated and index files are generated to relate gene IDs to taxa, KO terms, EC numbers, and preserve timestamps. usearch [13] is used with the –maxhits=50 and –minlen=20 options to compare predicted protein-coding genes to genes in this database. The top 5 hits for each gene are retained. Phylogenetic assignment is based on the top hit only. For assignment of KO terms, the top 5 hits to genes in the KO index are used. A hit results in an assignment if there is at least 30% identity and greater than 70% of the query protein sequence or the KO gene sequence are covered by the alignment. The same rule is used together with the EC index to assign EC numbers to genes.

Product Names

Protein product names are assigned to genes as follows:

1. First, product names are assigned based on the name of their **COG** hits provided that
    - Check that the gene has a COG assigned (in IMG assigned by RPS-BLAST with e-value 1e-2 retaining top hit only)
    - Check that the gene has at least 25% identity to COG PSSM and alignment length is at least 70% of the COG consensus length

      If both conditions are satisfied, COG name is assigned as product name; if COG name is “uncharacterized conserved protein” or contains “predicted”, the name should be of the format “COG.cog_name, COG.cog_id”. If either % identity or alignment length condition is not satisfied, check whether there are any Pfams assigned to the gene:

- if the gene has assigned Pfams check the table of COG-Pfam correspondence (all_matching COGs_and_Pfams.txt)
- if the gene has a COG.cog_id and all corresponding Pfams (exact match, regardless of their order of occurrence in the gene and in the table) for this COG entry in the correspondence table, assign COG name as product name (irrespective of the % identity and alignment length for both COG and Pfams)

2. For genes that were not associated with a product name using COG, product names are assigned based on the name of their **Pfam** hit:
   - check that the gene has at least one Pfam assigned (by RPS-BLAST with e value of 1e-5, retaining top hits overlapping by no more that 30% of minimum of query or subject sequence length).
   - check that the gene has at least 25% identity to all PSSMs of assigned Pfams and alignment length is at least 70% of each of the Pfams consensus length

     If both conditions are satisfied, the product name will be a concatenation of Pfam family description (attribute “description” in pfam_family) with “protein”. If a protein has hits to multiple Pfams, their descriptions should be concatenated with “/” as a separator and a word “protein” added in the end.
   - if the gene has a COG.cog_id and all corresponding Pfams (exact match, regardless of their order of occurrence in the gene and in the table) for this COG entry in the correspondence table, assign COG name as product name (irrespective of the % identity and alignment length for both COG and Pfams)

Functional Annotation Sources

- KEGG Release 63.0, July 1, 2012
- PFAM 25.0, March 30, 2011

References

1. Hui-Hsien Chou and Michael H. Holmes (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**: 1093-1104.
2. Morgulis A, Gertz EM, Schaffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences.
3. Lowe, T.M. and Eddy, S.R. (1997) *Nucleic Acids Res*, **25**: 955-964.
4. SPARTAN: SPecific & Accurate rRna and tRNA ANnotation
5. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**(1):209
6. Edgar, R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats, *BMC Bioinformatics*, **8**:18.
7. Lukashin A. and Borodovsky M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, **26**(4): 1107-1115.
8. Hideki Noguchi, Jungho Park and Toshihisa Takagi. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*. **34** (19): 5623-5630.
9. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**(1):119.
10. Mina Rho, Haixu Tang, and Yuzhen Ye. (2010) FragGeneScan: Predicting Genes in Short and Error-prone Reads. *Nucleic Acids Res*.
11. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press.
12. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. (2001) *Nucleic Acids Res* **29**(1): 22-28.
13. Edgar,RC (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics* **26**(19), 2460-2461.