Standard Operating Procedure for the Annotations of Genomes and Metagenomes submitted to the Integrated Microbial Genomes Expert Review (IMG-ER) System

Natalia N. Ivanova[1], Konstantinos Mavromatis[1], I-Min A. Chen[2], Victor M. Markowitz[2], and Nikos C. Kyrpides[1]

[1]Genome Biology Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

[2]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA

Genomes that are submitted for inclusion into in IMG-ER without gene predictions and/or gene product assignments, undergo the following gene prediction and product name assignment process:

1. Gene prediction

(1) identification of tRNAs is performed using **tRNAScan-SE-1.23** (Lowe, T.M. and Eddy, S.R. 1997). The kingdom of the organism (Bacteria, Archaea) is a parameter that is required, all other parameters are set to default values.
(2) identification of rRNAs is performed using **BLAST searches** against a database of non redundant rRNAs(16S RNA, 23S RNA, 5S RNA) from complete, finished IMG genomes. This database is updated every six months or upon request. BLAST results are parsed using a script developed by Niels Larsen (Danish Genome Institute) that filters multiple overlapping hits produced by repetitive regions and fragments.
(3) identification of CRISPR elements is performed using the program **CRT** (Bland C. et al. 2007).
(4) identification of other rna genes. All models from Rfam (except for tRNA and rRNA) are used to search the genome sequence by the program **infernal** v0.81 (Griffiths-Jones S. et al. 2005). For faster detection the script **rfam_scan**, provided by the Sanger institute with some in house modifications and adaptations to the newest version of the infernal program, is used. This step is performed by default on draft and finished isolate genomes and on metagenomes upon request. We update the database of Rfam models as soon as a new database is becoming available.
(5) identification of protein coding genes is performed using **GeneMark** (Lomsadze A. et al. 2005). The regions identified previously as RNA genes and CRISPRs are masked with Ns in order to avoid prediction of overlapping genes. Genemark is run using the parameter “combine” which combines the GeneMarkS generated (native) and Heuristic model parameters into one integrated model. In the case of draft isolate genomes and metagenomes each contig is treated separately. At the end of the procedure the masked sequences are replaced with their original content.
(6) a genbank or embl file is generated by combining the information from the above steps using in house developed scripts.

2. Protein Product Assignment

After a new genome is included into IMG-ER the following computations are performed:

(1) RPS-BLAST against COG PSSMs from the CDD database using e-value cutoff of 1e-2 with the top hit retained;

(2) RPS-BLAST against PRIAM database using e-value cutoff of 1e-10, minimum percent identity of 45%, soft masking (-F 'm S') and alignment filter (>70% alignment length on query gene and PRIAM sequence) with the top hit retained;

(3) Hmmsearch against Pfam and TIGRfam databases after BLAST prefiltering. The latter is performed by running BLAST of the new genome against the seed sequences used to generate an HMM model with e-value cutoff of 10 and low complexity masking turned off. All hits with bit scores better than per family noise cutoff (--cut_nc) are retained;

(4) BLASTp against the IMG database using e-value cutoff of 10, soft masking (-F 'm S') with 20 top hits retained.

Product name assignment attempts to assign an **IMG term** (Ivanova N. et al 2007) as product name in the first pass; if no IMG term can be assigned, product name is assigned based on the TIGRfam hit. In the absence of IMG term and TIGRfam hit, product name is assigned based on the COG hit or Pfam hit.

Assignment of IMG terms as product names occurs as follows:

(1) verification that the CDS of interest has at least 5 homologs in IMG database with >50% identity and at least 2 of these 5 homologs have an IMG term assigned. An additional alignment length filter is applied to the homologs with an IMG term (alignment length is >70% of the length of both the query and target proteins) and the IMG terms assigned to these homologs are checked for consistency (i. e. whether the same combination of IMG terms is assigned to all homologs). If all conditions are satisfied, this IMG term (or a combination of IMG terms) is assigned to the CDS of interest as a product name. Multiple IMG terms assigned to the same CDSs are separated by "/".

(2) if assignment of an IMG term as a product name fails, annotation using TIGRfam hit is attempted. If a CDS has a hit above noise cutoff to only one TIGRfam, the name of this TIGRfam is assigned; if a CDS has hits to more than 1 TIGRfams, the name of a TIGRfam of the type "equivalog" is assigned. In the case of several hits to TIGRfams of the type "equivalog" all names of TIGRfams are concatenated into product name with individual TIGRfam names separated by "/".

(3) for the CDSs that were not annotated with either IMG terms or TIGRfam names, the name of their COG hit is assigned as a product name if the CDS has at least 25% identity to COG PSSM and alignment length is at least 70% of the COG PSSM length. If COG name is "uncharacterized conserved protein" or contains "predicted", COG name and COG ID are concatenated in the product name.

(4) if either percent identity or alignment length condition is not satisfied, the CDSs may still be annotated with this COG name provided that it has a hit to Pfam which corresponds to this COG according to the COG-Pfam Correspondence Table. The latter was compiled based by mapping Pfams onto COGs through the genes in the IMG database: if all genes in IMG database with a hit to a certain COG had also hits

to the same Pfam (or the same combination of Pfams), this COG and Pfam(s) were designated as “corresponding COG and Pfam”.

(5) For the genes that were not annotated with either IMG terms, TIGRfam or COGs the names of Pfam hits are used as product names. The product name in this case is a concatenation of Pfam family description (attribute “description” in pfam_family) with “protein”. If a CDS has hits above noise cutoff to multiple Pfams, their descriptions are concatenated using “/” as a separator and a word “protein” added in the end.

The same algorithm is used for assignment of product names to metagenomic proteins except that all filters comparing alignment length to the query and target gene lengths are removed.

References

1. Bland, C, Sabree, F., Ramsey, T.L., Lowe, M., Brown, K. Kyrpides, N.C., Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 208.
2. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**(Database issue):D121-4.
3. Ivanova N.N., Anderson I., Lykidis A., Mavrommatis K., Mikhailova, N., Chen, I.A., Szeto, E., Palaniappan, K., Markowitz, V.M., Kyrpides N.C. (2007) Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes, Lawrence Berkeley National Laboratory Technical Report LBNL-62292.
4. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**(20):6494-506.
5. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**(5):955-64.