The comprehensive microbial resource

Tanja Davidsen1,*, Erin Beck1, Anuradha Ganapathy2, Robert Montgomery1, Nikhat Zafar1, Qi Yang1, Ramana Madupu1, Phil Goetz1, Kevin Galinsky1, Owen White2 and Granger Sutton1

1J. Craig Venter Institute, Rockville, MD 20850 and 2Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Received August 20, 2009; Revised October 6, 2009; Accepted October 7, 2009

ABSTRACT

The Comprehensive Microbial Resource or CMR (http://cmr.jcvi.org) provides a web-based central resource for the display, search and analysis of the sequence and annotation for complete and publicly available bacterial and archaeal genomes. In addition to displaying the original annotation from GenBank, the CMR makes available secondary automated structural and functional annotation across all genomes to provide consistent data types necessary for effective mining of genomic data. Precomputed homology searches are stored to allow meaningful genome comparisons. The CMR supplies users with over 50 different tools to utilize the sequence and annotation data across one or more of the 571 currently available genomes. At the gene level users can view the gene annotation and underlying evidence. Genome level information includes whole genome graphical displays, biochemical pathway maps and genome summary data. Comparative tools display analysis between genomes with homology and genome alignment tools, and searches across the accessions, annotation, and evidence assigned to all genes/genomes are available. The code underlying the CMR website and the CMR database are freely available for download with no license restrictions.

INTRODUCTION

The large volumes of genomic data being produced with more efficient and cost effective new sequencing technologies will increase the rate of scientific discovery only if investigators are able to find the data that is of specific interest to their research. Today at GenBank (1), the central repository of genome data, there are over 900 complete, publicly available bacterial and archaeal genomes from hundreds of sequencing centers. The computations and searches across multiple genomes that allow scientists to effectively mine this genomic data are only possible when annotation is uniformly applied to all bacterial sequences. Since 2000, the J. Craig Venter Institute [JCVI; JCVI will be used throughout to denote The J. Craig Venter Institute or its predecessor organizations, including The Institute for Genomic Research (TIGR) that was merged with JCVI in 2006] has provided the Comprehensive Microbial Resource or CMR (http://cmr.jcvi.org). The CMR is a central repository containing the sequence and original annotation of complete prokaryotic genomes as well as standard automated annotation across all genomes and precomputed homology searches to allow meaningful genome comparisons. The CMR currently contains 571 genomes and over 50 tools are available to utilize and mine this genomic data. Some of these genomes have websites provided by their sequencing center, and all are available at GenBank. However, comparative genomics and searches across more than a single genome are enhanced when genomic data is located in a common repository with consistent annotation and bioinformatics tools that serve the diverse needs of the scientific community. The CMR has provided such a resource to the prokaryotic scientific community for over nine years.

PROKARYOTIC ANNOTATION DATA ISSUES AND RESOLUTION

Sequencing centers employ a variety of gene finding methods and data management strategies. Similarly, there is considerable variation among annotation groups regarding how function, gene symbols, Enzyme Commission (EC) numbers (2), Gene Ontology (GO) terms (3) and functional role category assignments are produced. Variable approaches to annotation are unavoidable given the size and diversity of microbial genomics; however, robust comparisons across all...
prokaryotic genomes are facilitated when annotation is systematically assigned by applying identical data types. An additional challenge to robust genome comparison are the inconsistencies (4) found across the annotation files submitted to GenBank.

To create consistent data in the CMR, a two-stage approach is employed. The first stage is to carefully extract data from the GenBank records of complete genomes and categorize them based on rigorous data types. Data types such as EC numbers can be located in different tagged fields in the records for different genomes; a customized text parser is used to capture critical elements such as genes, gene product and EC numbers, and load them into explicit fields of a relational database. The second stage is to assign additional, consistent annotation across all genomes using an automated pipeline. The automated annotation pipeline assigns function, gene symbols, JCVI functional role categories [based on Monica Riley’s *Escherichia coli* functional classification system (5)], EC numbers, GO terms and related evidence in a consistent manner, allowing for reliable comparisons across uniform annotation for all CMR genomes.

Annotation from the original sequencing center extracted from GenBank is defined as the ‘primary’ annotation for each genome. The primary annotation and the annotation from the JCVI automated pipeline (called secondary or JCVI annotation) are stored separately in the database, and genes agreed on by both the primary and secondary annotation are linked. Nearly all of the CMR tools can be based on either the primary or the automated annotation, with the option to toggle between the two types. However, because the secondary annotation is automatically generated, the primary annotation is preferentially displayed throughout the CMR.

AUTOMATED ANNOTATION FOR CMR GENOMES

To create the secondary or JCVI annotation for all CMR genomes, JCVI employs an automated annotation pipeline that identifies genome features in the raw DNA sequence, gathers evidence for function of the features, and assigns functional annotation based on the weight of the evidence. Annotation is an ongoing cyclical process and annotation of new genomes or re-annotation of older genomes is improved as new trusted evidence is produced. Figure 1 shows an overview of the JCVI annotation process.

DNA feature identification

Glimmer3 (6) is used to predict protein coding sequences (CDS), tRNAs are identified with the tRNAscan tool (7), rRNA genes and other structural RNAs are identified directly from BLAST (8) matches to Rfam (9), a database of non-coding RNA families.

Evidence for functional annotation

JCVI uses a combination of trusted evidence types which provide consistent functional annotation and can be transferred onto genes with high confidence in an automated fashion. The two major trusted evidence types used in the annotation pipeline are:

- **CHAR database**: JCVI’s CHAR is a curated database of experimentally verified proteins, source publications, and functional annotations. Each protein entry has detailed annotation including function, gene symbol, and GO terms and evidence codes.
- **Trusted protein families**: these families currently include JCVI’s TIGRFAM protein family models (10) and Pfams (11), both built on Hidden Markov Models (HMMs). JCVI is collaborating with other centers to consolidate, validate and incorporate similar high quality protein classification systems [e.g. NCBI’s PRK clusters (12)].

Supporting evidence for the annotation pipeline includes:

- **BLAST searches against PANDA**: PANDA is JCVI’s internal repository of non-redundant and non-identical protein and nucleotide data pulled from public databases that include the latest assembly and protein sequences (e.g. GenBank, RefSeq, UniProt, Protein Data Bank, CMR). PANDA is available on the JCVI FTP site (ftp://ftp.jcvi.org/pub/data/panda).
- **Computationally derived assertions**: computations integral to the pipeline include derived physical and chemical metrics including lipoprotein signals (LP) and transmembrane helices [TmHMM, (13)].

AutoAnnotate

AutoAnnotate weighs the evidence from a precedence-ordered list of evidence types—the CHAR database,
trusted protein families, best protein BLAST matches from PANDA and computationally derived assertions—to annotate each protein by assigning, where possible, a function, gene symbol, EC numbers, JCVI functional role category and GO terms. AutoAnnotate and the databases on which AutoAnnotate runs are freely available for download and installation via the open source repository SourceForge (https://sourceforge.net/projects/prokfunautoanno/).

THE CMR WEB RESOURCE

The CMR comparative database contains complete, public bacterial and archaeal genomes, with a web interface that allows for a wide variety of data retrievals pertaining to inter- and intragenomic relationships for comparative genomics, genome diversity and evolutionary studies. Retrievals can be based on a number of different properties, including molecular weight, hydrophobicity, GC-content, functional role assignments and taxonomy. The CMR interface is designed to make it easy for users to create complex database queries using menu-driven web pages. The CMR has special web-based tools to allow analysis using pre-computed homology searches (i.e. All versus All searches generated using BLASTP), whole genome dot-plots, batch downloading and searches across genomes using a variety of data types.

CMR data model

The Omniome is the production database underlying the CMR, and it holds all of the annotation for each of the CMR genomes, including DNA sequences, proteins, RNA genes and many other features. Associated with each of these DNA features in the Omniome are the coordinates, nucleotide and protein sequences (where appropriate), and the DNA molecule and organism with which the feature is associated. Also available are evidence types associated with annotation such as HMMs (TIGRFAMs and Pfams), BLAST, InterPro (14), NCBI COG (15) and PROSITE (16), individual gene attributes and identifiers from other centers such as GenBank and Swiss-Prot/UniProt (17), manually curated information on each genome and the precomputed All versus All searches.

The CMR tools

New users can learn about the functionality and navigation of the CMR website by utilizing the CMR user manual, Frequently Asked Questions and an on-line tutorial, all available off of the CMR home page. In addition, each page on the CMR provides detailed descriptions on the content and usage of the tool under an information icon ‘i’. Over 50 different tools are available on the CMR, broken into seven major categories:

Searches: Searches allow users to find genes, genomes, sequences, or text from data stored in the CMR. Searches to find genes based on annotation (i.e. locus identifier, functional name, gene symbol), alternate accessions (i.e. GenBank, SwissProt), and evidence or role categories (i.e. EC numbers, GO terms, TIGRFAMs, Pfams, Interpro, Prosite, NCBI COGs and JCVI functional role categories) across all genomes in the CMR are available, as are BLAST searches against the CMR database, HMM sequence searches against TIGRFAMs and Pfams, protein motif searches, and the ability to retrieve nucleotide sequence or list of genes between two coordinates.

Genome tools: Genome level information (Figure 2, Genome Tools) includes graphical displays showing genes placed linearly on regions of the chromosome, or as a complete circle. Other pages give overviews of pathways and subsystems utilizing JCVI’s Genome Properties database (18) and KEGG biochemical pathway maps (19). Codon usage tables, GC plots, computer generated 2D gels, restriction digest tools, JCVI functional role category graphs, and tables summarizing information such as average gene size or numbers of coding regions are also available.

Comparative tools: Comparative tools include homolog analyses and genome alignment displays (Figure 2, Comparative Tools). For example, a whole genome alignment between two bacteria using MUMmer (20) can be seen in a dotplot showing all the genes that are homologous between any two genomes the user selects; schematically, one can see large-scale conserved synteny as well as inversions and translocations in this display. Other displays show protein homology across genomes on a whole genome scale, or focus on a particular region of the genome. The Multi-Genome Homology tool calculates and displays homologous proteins across a single reference genome and up to 15 comparison genomes, with a summary of all proteins unique to the reference and in common with the comparison genomes. The Region Comparison tool aligns the overall best matching regions from other genomes to a user selected reference region.

Lists: The three main categories of lists available are gene lists (e.g. all genes by JCVI functional role category), lists of other genomic elements (e.g. all RNAs in a genome, all intergenic regions in a genome) and evidence lists (e.g. lists of all EC numbers, TIGRFAMs, Pfams and NCBI COGs in the CMR).

Downloads: Both Batch Download and the Gene Attribute Download tools are available. The Batch Download allows users to get a FASTA file of the nucleotide or protein sequence for a set of genes. The Gene Attribute Download allows users to download over 20 different gene attributes (e.g. coordinates, gene symbol, product name, EC number, GenBank ID) for a set of genes. To select genes for these tools the user can upload a list of accessions or select all genes from an organism and/or JCVI functional role category.

Carts: Two carts are currently available, a Genome Cart and a Gene Cart. The Genome Cart allows users to choose their genomes of interest; these genomes are then preselected when a user comes to an organism selection menu. The Gene Cart allows users to select genes of interest while perusing the CMR and for use in later retrievals. Users can select genes into the Gene Cart from any CMR page that shows a list of genes; selected genes can be viewed and downloaded using the Batch
Download and Gene Attribute Download tools from the Gene Cart page.

Gene pages: At the gene level (Figure 2, Gene Page Tools), users are able to view the annotation given to the gene both from the primary and automatic annotation including the product name, gene symbol, EC number, GO terms, JCVI functional role category assignment, DNA sequence and protein sequence. Users may view the TmHMM profile, links to other resources such as UniProt and GenBank, secondary structure, third position GC-Skew and many other displays.

CMR use cases

A review of articles referencing the CMR indicates a variety of use cases. Researchers retrieve gene sequences (21–23), download whole genomes (24) and BLAST against the CMR sequence databases (25). Many scientists use the CMR to provide JCVI functional role category classification across one or more organisms (26–30), showing the importance of having the standard functional classification across all genomes that the CMR provides. The CMR is used for functional classification of genes on microarrays (29) and for microarray design (31). Identifying codon usage and tRNA gene copy number (32), operon identification to provide a genomic link between genes to support laboratory results (33), and the identification of novel genes not called in the original GenBank annotation (34) are other ways the CMR is aiding researchers.

Scientists are using the CMR for comparative genomics including the analysis of potential intragenome transfers with the Multi Genome Homology tool (35), analysis of flanking DNA using the Region View tool (36) and identification of proteins in other bacteria that are similar to a test set using the Genomes Region Comparison tool (37).

CMR updates

New genomes are added to the CMR two to four times per calendar year. Genomes released at GenBank since the last update are downloaded and put through automated annotation and added to the CMR database, the Omniome. JCVI genomes published or released to GenBank since the last update are added to the Omniome, and All versus All homology searches are performed on all genomes. These precomputed BLASTP searches are used throughout the CMR for comparative analysis. Once all data in the Omniome is validated by a series of consistency checks it is tagged and released to the CMR website as a versioned data update. New releases are advertised on the CMR home page.

The CMR currently contains 571 organisms, while GenBank has more than 900 complete prokaryotic genomes. JCVI is currently working on updating the number of genomes in the CMR to reflect and keep up with the complete genomes at GenBank. JCVI expects the CMR to contain over 800 genomes by early 2010, and be caught up with GenBank by mid 2010.

CMR 3-tier system

The CMR is implemented in a ‘3-tier’ architecture written in Perl. The tiers of this architecture are the presentation tier (i.e. user interface), the functional process logic tier, and the data storage and access tier (i.e. database tier).
ACKNOWLEDGEMENTS

The authors would like to thank the J. Craig Venter Institute Information Technology and Bioinformatics Departments for their ongoing technical, engineering and scientific support including Michael Heaney, Dan Haft, Jeremy Selengut, Scott Durkin, Susmita Shrivastava, Lauren Brinkac, Roland Richter, Peter Rosanelli and Tom Emmel; as well as the support received from former employees of The Institute for Genomic Research including William Nelson, Sam Angiuoli and Anup Mahurkar.

FUNDING

Department of Energy [DE-FC02-95ER61962, DE-FG02-01ER63203]; and the National Science Foundation [DBI-0110270]. Funding for open access charge: JCVI.

Conflict of interest statement. None declared.

REFERENCES


