DEVELOPING BIOINFORMATIC RESOURCES FOR CORONAVIRUSES

Susan C. Baker, Dalia Jukneliene, Anjan Purkayastha, Eric E. Snyder,
Oswald R. Crasta, Michael J. Czar, Joao C. Setubal, and Bruno W. Sobral*

1. INTRODUCTION

Virginia Bioinformatics Institute (VBI) and its partners have been awarded a 5-year contract from NIH-NIAID to establish a national Bioinformatics Resource Center (BRC) to facilitate research on microbial pathogens. As part of this initiative, VBI is developing the PathoSystems Resource Integration Center (PATRIC), a multi-organism relational database to support infectious disease research, especially as it affects biodefense and research on emerging infectious diseases (<http://patric.vbi.vt.edu>). We expect PATRIC to be used as a computational resource to gain insight into mechanisms of microbial pathogenesis and to hasten the development of improved vaccines, diagnostics, and therapeutics. The database will contain high-quality curated data: sequence annotations from published whole and partial genomes; relevant experimental data; metabolic pathway data; taxonomic data; literature citations; and a suite of visualization and analysis tools. Research experts and members of the scientific community will be closely involved at each step of the curation/annotation process. VBI is curating information on a set of eight different pathogen classes that include both bacteria and viruses. Included in this set is the genus *Coronavirus* (family *Coronaviridae*). At present we have archived the annotations of the 153 coronavirus species. These include both whole-genome (130) and partial-genome (23) annotations. This sequence archive represents the initial step in our efforts to curate data on *Coronavirus* species. We welcome active participation by the *Coronavirus* research community in developing PATRIC as a useful computational resource for infectious disease research.

* Susan C. Baker, Dalia Jukneliene, Loyola University Chicago Stritch School of Medicine, Maywood, Illinois.
Anjan Purkayastha, Eric E. Snyder, Oswald R. Crasta, Michael J. Czar, Joao C. Setubal, Bruno W. Sobral,
Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

2. INTEGRATION OF ORGANISM INFORMATION

To facilitate the large-scale annotation/curation project that we have undertaken, we have built an annotation pipeline and associated curation tool interface. The annotation pipeline is composed of gene-prediction programs, similarity search algorithms, and protein structure and function prediction programs. The results of these programs and searches assembled by the annotation pipeline are used to propose biological features that are also stored in the curation database that uses the Genomics Unified Schema (GUS). The scenario for user interaction with the tools is presented in Figure 1. During the manual curation/annotation process, the curation tool interface retrieves the results of the automated annotation process [along with the proposed biological features] and presents them to a curator. Curators review the computational evidence in light of their collective expertise and accept proposed features or edit/remove them.

3. REFERENCE GENOMES

PATRIC genomes are organized into categories based on phylogenetic relationships. The simplest of these PATRIC categories consists of a relatively small number of sequenced genomes from a bacterial or viral family or genus. For the purposes of defining minimal, non-redundant set of genes characteristic of the category, one genome (usually the best-known or best-characterized) is identified as the “reference genome”; the remaining members of the class are called “associated genomes.” For example, the Tor2 and Urbani isolates were the first two SARS coronavirus genomes to be sequenced and therefore were named as reference genomes. Efforts are underway to coordinate our system of reference and associated genomes with the RefSeqs from NCBI.¹

For each organism category, a “reference gene set” is constructed consisting of a single representative of each orthologous group and is built by progressive identification of unique genes from the category’s genomes. The reference genome has the highest precedence and therefore contributes its entire gene complement to the reference gene set. The reference set is then compared at the protein level to the first associated genome and vice versa. Genes from the associated genome identified as orthologs according to the “bidirectional best hit” test are annotated as such. This allows high-value, manually curated information from the corresponding reference genes to be automatically linked to the associated genes, provided minimal similarity criteria based on automated sequence analysis are satisfied. However, because the orthologous genes from the reference genome are already present in the reference gene set, only genes that fail the orthology test are added to the reference set. These genes are presumed to be novel and characteristic of the associated genome. This process is repeated for the remaining associated genomes.

4. PATRIC'S GENOME ANALYSIS PIPELINE (GAP)

The GAP is an automated system for annotating prokaryotic and viral genomes. It consists of two conceptual units, the Genomic Sequence Analysis Pipeline (GSAP) and Protein Analysis Pipeline (PAP) and is configured using GAPML, an XML-based pipeline description language. Submission of a genomic sequence to the database triggers

Figure 1. PATRIC browser / curation tool.

pipeline execution. Analysis begins in the GSAP with programs to identify tRNA, rRNA, and protein-coding genes. The programs tRNAscanSE, BLASTN, Glimmer, and GeneMark, respectively, make the gene predictions. The sequence is processed by the “putative gene interval” (PGI) parser to segment the genome into fragments containing a single gene. This breaks the genome into a manageable size for similarity searches and simplifies interpretation of their results. Because noncoding sequence is included within PGIs, genomic features such as putative RNA secondary structures, transcription regulatory sequences, and other features are annotated and queued for curatorial review. Curators make the final call on the predicted gene coordinates and translation and review the other results prior to submission to the GUS database. The translations are then passed to the PAP where it is first classified with respect to the Reference Protein Set, a

collection of canonical proteins for each category of PATRIC organisms. If the protein is found to be similar to an existing entry in this database based on BLASTP search and very stringent cutoff, it is linked to that sequence and inherits its annotation. Inappropriate links can be broken at any point by curators. If no match is found, the protein is added to the Reference Set. Characterization continues with similarity searches to other databases such as SwissProt, GenPept and PIR. The sequence is then analyzed to identify physical characteristics such as signal peptides, transmembrane segments, secondary structure and PROSITE motifs. The final step involves characterization of functional domains using Pfam and TIGRfam, HMM libraries, SCOP, SMART, and BLOCKS. These programs/databases not only predict features but are also used by the curators to infer functions. Functions are encoded using terms from Genome Ontology (GO) and Enzyme Commission (EC) numbers. Features and functional assignments are then written to the database where they are used to infer pathway membership.

5. FUTURE DIRECTIONS

The information presented above reflects our immediate plans for basic genome annotation. This lays the foundation for our future work, which will include the analysis of metabolic and regulatory pathways and comparative genomics. In addition, we plan to relate this information to RNA and protein expression as data becomes available. Ultimately, the goal of this work is to help the biomedical research community leverage genomic information to better understand the physiology of these organisms and their interaction with their human and animal hosts. In time, this will lead to improved treatment and prophylaxis of disease caused by these potentially deadly organisms.

6. ACKNOWLEDGMENT

This project is funded by NIAID / NIH contract HHSN26620040035C to Bruno Sobral.

7. REFERENCES

1. Bao, Y., Federhen, S., Leipe, D., Pham, V., Resenchuk, S., Rozanov, M., Tatusov, R., and Tatusova, T., 2004, National Center for biotechnology information viral genomes project, *J. Virol.* **78**:7291-7298.