Nucleic Acids Research, 2001, Vol. 29, No. 1 123–125

The Comprehensive Microbial Resource

Jeremy D. Peterson, Lowell A. Umayam, Tanja Dickinson, Erin K. Hickey and Owen White*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 1, 2000; Revised and Accepted November 1, 2000

ABSTRACT

One challenge presented by large-scale genome sequencing efforts is effective display of uniform information to the scientific community. The Comprehensive Microbial Resource (CMR) contains robust annotation of all complete microbial genomes and allows for a wide variety of data retrievals. The bacterial information has been placed on the Web at http://www.tigr.org/CMR for retrieval using standard web browsing technology. Retrievals can be based on protein properties such as molecular weight or hydrophobicity, GC-content, functional role assignments and taxonomy. The CMR also has special web-based tools to allow data mining using pre-run homology searches, whole genome dot-plots, batch downloading and traversal across genomes using a variety of datatypes.

INTRODUCTION

The CMR data is stored in a database called the Omniome. The annotation in the Omniome was derived from the complete bacterial sequences generated by this and several other sequencing centers. See http://www.tigr.org/tdb/mdb/mdbcomplete.html and http://www.tigr.org/tdb/mdb/mdbinprogress.html for the bacterial projects that are completed and underway, respectively. Some annotations of bacterial genomes, which have been sequenced here, are available as text-based information that is resident at other resource centers such as NCBI, EMBL and SWISS-PROT. Additionally, Omniome information is available at our site in a complete, highly structured database and its contents are described here in the following sections: (i) annotation datatypes, (ii) data presentation (iii) manual curation, (iv) automated assignments and (v) data improvement.

ANNOTATION DATATYPES

Gene- and genomic-level datatypes for microbial annotation in the Omniome are presented in Table 1. Genes also have role assignments, coordinates, pI, MW, hydrophobicity values, NCBI gi numbers and links to other sites. The omniome also contains 2104 COGs (1), 996 TIGRFAMs (2) and the results of pre-run searches of all proteins searched against each other. Each organism has taxonomic information, links to other web sites and the source of sequence funding. The data is entirely non-redundant. Some data that is displayed on the CMR such as GC-content, dot-plot information, restriction sites or graphics such as circular genome depictions, are generated on-the-fly and not explicitly stored in the Omniome database. The Omniome is implemented using a commercial Sybase relational database and occupies 951 megabytes of disk space. Description of the Omniome's relational schema is available on request.

Table 1. Datatypes contained in the Omniome

Datatype	Number
Genomes	31
Plasmids, megaplasmids, chromosomes	60
Nucleotides	63 013 470
Sequencing centers	20
Genes	
Function assigned	26 393
Unassigned	39 432
	65 825
EC number assigned	815
Rho-independent terminator assigned (9)	2 983
tRNAs	821
rRNAs	177
sRNAs	10
Repeats	675

OMNIOME DATA PRESENTATION

All Omniome data is available via a World Wide Web interface. Currently the format of retrieval is exclusively text-based information via a web-browser, with some selected information also available by ftp. Web access to sql transactions to the Omniome is available on request. Translation to a free, MySql version of the entire Omniome database is underway. Because the underlying datatypes of the Omniome have been uniformly assigned, the web display is able to smoothly extract data across all bacterial genomes. For example, [for this and all other examples, we have placed a page containing the text for this section with hotlinks to the pages described: http://www.tigr.org/tigr-scripts/CMR2/NAR_examples.html] it is possible to retrieve genes from all completed bacterial

*To whom correspondence should be addressed. Tel: +1 301 838 0200; Fax: +1 301 838 0209; Email: owhite@tigr.org
Present address:
Erin K. Hickey, Motorola, 4088 Commercial Avenue, Northbrook, IL 60062, USA

genomes that have been assigned the same biological role, (e.g., 'Display all genes involved in amino acid biosynthesis'). Similarly, all genes having the same EC number (EC#), common name or gene symbol can be retrieved. Retrieval of genes from all genomes based on protein properties such as pI, molecular weight, GC-content and membrane spanning regions are also available. Complex queries that use many of the above attributes, as well as attributes such as taxon, paralogous gene families, similarity to other proteins, gram-staining, or chromosome topology allow retrievals like 'Display all transporters with >5 membrane spanning domains and have a MW of 36–51 kilodalton'. Every gene has a page displaying a matrix that links to other genes according to different lines of evidence. This page shows associations to other genes based on its membership to a TIGRFAM, COGs, EC#, role and protein similarity. Graphical displays are provided for gene hydrophobicity, as are alignments of those genes to the protein used for its functional assignment. Links to other annotation centers are provided on individual genes or whole genomes. For every microbial gene sequenced here, small-insert library clones can be requested from the TIGR/ATCC clone collection. Custom nucleotide and protein searches are provided, as is the precomputed search of every gene. The precomputed searches make it possible to display candidates from recent duplications, as well as whole genome comparisons. Alignment of the DNA sequences of two complete genomes using the Mummer algorithm (3) is graphically displayed to allow viewing of similar regions in the context of annotated genes. Depictions of genes placed circularly or linearly on the chromosome, restriction digests and overall summary statistics of the Omniome are also provided.

PRODUCTION OF OMNIOME DATA BY MANUAL CURATION

Bacterial genomes sequenced at TIGR have been annotated using computer analyses such pair-wise searches and TIGRFAM comparisons in combination with systematic manual evaluation. This administration of analysis has served to generate highly uniform annotation for 14 complete bacterial genomes. The overall process of curated annotation has now been formalized into a set of documented Standard Operational Procedures (SOPs). SOPs should not be considered a set of computer programs applied against sequence; SOPs represent a structured effort to rigorously analyze and interpret standard software applications for uniform annotation.

Annotation by SOP begins when anonymous DNA sequence is initially searched using Glimmer, a program that assigns probabilities to potential coding regions (4). Glimmer has a ~99% sensitivity for identification of known genes. Predicted coding regions are identified and searched against the non-redundant database of publicly available proteins using the BLAST algorithm. BLAST matches are collected in a subset of proteins. An extended portion of the predicted coding region is then aligned at the DNA-level to hits from the protein subset using PRAZE, a pattern-matching program that employs a modified Smith–Waterman algorithm. PRAZE generates alignments across gapped regions, and into other frames, and is therefore particularly useful to identify frame shifts. Predicted coding regions are also searched against probability tables called Hidden Markov Models (HMMs; 5) that represent information in multiple alignments. The HMMs sets are from two sources, TIGRFAM and PFAM, and provide a sensitive and selective method for functional assignment.

The Glimmer program identifies predicted coding regions; however, additional steps are required for identification of the final sets of genes in a bacterial genome. In some cases (such as regions of the genome that have been laterally transferred), genes that have a sufficiently unusual composition are not detected by Glimmer. To correct for this, the genome is scanned for regions that either contain ORFs without any similarity matches, and for those regions that do not contain ORFs. All six reading frames from these 'intergenic' regions are examined for sequence matches and if any are found within a translation, the endpoints of an ORF are determined from the position of the pair-wise alignment in the region. Candidate genes are then evaluated prior to placement into final annotation. tRNAs are identified by tRNAscan. rRNA genes and other structural RNAs are identified manually. Translational start site accuracy is currently ~75%. Annotators inspect the Glimmer results, compare the match against lengths of orthologous proteins and examine upstream genes to best identify potential starts of translation. Regions containing potential frame shifts are identified and typically are resequenced using alternative sequencing chemistries. Electropherograms are examined in context of the overall assembly, and authentic frame shifts are repaired. Approximately 200 frame shifts are found and resolved in a typical bacterial shotgun sequencing project.

Potential replication origins in microbial genomes are located by a method that examines short oligomers whose orientation is preferentially skewed around the origin (6). These regions are also examined in the context of genes that are frequently observed near origins, and potential replication origins are assigned.

Paralogous genes represent gene duplications within an organism. Identification of such genes is important because increased duplication of genes is associated with biological activity that is specific to that organism's environment niche (7,8). Collection of genes into paralog families increases the confidence each individual gene's assignments. Methods for the identification and annotation of paralogous genes are simple and involve searching against all proteins from the candidate organism using fairly stringent search parameters and inspecting the results. However, no single match criteria is used to collect proteins into paralogous families. The degree of similarity between paralogous genes is the result of duplication that occurred over many different evolutionary time periods, is still unavoidably the subject of interpretation, and varies for each gene family and for each organism.

At least two passes through the predicted coding regions are made: an initial assignment selecting a canonical pair-wise database match based on pair-wise and TIGRFAM HMM searches, and a pass through a gene list grouped by cellular roles. For the second pass, annotators inspect each predicted coding region, weigh various forms of evidence and make a functional assignment for each coding region. The identification of signal peptides and membrane spanning domains involves examination in context of the database matches to identify biologically relevant genes. Annotators assess whether identified role categories are complete, and if not, whether the 'missing' proteins can be found. Biological characterizations of the

studied organism are compared against the gene list and points of potential disagreement are further evaluated.

AUTOMATED ASSIGNMENTS OF OMNIOME DATA

Annotation using the current set of SOPs is labor-intensive. It requires four full-time curators and roughly 1 month to annotate a microbial genome. Continuous application of all SOPs against the world-wide effort in bacterial genomes production is not feasible, and an automated method of annotation was developed. This method uses genes that were annotated manually at this laboratory in a combination of pair-wise searching heuristics and TIGRFAM HMM searches against genes from a new genome. For our initial tests the automated assignments were evaluated by comparing them against final, manually prepared annotation. Correct assignments were of three kinds. One correct set of genes was assigned the exact gene name that was made from a different pair-wise match than the manual annotation, another set of genes had a legitimately synonymous name made from a different pair-wise match than manual annotation, and a final set of genes was assigned the exact gene name and given the exact same pair-wise match as manual annotation. Based on these classes of correct assignments, 93 and 95% of the genes that had been manually curated received the correct assignment using an automated analysis for *Chlamydia trachomatis* and *Vibrio cholerae*, respectively. This methodology has been applied to annotate genomes from other centers. First the data of these genomes was imported to the Omniome. In some cases from annotation retrieved directly those centers, or in other cases was derived from GenBank. Annotation from other centers such as functional assignments, common names and genetic symbols was captured and stored explicitly as original information. The anonymous DNA sequence from these genomes was analyzed using Glimmer and those gene calls were stored in the database. ORFs were then subjected to automated annotation methodology, placed in the Omniome and presented in the Comprehensive Microbial Resource (CMR). The original annotation of the genes is also presented, wherever possible, in the CMR.

ADDITIONAL DATA IMPROVEMENT

Sequence similarity is the most commonly used method for assignment of putative function to a newly discovered gene. Other sequence-based strategies for functional prediction, such as protein motif searching and specialized composition algorithms (e.g., those that measure signal peptide or membrane-spanning domains) supplement similarity. At present, however, most function assignments are the result of 'transitive' assignment by pair-wise comparisons of anonymous genes against the public protein archives. However, in the absence of experimental confirmation of genes that have resulted from high throughput genomic sequencing, many genes from subsequent sequencing projects have been misassigned by error propagation in this process. To systematically overcome ambiguous function calls due to incorrect transitive assignment, we placed genes from completed bacterial genomes into families that are related by function. Misassigned functions typically associated with transitive annotation are corrected during this process. Improved gene annotation is represented in the TIGRFAM collection and in the presentation of those data on the CMR.

ACKNOWLEDGEMENTS

Supported by the US Department of Energy, Office of Biological and Environmental Research, Cooperative Agreement DE-FC02-95ER61962 amendment number 8.

REFERENCES

1. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 22–28.
2. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.
3. Delcher,A.L., Kasif,S., Fleischmann,R.D., Peterson,J., White,O. and Salzberg,S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.
4. Delcher,A.L., Harmon,D., Kasif,S.F., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
5. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
6. Salzberg,S.L., Salzberg,A., Kerlavage,A. and Tomb,J.-F. (1998) Skewed Oligomers and Origins of Replication. *Gene*, **217**, 57–67.
7. Klenk,H.-P., White,O. Tomb,J.-F., Clayton,R.A., Nelson,K.E., Ketchum,K.A., Kerlavage,A.R., Graham,D.E., Kyrpides,N.C., Dodson,R.J. *et al.* (1997) The complete genome sequence of the hyperthermophilic, sulfate-reducing archeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.
8. White,O., Eisen,J.A., Heidelberg,J.F., Hickey,E.K., Peterson,J.D., Dodson,R.J., Haft,D.H., Gwinn,M.L., Nelson,W.C., Richardson,D.L. *et al.* (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, **286**, 1571–1577.
9. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of Transcription Terminators in Bacterial Genomes. *J. Mol. Biol.*, **301**, 27–33.