Microbial Genome Functional Annotation & Data Integration Standard Operating Procedure for the IMG System

Background

The **microbial genome annotation** consists of two stages: **gene calling** and **functional annotation**. Gene calling and repeat identification produce 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-ER. Functional annotation and data integration into IMG-ER occurs every two to three weeks.

Gene prediction

Gene prediction starts with breaking scaffolds into contigs, finding the origin of replication and permuting the genome.

Next, CRISPR elements are detected using CRT [6] and PILERCR [7]. Predictions from both methods are concatenated and in case of overlapping elements, the shorter one is removed.

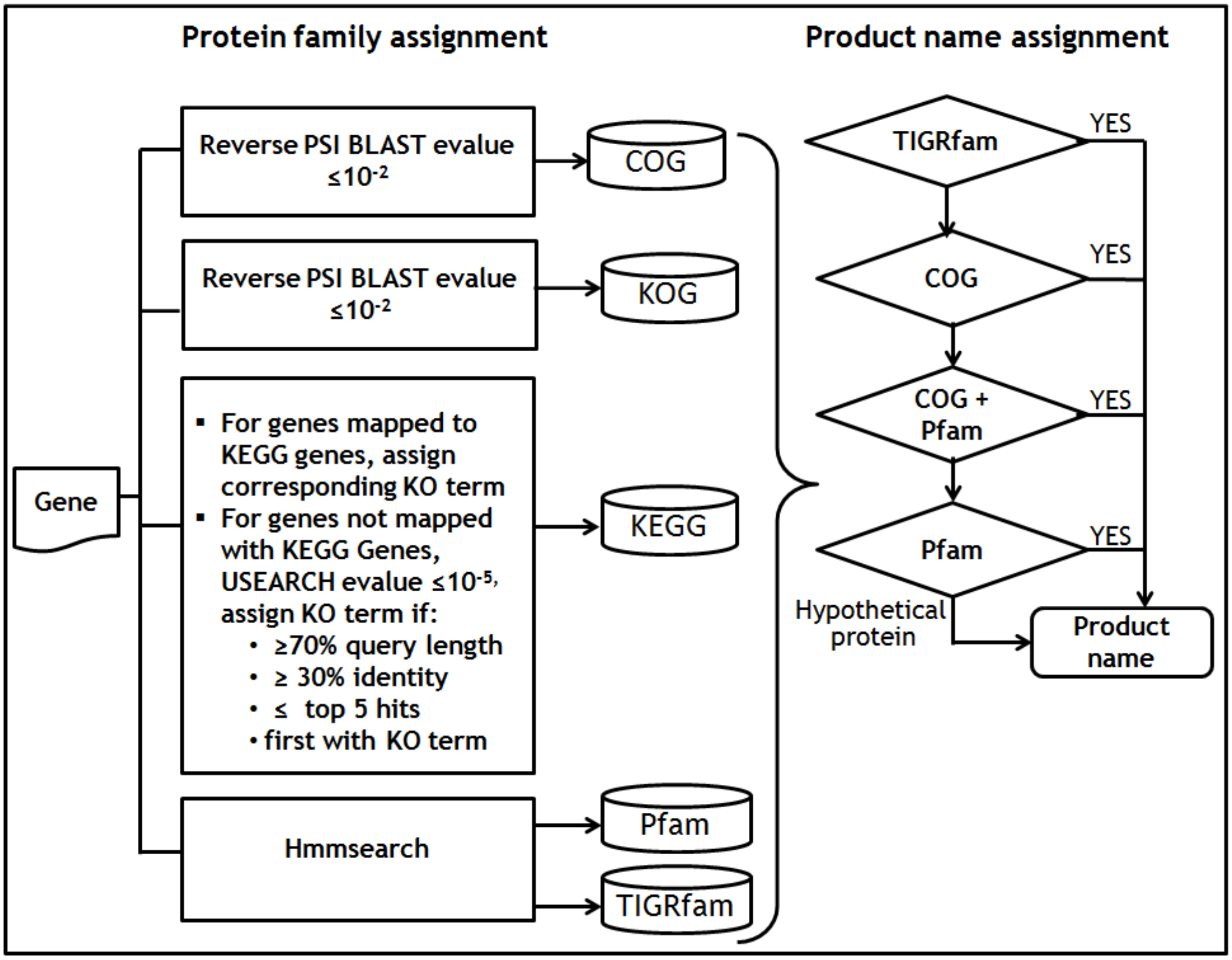
Identification of tRNAs is performed using tRNAScan-SE-1.23 [2]. The domain of the organism (*Bacteria*, *Archaea*) is a parameter that is required; all other parameters are set to default values. Ribosomal RNA genes (5S, 16S, 23S) are predicted using the program RNAmmer [3] using the standard sets of HMMs for RNA genes, provided by the authors. With the exception of tRNA and rRNA, all models from Rfam [4] are used to search the genome sequence. For faster detection, sequences are first compared to a database containing all the ncRNA genes in the Rfam database using BLAST, with a very loose cutoff. Subsequently, sequences that have hits to any genes belonging to an Rfam model are searched using the program INFERNAL [5].

Identification of protein-coding genes is performed using Prodigal [8] which is an *ab initio* gene prediction program. The regions identified previously as RNA genes and CRISPRs are masked with Ns in order to avoid prediction of protein coding genes that overlap RNA genes. In the case of draft isolate genomes each contig is treated separately. At the end of gene prediction, masked sequences are replaced with their original content. In the case of an overlap between a protein coding-gene and an RNA gene, the protein coding gene is truncated to the first start codon (ATG, GTG, TTG) in the same frame that eliminates the overlap or makes it shorter than 30 bp. If this is not possible, the predicted protein-coding gene is removed.

Every annotated gene is given a locus tag of the form PREFIX_#####. Each locus tag is guaranteed to identify a unique gene within this 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 a multiple of 10 allowing the future addition of new genes with loci between the existing ones. Loci are simply identifiers and are not guaranteed to have any particular order or internal structure. The output of this stage is a Genbank format genome file, which is the input for the functional annotation stage.

Functional Annotation

After a new genome is processed, the protein-coding genes are **compared to protein families** (e.g., COGs, Pfam) and the proteome of selected "core" genomes, which are publicly available, and the product name is assigned based on the results of these comparisons, as illustrated in the diagram below.


Protein family assignment
Product name assignment
Gene
Reverse PSI BLAST evalue ≤10-2
COG
Reverse PSI BLAST evalue ≤10-2
KOG
▪ For genes mapped to KEGG genes, assign corresponding KO term
▪ For genes not mapped with KEGG Genes, USEARCH evalue ≤10-5, assign KO term if:
• ≥70% query length
• ≥ 30% identity
• ≤ top 5 hits
• first with KO term
KEGG
Hmmsearch
Pfam
TIGRfam
TIGRfam
YES
COG
YES
COG + Pfam
YES
Pfam
YES
Hypothetical protein
Product name

Protein Families

1. **COG & KOG assignment**: protein sequences are compared to COG PSSMs obtained from the CDD database [9] using the program RPS-BLAST at an e-value cutoff of 1e-2, with the top hit retained.
2. **KO assignment**: IMG genes are associated with KO terms [10] as follows:
   - First, IMG genes that can be mapped to genes in KEGG's list of genes are assigned the KO terms associated with the corresponding KEGG gene. The IMG to KEGG gene mapping is based on using NCBI's GI numbers and GeneIDs.
   - For IMG genes that are not mapped to KEGG genes, USEARCH is run against the database of KEGG genes. The results of this search are organized in a list of candidate KO assignments, where an E-value cutoff of 1e-2 is employed. KO terms are assigned

to IMG genes using a subset of this list, where the threshold defined by an E-value cutoff of 1e-5, KO assignment rank of 5 or better, 30% or better alignment sequence identity, and alignment percentage of at least 70% over the length of the IMG query gene and KEGG subject gene.

3. **MetaCyc assignment**: IMG genes are associated with MetaCyc reactions as follows:
   - First, IMG genes are mapped to KO terms as mentioned above.
   - KO terms are associated to Enzymes EC# using the KEGG KO_term to Enzyme relationship provided by KEGG.
   - IMG genes are linked to Enzyme EC#s via their relationship to KO terms (which are associated to Enzyme EC#s), and are thus associated to MetaCyc reactions via EC#s.

4. **Pfam & TIGRfam assignment**: IMG protein sequences are searched against the Pfam [11] and TIGRfam [12] databases using HMMER 3.0 [13]. For TIGRfam, the noise cutoff (--cug_nc) is used. Hits below the trusted cutoff and at or above the noise cutoff are flagged as "marginal" hits. For Pfam, the gathering threshold (--cut_ga) is used inside the pfam_scan.pl script from Sanger. The script also helps resolves overlaps in the final Pfam.

Product Names

Protein product names are assigned to genes as follows:

1. First, 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. Note: in IMG TIGRfam assignment is performed using hmmsearch with noise cutoff and positive bit score.
   - 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.
2. For genes that were not associated with a product name using TIGRfam names, product names are assigned based on the name of their **COG** hits:
   - 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)

3. For genes that were not associated with a product name using TIGRfam or COG names, 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.

A translation table for protein product names based on TIGRfam, COG and Pfam descriptions in GenBank is constantly formatted throughout the document. This table has been compiled to make the final product names compatible with GenBank requirements and is used upon submission of the genome to GenBank.

Sequence Feature Annotation

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

- KEGG Release 63.0, July 1, 2012
- PFAM 25.0, March 30, 2011
- TIGRfam Release 11.0, August 3, 2011

Data Integration

The integration of new genomes into IMG ER involves computing protein sequence similarities between their genes and genes of all other (new or existing) genomes in the system, assigning IMG terms to the genes of the new genomes, revising protein product names based on IMG terms, identifying fusions, and computing horizontally transferred genes. Note that the phylogenetic distribution for new genomes is not computed on a regular basis, but can be computed on request.

Gene Similarities

Protein sequences of the genes of each new genome are compared to the genes of other new genomes as well as the genes of existing genomes in IMG ER using USEARCH. Protein sequence similarity comparisons are performed incrementally in the following order: (1) genes of new genomes are compared to genes of other new genomes; (2) genes of new genomes are compared to genes of existing genomes in IMG ER, and (3) genes of existing genomes in IMG ER are compared to genes of new genomes.

The results of protein sequence similarity comparisons are recorded in “genome A vs. genome B” files. These files allow extracting: (i) sequence similarity hits ordered by query gene and listing “top hits” in same or other genomes ordered by descending bit score; (ii) bi-directional best hits ordered in query gene top hits format; (iii) genes within a genome hitting the same genome are clustered into paralog groups.

A similar procedure is employed for computing homologs for RNA genes as DNA sequences, with the results recorded the following files: (i) genes of new genomes vs. genes of other new genomes, (ii) genes of new genomes vs. genes of existing genomes; (iii) genes of existing genomes vs genes of new genomes. The RNA homolog data is stored in query gene top hits format.

Fusions

Fusions are identified based on computing USEARCH similarities between genes. Only genes from **finished** genomes are considered as putative components in order to avoid false predictions from fragmented genes in draft genomes. Furthermore, genes that are frequently appear as fragmented in finished genomes, such as *transposases* and *integrases*, as well as *pseudogenes* are excluded from fusion calculations.

Fusions are identified as follows:

- Starting from a *candidate fused gene, x,* in a given genome, *G*, check the similarities to all other genes in other genomes, G_i: for each genome G_i, *candidate component genes* are identified by finding G_i genes that have alignment longer than 80% of their size to gene *x*. Only candidate component genes that overlap for less than 10% of the size with the shortest candidate component gene are kept. Additionally, candidate components should not be paralogs.
- For each *candidate fused gene, x,* in a given genome, *G*, that has candidate component genes in genomes, G_i, *x* is *accepted as a valid fusion* only if

  - The candidate components found in each genome G_i cover more than 80% of *x* (see Figure 1.1 and 1.3).
  - The same combination of component genes is found in at least two genomes (see Figures 1.1 and 1.2), and in at least one of these genomes the components are *not in tandem*, i.e. in at least one genome one or more genes are found between the components (see Figure 2). The second condition eliminates cases of consistent frameshifts in a group of genomes.
- For each *accepted* fused gene, *x*, groups of component genes from *multiple* genomes that hit the same region of *x* are identified: for each of these groups, the gene that has the maximum bitscore/alignment length value is used as the representative component for the group and used as *exemplar*, as illustrated in Figure 3.

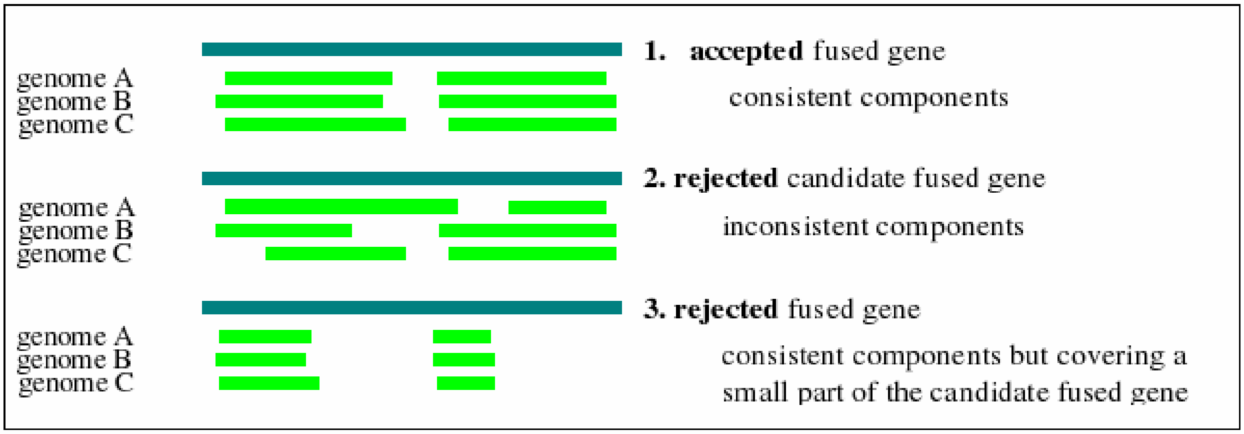


Figure 1. Fused genes and their components.

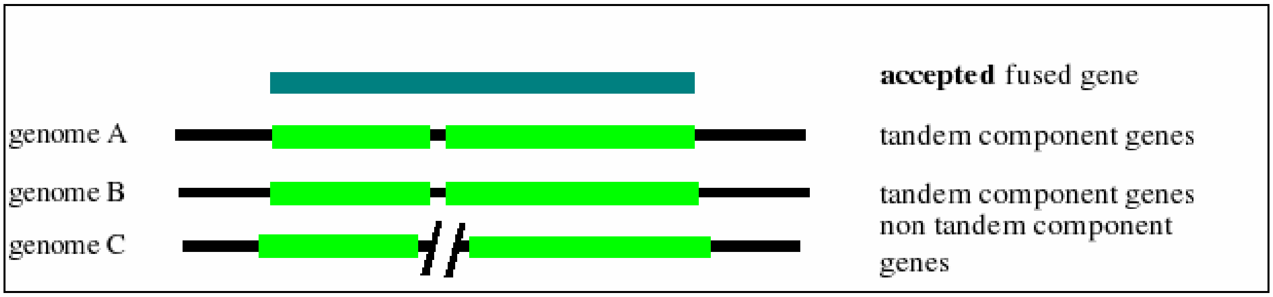


Figure 2. Component genes should appear not in tandem in at least one genome.

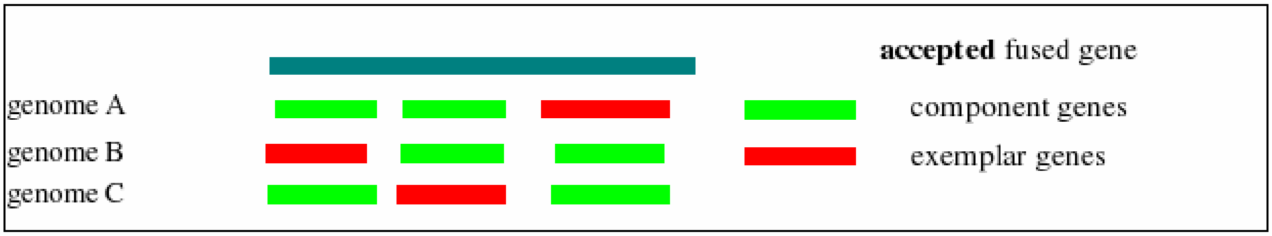


Figure 3. Component and exemplar genes for a fused gene.

IMG term assignment

IMG terms [14] are assigned to genes in IMG ER by domain experts at JGI. IMG terms are then propagated automatically to other genes, including genes of new genomes included into IMG ER, following two complementary methods applied in succession:

1. **Method 1** is based on a set of rules devised by domain experts at JGI. 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*)".
2. **Method 2** 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 overlapping >= 70%, and at least one of these BBH genes having percent identity >= 25%. No IMG terms can be assigned to gene *g* if we cannot find at least 5 of such BBH genes.
    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 T1 in *Set T*:
        - If the 5 BBH genes have conflicting term assignments (e.g., some were assigned term T1, while others were assigned term T2), then no terms in *Set T* can be assigned to gene *g*.
        - If there are no conflicting IMG term assignments and at least 2 of the 5 BBH genes have terms T1, then assign T1 to gene *g*.
        - 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.

Protein product name revision

Gene product names in IMG can be revised or improved after IMG terms are assigned to genes. Note that protein product name revision is **not part** of **IMG's data integration SOP** and is carried out by request only.

The protein product of a gene *g* is revised as follows:

1. If gene *g* is associated with one or more IMG terms, then the IMG term becomes the new product name of *g*.
2. If the gene *g* is not associated with any IMG terms, then check whether *g* has any homolog genes associated with IMG terms as follows: find *g*'s homologs in IMG using USEARCH with e-value 1e-2, retain its top 20 homologs, and
    a. check that at least 5 homologs have >50% identity to the query gene ("good homologs");
    b. check that at least 2 of these "good homologs" are associated with IMG terms;
    c. check that all IMG terms (or all combination of IMG terms) assigned to "good homologs" with terms are consistent;
    d. compare alignment length and protein length: alignment length should be at least 70% of both query gene length and the lengths of each of the "good homologs" with term.

If all conditions above are satisfied, IMG term (or a combination of IMG terms) is assigned as product name. If multiple IMG terms have been found, they are concatenated with "/" as separators.

3. If the gene *g* gets a new product name from an IMG term T1 in Step 2, then assign term T1 to *g* too.

If protein product names are revised, then the following files associated with IMG need to be refreshed to reflect the product name revisions: FASTA files, BLAST databases, IMG data distribution files available via JGI's Genome Portals.

IMG pathway assertion

An IMG pathway is defined as a list of IMG reactions, whereby each IMG reaction is associated with one or more IMG terms.

Consider IMG pathway *P* consisting of IMG reactions *R1*, ..., *Rn*; each reaction *Ri* (1 <= i <= n) may consist of alternative reactions *X1*, ..., *Xm*; each *Xj* is associated with one or more IMG terms. Pathway *P* is **asserted** for a genome *G* if each of *P*'s mandatory reactions *Ri* is asserted, whereby each reaction *Ri* is asserted if at least one of the alternative reactions *X1*, ..., *Xm* is asserted.

Assume that reaction *Xj* (1 <= j <= m) is associated with IMG terms *T1*, .., *Ts*. Then *Xj* is considered to be asserted if: (i) *Tk* (1 <= k <= s) is a *gene product* term, and *G* has at least one gene associated with *Tk*; (ii) *Tk* is a *modified protein* or *protein complex* term with child terms, and *G* has at least one gene associated with each of *Tk*'s child terms.

If genome *G* has genes associated with all the terms required for pathway *P*, then *P* is asserted for *G*. If genome *G* does not have genes with all the terms required for pathway *P*, but there are ortholog genes for the genes of *G* associated with each of the "missing" terms, then the assertion for *G* is set as "unknown". Otherwise, pathway *P* is not asserted for *G*.

Phenotype prediction

Phenotype prediction in IMG is specified using a set of AND-OR rules. Each sub-condition in the rule checks whether an IMG pathway is asserted or not. The evaluation is based on 3-valued logic:

NOT: (e.g., not P1)

Pathway Assertion	Evaluation Result
P1 asserted	false
P1 not asserted	true
P1 unknown	unknown

AND: (e.g., P1 AND P2)

	P2 asserted	P2 not asserted	P2 unknown
P1 asserted	true	false	unknown
P1 not asserted	false	false	false
P1 unknown	unknown	false	unknown

OR: (e.g., P1 OR P2)

	P2 asserted	P2 not asserted	P2 unknown
P1 asserted	true	true	true
P1 not asserted	true	false	unknown
P1 unknown	true	unknown	unknown

For example, Phenotype Prediction Rule for Aerobe is specified as:

- IMG Pathway 768 (Ubiquinol oxidation with oxygen (with proton transport) **OR**
- IMG Pathway 769 (Menaquinol oxidation with oxygen) **OR**
- IMG Pathway 770 (Plastoquinol oxidation with oxygen)

A genome *G* is predicted to have the phenotype "Aerobe" if any of the 3 IMG pathways is asserted for *G*. The genome is predicted not to be aerobic if all 3 IMG pathways are not asserted. Otherwise, the prediction result is unknown.

Horizontally Transferred Genes

Putative horizontally transferred genes are defined as genes that have best hits (best bit scores) to genes that don't belong to the phylogenetic group of the query genome. In computing horizontally transferred genes we use not only the best hit (i.e. the hit with the best bit score) but all the hits that have bit score equal or greater than 95% of the best hit.

Putative horizontally transferred genes are computed as follows:

1. Given a query gene, find a block of "top hits" by taking the maximum bit score, then a block of genes where the bit score is >= 0.95 * max(bit score).
2. Within this group of hits, for each hit, find the highest phyla that differs from the query gene's genome phyla on the levels of domain, phylum, class, and order. The block of hits have to be consistent on this external phyla differing from the query gene's genome. (The query gene's genome cannot be the only member of that phyla else it will always have an external phyla simply due to insufficient data in the database.) If this criteria is met, the query gene is deemed putatively horizontally transferred.
3. Also report reverse hits from block of genes hit if they hit the same phyla as the original horizontally transferred query gene genome's phyla.

Phylogenetic Distribution of Genes

The phylogenetic distribution of genes in a genome against a set of reference isolates is computed with the following parameters:

1. The reference isolates are public genomes.
2. The hit has to be >= 30% identity.
3. E-value <= 1e-2
4. The top hit cannot come from the same genome (self hit or paralog).

The distribution of top hits is computed across the taxonomy and noted at the phylum level. In the case of Proteobacteria and Firmicutes, the phylum level is expanded into including the class level.

Transporter Classification assignment

IMG genes are assigned to TC Families (http://www.tcdb.org/) using a TCDB provided cross mapping of TC Families to PFAM, COG, TFAM identifiers, which were further curated by GBP experts (Iain) to include IMG Terms derived mapping rules as well. The mapping process involves a hierarchical assignment method:

1. First, gene (g) is assigned to a TC Family transitively through its associated PFAM → TC Family mapping data. (PFAM based rule).
2. Remaining un-assigned gene (g) is assigned to a TC Family transitively through its associated TFAM → TC Family mapping data. (TFAM based rule).
3. Remaining un-assigned gene (g) is assigned to a TC Family transitively thru its associated IMG Term → TC Family mapping data. (IMG Term based rule).
4. Remaining un-assigned gene (g) is assigned to a TC Family transitively thru its associated COG → TC Family mapping data. (COG based rule).

SEED product name assignment

IMG genes are assigned to SEED product names and SEED functional roles using md5 checksum computed on the protein sequences and mapped to md5 → SEED provided product names. This assignment is done during major IMG updates and requires asking SEED to periodically dump the necessary data. (We have not updated this more than a year now and SEED group did not respond to my latest email request for IMG 4.0 refresh).

References

1. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, et al. (2012) IMG: the integrated microbial genomes database and comparative analysis system, *Nucleic Acids Res.*, **40**: D115-D122.
2. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997; **25**:955-964.
3. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007; **35**:3100-310.
4. 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**:D121-124.
5. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009; **25**:1335-1337.
6. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007; **8**:209
7. Anonymous. PILER Genomic repeat analysis software. 2009 http://www.drive5.com/pilercr
8. 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(1):119.
9. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, et al. CDD: a conserved domain database for inter-active domain family analysis. *Nucleic Acids Res* 2007; **35**:D237-240.
10. Kanehisa, M, Goto, S, Sato, Y, Furumichi, M and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.*, **40**: D109-114.
11. Punta, M, Coggill, PC, Ebenhardt, RY, Mistry, J, Tate, J et al. (2012) The Pfam protein families database. *Nucleic Acids Research* **40**: D290-D301.
12. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, et al. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Res.* **35**, D260-D264.
13. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998; **14**:755-763
14. Ivanova NN, Anderson I, Lykidis A, Mavrommatis K, Mikhailova N, Chen IA, Szeto E, Palaniappan K, Markowitz VM, Kyrpides NC. Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes. Lawrence Berkeley National Laboratory Technical Report LBNL-62292. 2007.