

The Standard Operating Procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4)

Marcel Huntemann¹, Natalia N. Ivanova¹, Konstantinos Mavromatis^{1,3}, H. James Tripp¹, David Paez-Espino¹, Krishnaveni Palaniappan², Ernest Szeto², Manoj Pillay², I-Min A. Chen², Amrita Pati¹, Victor M. Markowitz², Nikos C. Kyrpides¹

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

²Biosciences Computing, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA

³Current address: Computational Biology Group, Celgene Corporation

Abstract

The DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4) performs structural and functional annotation for microbial genomes datasets that are then included into the Integrated Microbial Genome (IMG) comparative analysis system. MGAP is applied on assembled nucleotide sequence datasets that are provided via the IMG submission site (<http://img.jgi.doe.gov/submit>). Dataset submission for annotation first requires project and associated metadata description in GOLD (<http://www.genomesonline.org/>). The MGAP sequence data processing consists of feature prediction including identification of protein-coding genes, non-coding RNAs and regulatory RNA features, as well as CRISPR elements. Structural annotation is followed by assignment of protein product names and functions.

Introduction

The DOE-JGI Microbial Genome Annotation Pipeline (MGAP) performs structural and functional annotation of bacterial and archaeal genomes included into the Integrated Microbial Genome (IMG) system [1]. Annotation consists of the identification of RNA and protein-coding genes and repeats, as well as the prediction of functions for each gene (product name assignment). The annotated microbial genomes datasets produced by the MGAP are integrated into IMG, where they can be analyzed or manually edited in the context of a comprehensive set of publicly available genomes.

The MGAP requires a multi-FASTA file of assembled nucleotide sequences as an input for gene calling. In addition, each sequence dataset submitted for annotation needs to be associated with an analysis project that has already been specified in the Genomes OnLine Database [2]. Microbial genome annotation consists of three stages: sequence data pre-processing, feature prediction, and functional annotation. Feature prediction (which includes gene prediction and repeat identification) produces a GenBank file that does not have any functional information for the predicted genes. Subsequently, these genes are assigned functions and are integrated into IMG.

Implementation

The MGAP stages and individual steps are further described below.

Sequence Data Preprocessing

All genome datasets undergo preprocessing in order to ensure that only good quality sequences are processed in the gene prediction stage, as illustrated in Figure 1(i). First, all ambiguous nucleotides in the sequence datasets are replaced by N's and sequences with characters that do not belong to the {A,C,G,T,N} set are not considered further. Additionally, the headers in multi-FASTA files are changed to ensure that all contig and scaffold names are unique and compatible with the tools employed in subsequent stages. The pipeline creates a mapping file, which records the correspondence between old and new sequence headers. Furthermore, sequences shorter than 150 nt are removed. Second, the sequences are trimmed in order to remove trailing 'N's. The trimmed sequences then have to pass a low complexity filtering where low complexity noisy sequences are identified using the DUST [3] application and eliminated. Finally, for finished circular genomes the pipeline attempts to detect the origin of replication by running BLASTx against an in-house curated database of genes located near the origin of replication. If an origin of replication is successfully detected, the sequence is permuted so that it starts at that position.

Feature Prediction

As illustrated in Figure 1(ii), feature prediction starts with the detection of CRISPR arrays, followed by non-coding RNA genes (tRNA, rRNA and other RNA genes), and finally the prediction of protein coding genes.

CRISPR elements are identified using the programs CRT [4] and PILER-CR v1.06 [5]. For PILER-CR the maximum spacer length is set to 100 and the CRISPR element needs to have at least 5 repeats, which have at least 90% identity to each other and at least 75% identity to the consensus sequence of the repeat. For CRT the pipeline runs a modified version of the latest official CRT-CLI 1.2 version. Specifically, the modified CRT has the capability to read multi-FASTA files, detect truncated repeats at the ends of the contigs/scaffolds and deal with spacer artifacts and repeats that contain Ns. This version also executes checks for repeat and spacer length ratios, while the length and similarity checks are performed as part of "all vs. all" spacer and repeat comparisons. Additionally, the progression step of the sliding search window was reduced to 1, while threshold values and search ranges, which were strictly defined in the original software, can be changed from default values on the command line together with the new options and arguments.

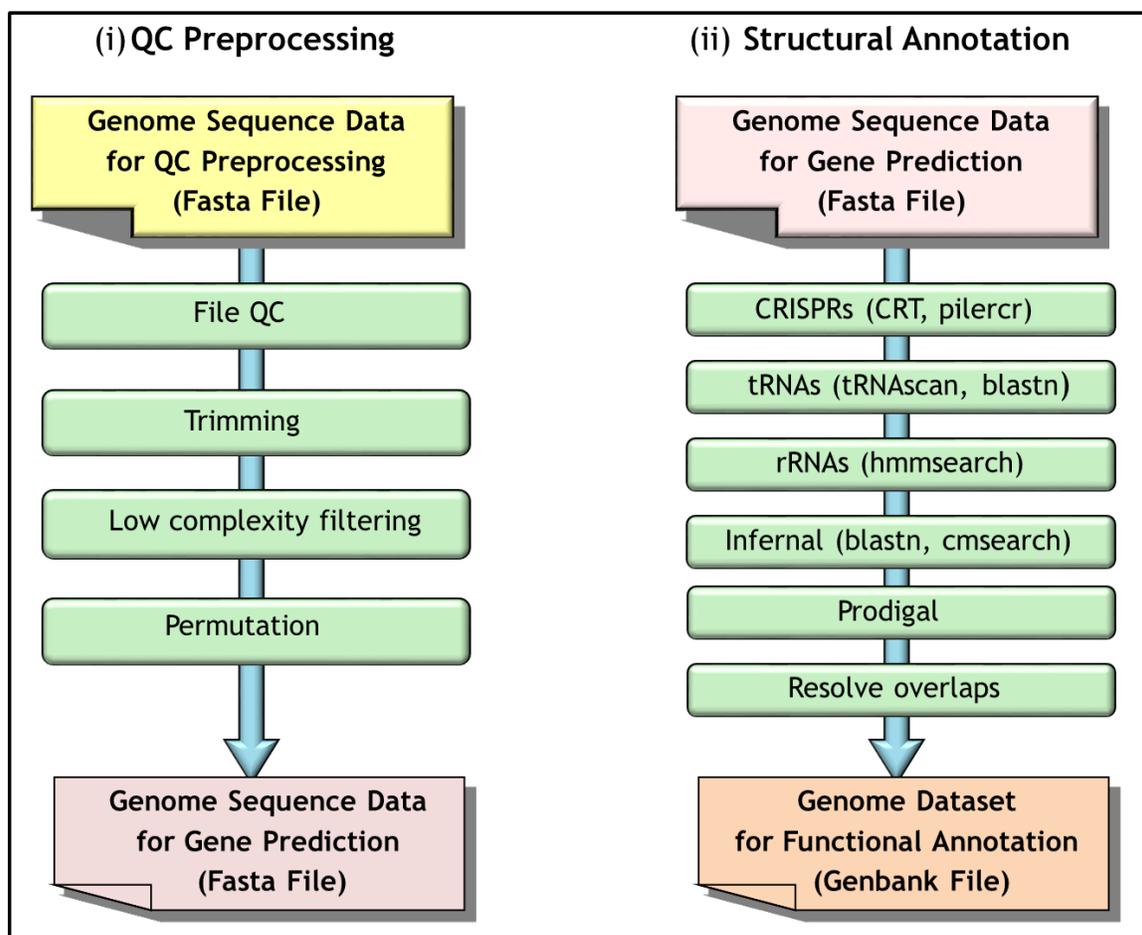


Figure 1. Genome sequence data preprocessing and structural annotation steps.

In our modified CRT the default values for the minimum and maximum repeat lengths are set to 20 and 50 bp, respectively, while the minimum and maximum spacer lengths are set to 20 and 60 bp, respectively. The ratio of the spacer lengths to the repeat lengths have to be between 0.6 and 2.5. The search window is 7 bp long and an element needs to have at least 3 repeats that have a minimum of 70% identity. The predictions from PILER-CR and CRT are concatenated and, in case of overlapping predictions, the CRT prediction is retained.

Protein-coding genes and non-coding RNA genes are identified using a combination of Hidden Markov Models and sequence similarity-based approaches. The first category of non-coding RNAs, tRNAs, are predicted using tRNAscan-SE 1.3.1 [6], and the best scoring predictions are selected. Since tRNAscan fails to detect fragmented tRNAs at the ends of sequences, contig/scaffold sequences are compared to a BLAST database of nt sequences of tRNAs identified in isolate genomes. For sequences longer than 300 nt only the first and the last 150 nt are matched. Hits with high similarity (at least 85% identity and a minimum alignment length of 40 bp) are kept. Ribosomal RNA genes (5S, 16S, 23S) are predicted using hmmsearch tool from the package HMMER 3.0 [7] and a set of in-house curated HMMs derived from an alignment of full-length rRNA genes selected from IMG isolate genomes. Both tRNAscan-SE and hmmsearch use a domain-specific

set of models for Bacteria and Archaea, which is selected based on the taxonomic information provided in the corresponding GOLD Analysis Project.

MGAP also predicts other non-coding RNAs and regulatory RNA features, such as riboswitches. With the exception of tRNAs and rRNAs, all models from Rfam 10.1 [8] are used to search the genome sequences. For faster detection, sequences are first compared to a database containing all the ncRNA genes and other RNA features in the Rfam database using BLASTn, with a very loose e-value cutoff ($1.0e^{-10}$). Subsequently, contigs/scaffolds with hits to this database are searched against Rfam covariance models using the program cmsearch from the INFERNAL 1.0.2 package [9].

The identification of protein-coding genes is performed using the Prodigal v2.50 *ab initio* gene prediction program [10]. Overlaps between predicted features of different type (e. g. ncRNAs and protein-coding genes) get resolved based on an in-house curated set of rules. Every annotated gene is assigned a locus tag of the form PREFIX_SUFFIX where the prefix is the identifier of the GOLD Analysis Project associated with the genome dataset and the suffix is a number that identifies a certain gene on a particular sequence. This assignment scheme guarantees that each gene within a sequencing project gets a unique locus tag. The output of this stage is a GenBank format genome data file.

Every genome data GenBank file must pass an additional validation step before it is forwarded to the next stage for functional annotation. The validation involves checking the file format, matching the gene coordinates to their translations and sequence lengths. The validation also removes phage PhiX sequences, which were identified as a common contaminant in isolate genomes sequenced with Illumina technology. These are identified by running BLASTn against the PhiX genome sequence with an e-value of 0.01 and 90% identity. If a hit covers 80% or more of a query contig/scaffold, the latter gets removed. As a final step the validation script also assesses the quality of a genome, which determines whether it will be included as a reference genome for taxonomic comparisons with other genomes and metagenomes. A genome could get marked as “low quality” and excluded from taxonomic reference database if (a) it lacks phylum-level taxonomic assignment or (b) its coding density (defined as total length of nucleotide sequence of predicted genes divided by the total length of nucleotide sequence) is less than 70% or greater than 100% or (c) there are more than 300 sequences per million base pair or (d) the number of genes per million base pair is less than 300 or greater than 1200. These values were set after manual analysis and benchmarking and are intended to prevent highly fragmented genomes and/or genomes with high rate of sequencing artifacts from being included in a taxonomic reference database.

Functional Annotation

After a genome dataset undergoes structural annotation, the resulting protein-coding genes are compared to protein families and the proteome of selected “core” genomes which are publicly available, and a protein product name is assigned to each gene as discussed below.

Protein Families

1. **COG & KOG assignment:** protein sequences are compared to COG PSSMs obtained from the CDD database [11] using the program RPS-BLAST at an e-value

cutoff of $1e-2$, with the top hit retained. The alignment length needs to be at least 70% of the consensus sequence length.

2. **KEGG Orthology (KO) term assignment:** Genes are associated with KO terms [12] as follows. First, the genes that can be unambiguously mapped to the entries in KEGG Genes database are assigned the KO terms associated with the corresponding KEGG gene. The gene to KEGG gene mapping is based on NCBI's GI numbers and GeneIDs. For genes that are not mapped to KEGG genes, USEARCH is run against the database of KEGG genes by applying UBLAST [18]. The results of this search are organized in a list of candidate KO assignments. KO terms are assigned to genes using a subset of this list, whereby the threshold is defined by an E-value cutoff of $1e-5$, KO assignments are selected from the top 5 hits, with 30% or better alignment sequence identity, and alignment percentage of at least 70% over the length of the query gene and KEGG subject gene.
3. **MetaCyc assignment:** genes are associated with MetaCyc [13] reactions as follows. First, genes are mapped to KO terms as described above, whereby KO terms are associated with Enzyme Commission numbers (EC numbers) using the KEGG KO term to Enzyme relationship provided by KEGG. Next, genes are associated with MetaCyc reactions via EC numbers.
4. **Pfam & TIGRfam assignments:** protein sequences are searched against Pfam [14] and TIGRfam [15] databases using HMMER 3.0. For TIGRfam, the noise cutoff (`--cug_nc`) is used, with hits below the trusted cutoff and at/above the noise cutoff flagged as “marginal”. For Pfam, the gathering threshold (`--cut_ga`) is used inside the `pfam_scan.pl` script (see: ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/OldPfamScan/HMMER2/pfam_scan.pl). The script also helps resolving overlaps between hits to Pfam models from the same clan in order to generate final Pfam assignments.
5. **InterPro Scan:** Additional protein family annotations for SMART, PrositeProfiles, PrositePatterns, and SuperFamily are provided by InterPro Scan (run with default parameters) [16].

Protein Product Names

There are two gene product names associated with every gene in IMG: (i) original gene product name, which is included from the original genome datasets archived at GenBank or submitted by users for inclusion into IMG without IMG's product name assignment; (ii) IMG product name, generated by the IMG's native product name assignment procedure described below. Every gene in IMG with IMG product name is associated with an “IMG Product Source” attribute, which specifies the source of the IMG assigned product name. The value of this field can be:

- **ITERM:xxxxx:** The source of the gene product name is its associated IMG term xxxxx;
- **TIGRxxxxx:** The source of the gene product name is its associated TIGRfam xxxxx;
- **COGxxxx:** The source of the gene product name is its associated COGxxxx;
- **PFAMxxxx:** The source of the gene product name is the associated PFAMxxxx.

IMG Protein product names are assigned to genes in two stages: (i) product names are assigned based on IMG terms whenever they are available, (ii) if IMG terms are not available then protein family associations for genes in individual genome datasets are employed for assigning product names. Protein product names assignments based on IMG terms rely on protein sequence similarities between the genes of the new dataset and genes of all other genomes in the IMG data warehouse.

IMG terms. IMG terms [17] are created and originally assigned to genes by domain experts at the JGI. IMG terms are then propagated automatically to other genes following three complementary methods applied in succession:

1. **Method 1** is based on manually curated IMG terms that are preserved in a BLAST-able database. Genes of new genomes are compared against this database: a gene g with hit to IMG term t sequence in the database satisfying the following criteria will be assigned this IMG term: top hit, $e\text{-value} \leq 1e-5$, $\geq 90\%$ identity, alignment $\geq 80\%$ on both query and subject sequences, smallest to largest sequence length ratio of query and subject sequence $\geq 70\%$
2. **Method 2** is based on a set of rules devised by domain experts at JGI for mapping functional annotations (COG, Pfam, TIGRfam) to IMG terms. An example of such a rule is: "assign IMG Term 6 (*replicative DNA helicase loader DnaB*) to a gene if the gene is annotated with COG3611 (*replication initiation/membrane attachment protein*)".
3. **Method 3** is based on gene bi-directional best hits (BBHs). For a gene g that is not associated with IMG terms:
 - a. Get g 's top 5 BBH genes satisfying the following conditions: sequence alignment length $\geq 70\%$, percent identity $\geq 25\%$. No IMG terms can be assigned to gene g unless there are at least 5 BBH genes satisfying these conditions.
 - b. Let *Set T* be the set of all manually assigned IMG terms (i.e., not automatically populated terms) of any of the 5 BBH genes above. Check each term T_1 in *Set T*: (i) if the 5 BBH genes have conflicting term assignments (e.g., some were assigned term T_1 , while others were assigned term T_2), then no terms in *Set T* can be assigned to gene g ; (ii) if there are no conflicting IMG term assignments and at least 2 of the 5 BBH genes are associated with term T_1 , then assign T_1 to gene g ; (iii) if there are no conflicting IMG term assignments but no IMG terms are assigned to gene g , then repeat this step with top 10 BBH genes.

Bi-directional best hits are computed between protein sequences of pair of genomes using USEARCH (<http://www.drive5.com/usearch/manual/>) by applying UBLAST with a nominal e-value cutoff of $1e-2$. An effective database size (`-ka_dbsize 700000000`) is used in order to make the e-values comparable across pairs of genome computations.

IMG Protein product names. The IMG protein product of a gene *g* is assigned as follows:

1. If gene *g* is associated with one or more IMG terms, then the IMG term becomes the new IMG product name of *g*.
2. For genes that were not associated with a product name using IMG terms, assignment of TIGRfam names as product names is attempted: the gene without a product name is assigned a name of a TIGRfam if it has a TIGRfam hit. If a gene has a hit to only one TIGRfam, the name of this TIGRfam is assigned; if more than 1 TIGRfam is assigned, the name of a TIGRfam of the type “equivalog” is assigned.
3. For genes that were not associated with a product name using IMG terms or TIGRfam names, product names are assigned based on the name of their COG hit. If the COG name is “uncharacterized conserved protein” or contains “predicted”, the name has the format “COG.cog_name, COG.cog_id”.
4. For genes that were not associated with a product name using IMG terms, TIGRfam or COG names, product names are assigned based on the name of their Pfam hit, where the product name is a concatenation of Pfam family description (attribute “description” in pfam_family) with “protein”. If a protein has hits to multiple Pfams, their descriptions are concatenated with “/” as a separator and a word “protein” added in the end.

A translation table for protein product names based on IMG terms, TIGRfam, COG and Pfam descriptions is employed in order to ensure that the product names are compatible with GenBank requirements when the datasets are submitted to GenBank.

Sequence Features

Signal peptide feature prediction employs SignalP 3.0. The model used is determined by the gram stain annotation field for the genome (gram+, gram-, Euk). If the gram stain field is not specified, try all three models. Take any hit first from the HMM model, second from the NN (neural net).

Transmembrane helices are predicted using TMHMM2.0c.

SignalP and TMHMM tools are provided by the Center for Biological Sequence Analysis.

Functional Annotation Sources

- COG 2014 Version (November 2014)
- KEGG Release 71.0, July 2014
- MetaCyc Release 18.1, June 2014
- PFAM 27.0, March 2013
- TIGRfam Release 14.0, January, 2014
- InterPro Data Release 48

Summary

The DOE-JGI MGAP performs annotation for bacterial and archaeal genomes. The pipeline consists of custom scripts and publicly available tools. Consistency and reproducibility of the results produced by the MGAP depend on the tools and annotation resources used in the pipeline. Thus, updated versions of resources such as Rfam, Pfam, and KEGG may improve the breadth and depth of functional annotations.

In order to apply the DOE-JGI MGAP on their datasets, users need to first specify their analysis projects in GOLD and then provide their genome datasets via IMG's data submission site (<http://img.jgi.doe.gov/submit>).

We will continue to extend the MGAP with the goal of improving the identification and characterization of genes in the genome datasets it processes.

Acknowledgements

This work is funded by Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy (Contract No. DE-AC02-05CH11231).

References

1. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 2014;42:D560-7.
2. Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, et al. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* 2015;43:D1099-D1106.
3. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *J Comput Biol.* 2006;5:1028-40.
4. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics.* 2007;8:209.
5. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics.* 2007;8:18.
6. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25:955-64.
7. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol.* 2011;7:e1002195
8. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005;33:D121-4.

9. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009;25:1335-7.
10. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119.
11. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, et al. CDD: a conserved domain database for inter-active domain family analysis. *Nucleic Acids Res*. 2007;35:D237-40.
12. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42: D199–205.
13. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*. 2014;42:D459-71.
14. Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam Protein Families Database. *Nucleic Acids Res*. 2012;40:D290-301.
15. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res*. 2007;35:D260-4.
16. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236-40.
17. Chen IM, Markowitz VM, Chu K, Anderson I, Mavromatis K, Kyrpides NC, et al. Improving microbial genome annotations in an integrated database context. *PLoS One*. 2013;8:e54859.
18. Edgar RC Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460-1.