"Quality assessment of predicted gene models in microbial genomes"

Kim E.*, Dunham B.J., Bogorad I.W., Croitor L.C., Kuo J.H., Zhao X., Anderson I.J., Lykidis A., Mavromatis K., Ivanova N., & Kyrpides N.C.

Production Genomics Facility, DOE/Joint Genome Institute. Walnut Creek, California 94598

Introduction

The Joint Genome Institute (JGI) is developing the Integrated Microbial Genomes (IMG) system (Fig.1) which aids users in the visualization and comparative analysis of genomes from a functional and evolutionary perspective. All IMG genomes pass through the gene model quality assessment (GeneQA) pipeline which not only prevents the propagation of false-positive results produced by computational gene prediction algorithms, but also improves the accuracy of evidence-based ranking of genes. In particular, the GeneQA pipeline is designed to correct inaccurately-predicted start codons, examine overlapping genes, and identify overlooked genes in intragenic regions. Finally, the pipeline attempts to disambiguate real pseudogenes from those resulting from sequencing errors, which in turn prevents fully-functional genes from being disregarded in future comparative analysis.

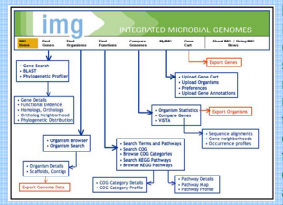

Fig. 1. A map of the Integrated Microbial Genomics (IMG) system. Users can navigate through multiple dimensions of the microbial database: entire genomes, functional annotations organized by pathways and lineages, and individual genes.

Data Cleaning

To better data quality, IMG corrects two kinds of errors: overlapping and non-overlapping. With the aid of BLAST (Fig. 2a) and Artemis v.7 (Fig. 2b), the GeneQA pipeline edits overlapping CDSs, fixes non-overlapping CDS with incorrect start codons, and finds missing CDSs. The correctable errors are compiled into a QA error report (Fig. 4) and manually examined to apply any necessary revisions.

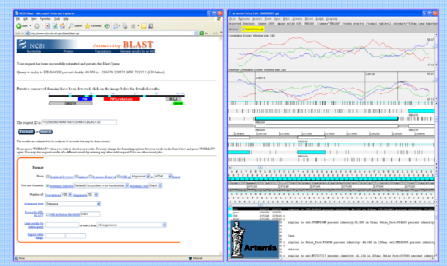

Fig. 2(a-b): Tools used in the GeneQA pipeline. a) Basic Local Alignment Search Tool (BLAST) from the National Center for Biotechnology Information to compare genes to a national database b) Artemis v.7, a DNA sequence viewing and annotation tool, from the Sanger Insitutute

Overlapping Errors:

Overlapping errors are corrected by manually trimming or extending the genes in question. There are three types of overlaps:

- Same-direction (Fig. 3a)
- Convergent (Fig. 3b)
- Divergent (Fig. 3c)

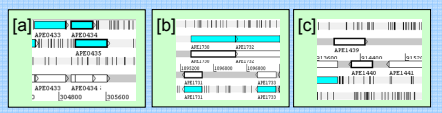

Fig. 3(a-c): Screenshots of overlaps in *Aeropyrum pernix* genome. [a] same-direction (type I). [b] divergent (type II). [c] convergent (type III)

In the case of overlapping genes, either one or both CDS is adjusted at the 5'-end, deleted altogether, or flagged as a pseudogene according to the error that occurs. In order to find the error, the top 10 hits of a BLAST query for each overlapping CDS are compared to pinpoint irregularities in the length of homologs .

Non-overlapping Errors:

a) Incorrect start codon (Fig. 4a)

BLAST queries are run and the 10 best hits are recorded. If a gene is either 30 amino acids longer or shorter than the 10 best hits, then the gene is flagged for future trimming or extending. Pseudogenes are often discovered while searching for incorrect start codons.

b) Potential CDS in intragenic regions (Fig. 4b)

To find missing CDS in intragenic regions, BLAST queries are run for DNA fragments that do not exhibit conserved proteins (usually found between CDS). The DNA fragments should be more than 200 base pairs and its BLAST hits should have an E-value of at least 10^-5.

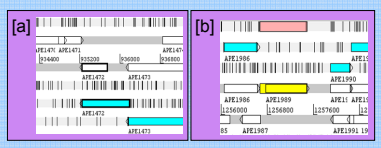

Fig. 4(a-b): Screenshots of non-overlapping errors in *Aeropyrum pernix* genome. [a] A modified start codon. [b] Plausible CDS found within an intragenic region.

Five new tags in IMG nomenclature:

If a gene does not appear to be real and certain criteria are met, then questionable genes are tagged with the use of Artemis v.7 (Fig. 5) :

/short:

- CDS is shorter (less than 80% of the full-length hit) than 90% of their homologs, and
- Cannot be extended to the full length due to the absence of a valid start codon, and
- Missing fragments cannot be identified in upstream or downstream DNA, and
- Short CDS corresponds to a truncated COG or Pfam hit (between 30% and 80%), and
- CDS is at least 100 amino acids in length

/long:

- CDS is longer than 90% of their homologs on the 3'-end, and
- Overlaps with another gene and overlap cannot be corrected with gene truncation or without changing the nucleotide sequence,
- Overhanging part of the gene does not have a COG or Pfam hit, or hits to proteins found outside the genus

/frameshift:

- Two adjacent CDSs are homologous to different parts of the same gene
- Do not have a full-length COG or Pfam hit
- Fragments of this length are present in only one genome or different strains of the same species

/img_pseudo:

- CDS is interrupted by more than one stop codon or frameshift, or
- Two fragments of the CDS are separated by at least one other valid CDS, or
- Single fragment of a CDS corresponds to a severely truncated COG or Pfam (less than 30% of the full-length hit)

/dubious:

- CDS that overlaps with other genes that are determined to be real (code for structural RNA, have COG or Pfam hits, or have homologs outside the genus)
- Codes for short proteins less than 70 amino acids, which are usually unique within a database or have hits of the same length within the same organism\

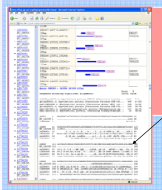

Fig. 4: GeneQA Error Report. This framed interface shows the genomes to be corrected, their individual error reports, and also all associated BLAST results.

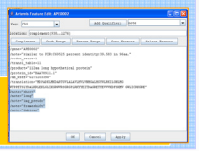

Fig. 5: Edit selected gene feature on Artemis v.7, in which genes can be tagged with the appropriate warning tag.

Observations

Running several archaeal genomes through the pipeline showed varied results. Depending on the source of the sequenced genome, the number of corrections varied either due to evolutionary changes or merely simple sequencing and annotation errors. While genomes exhibited a variety of errors, each genome generally had the same type of error throughout the sequence. Also, the length of the genome cannot be used to predict the number of correctable genes that are found. Additionally, the error report merely records the reason for which a particular gene was considered erroneous; however, this does not indicate the type of tag that a gene may or may not receive. In addition to the tag criteria, examining the data around the erroneous gene is just as important as reading the actual BLAST hits because the presence or absence of homologous fragments differentiates between genes tagged "/short" and those that are "/frameshift" or "/img_pseudo".

Conclusion

The annotation of IMG genomes increases the potential for finding more proteins within a genome as well as improving the ability to establish the evolutionary and functional significance of predicted proteins. The identification of previously missed genes is especially helpful as the number of fully-sequenced genomes rises and hypothetical proteins become associated with more homologs. Ultimately, the comparison of these ambiguous proteins will lead to accurate descriptions of their roles in organisms. Finally, the reduction of the propagation of inaccuracies within databases can be eliminated as any questionable genes are properly tagged. The manual re-annotation of microbial genomes using Gene QA pipeline and aforementioned nomenclature will increase the consistency of gene models in different genomes, thereby facilitating their comparative analysis and functional genomics research.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.