

Developing Bioinformatic Resources for the Coronaviridae

Susan C. Baker¹, Dalia Jukneliene¹, Anjan Purkayastha², Oswald Crasta², Joao Setubal² and Bruno Sobral²

¹Loyola University Chicago Stritch School of Medicine, ²Virginia Bioinformatics Institute, Blacksburg, VA

Abstract

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) that consists of a multi-organism relational database to facilitate research on microbial pathogens (<http://patric.vbi.vt.edu>). 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. We expect PATRIC to be used as a computational resource to gain insight into the pathogenic mechanisms of microbes and in the development of improved vaccines, diagnostics and therapeutics. The database will contain high-quality curated data: sequence annotations from published whole and partial genomes; all relevant experimental data; metabolic pathway data; taxonomic data; relevant literature, 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 the *Coronaviruses*. We welcome active participation by the *Coronavirus* research community in developing PATRIC as a useful computational resource for infectious disease research.

Integration of Organism Information

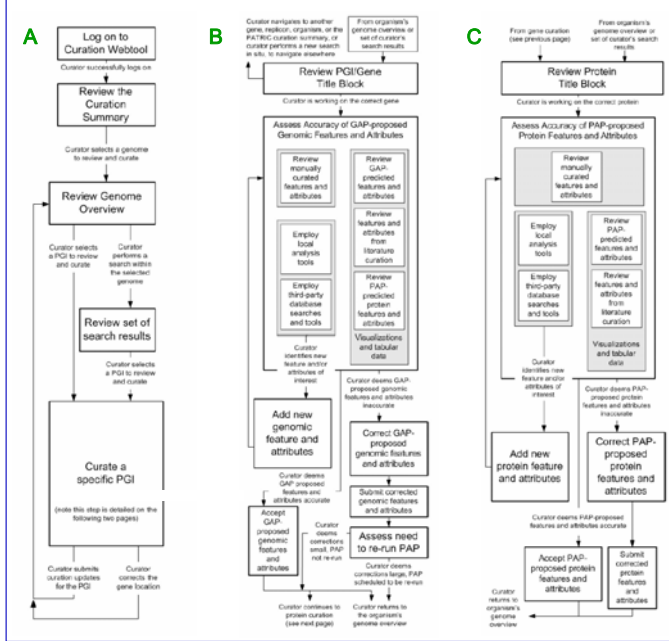
To facilitate the large-scale annotation/curation project that we have undertaken, we are building an annotation pipeline and associated curation tool interface. The annotation pipeline will comprise gene-prediction programs; similarity search algorithms and protein structure and function prediction programs. The results of these programs and searches assembled by annotation pipeline will propose biological features that will be stored in the curation database. The scenario for user interaction with the tools is presented below (Figure 1). During the manual curation/annotation process the curation tool interface will retrieve the results of prediction programs and searches, along with the proposed biological features and present them to a curator. The curator will review all the computational evidence and accept proposed biological feature or edit/remove it.

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 category SARS-coronaviruses, the Tor2 and Urbani isolates were the first two genomes to be sequenced in its entirety and named as RefSeq genomes.

For each organism category, a "reference gene set" is constructed consisting of a single representative of each orthologous group. The reference set is built by progressive identification of unique genes from the category's genomes following the precedence indicated in the figure. 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 to 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, since 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. If all genomes of the class are equally divergent, the number of unique genes identified from each successive associated genome will decrease at each step.

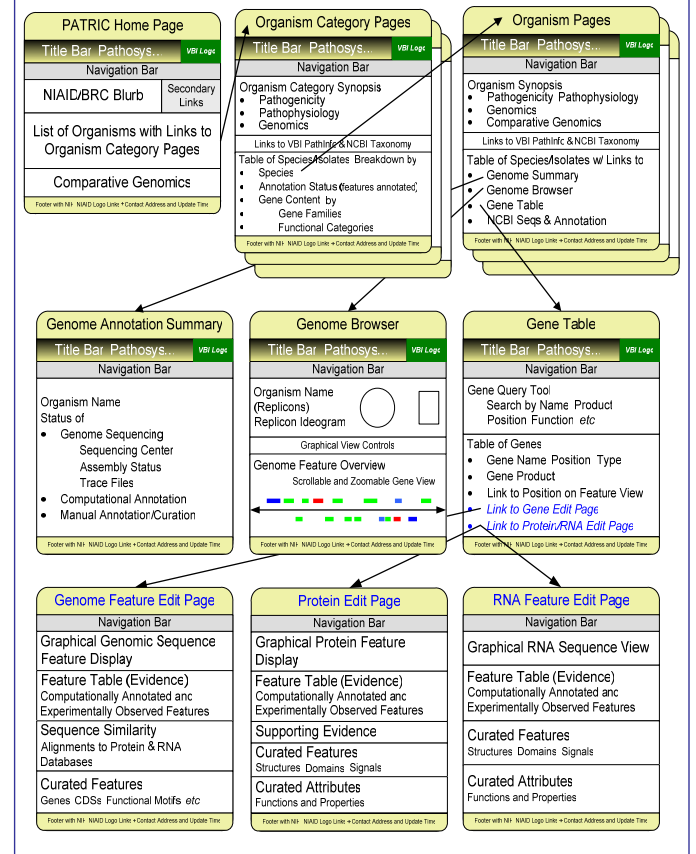
Figure 1: User interaction with the curation interface is described in these three flow-diagrams. Diagram-A provides an overview of the curation process. Diagram-B describes the interaction with the Genome Annotation Tool. Diagram-C describes the interaction with the Protein Annotation Page.



PATRIC's Genome Analysis Pipeline

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 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. These are 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. The PGIs are annotated with putative ribosome binding sites, repetitive sequences and other features 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 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 binned with that sequence. 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.

PATRIC Browser / Curation Tool



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.

Acknowledgement

This project is funded by NIAID / NIH contract HHSN2662040035C.