Poster B-46

GEMINA: Genomic Metadata for Infectious Agents

Authors:
Aaron Gussman (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Sam Angiuoli (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Lynn M. Schriml (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.The Institute for Genomic Researc*)
Kumar Hari (*Ibis Division of Isis Pharmaceuticals, 1891 Rutherford Rd., Carlsbad, CA 92008, USA.*)
Alan Goates (*Ibis Division of Isis Pharmaceuticals, 1891 Rutherford Rd., Carlsbad, CA 92008, USA.*)
Kathy Phillippy (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Tanja Davidsen (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Anu Ganapathy (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Steven Salzberg (*Center for Bioinformatics and Computational Biology, University of Maryland Institute for Advanced C*)
Owen White (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)
Neil Hall (*The Institute for Genomic Research, Rockville, Maryland 20850, USA.*)

Short Abstract: The Gemina (Genomic Metadata for Infectious Agents) system, developed at TIGR, integrates pathogen and epidemiological information with genomic sequence and annotation data. It enables the identification of microbial pathogen species based on their epidemiology and pathogenesis, and supports the development of nucleotide signature-based assays for the detection of pathogens.

Long Abstract:

Background

Scientific inquiry has long been directed towards the investigation of microbial pathogens. Years of research have yielded the sequenced genomes for several strains of major bacterial and viral pathogens. Additionally, there exists a wealth of information on the methods by which pathogens are transmitted to hosts and the physiological effects of an infection.

The Gemina (Genomic Metadata for Infectious Agents) system, developed at TIGR, integrates pathogen and epidemiological information with genomic sequence and annotation data stored at TIGR. The Gemina data model enables the curation, storage, and retrieval of epidemiological data at the level of an individual pathogen, including associated hosts, transmission methods, vectors, environmental reservoirs, historical incidence data, diseases stemming from infection, symptoms, and affected components of the human anatomy.

The Gemina system has been developed to identify microbial pathogen species, strains, genomes, and phenotypes based on their epidemiology, pathogenesis, and related data. It will also support the development of nucleotide and protein signature-based assays for the detection of pathogens or sets of pathogens. This system, freely available to the scientific community, has a broad range of applications, including biodefense, diagnostics, pathology,

and clinical research. Here we describe Gemina's integration with the University of Maryland's Insignia signature pipeline. This system enables the identification of nucleotide signatures which differentiate user-defined target and background sets of organisms.

Epidemiological and sequence data undergo manual and automatic curation before inclusion in the Gemina system. Controlled vocabularies are utilized for the storage and retrieval of epidemiological data. These detailed controlled vocabularies are under active development at TIGR and are freely available to the scientific community. Nucleotide sequences stored in Gemina are systematically organized into tiers of sequence quality. This enables the preferential retrieval of the most-complete non-redundant set of genomic sequence available.

Methods

Gemina incorporates three core components for data modeling and storage: The Microbial Rosetta Stone dataset and schema for epidemiological and incidence data, a Chado database for controlled vocabularies and sequence annotation, and a Panda database for sequence data. These are linked through the use of common taxonomy identifiers. Typically this will be the NCBI taxonomy identifier, but the use of alternative (or multiple simultaneous) taxonomies is supported.

The Microbial Rosetta Stone (MRS), a system under development and currently co-opted for use from the Ibis division of Isys pharmaceuticals, provides a database schema for the storage of epidemiological data. MRS is organized around the concept of a "threat system". Each threat system describes a transmission event between a pathogen and host. This includes the taxonomy of the pathogen and host, the method of transmission, the disease and associated symptoms of the infected host, and the affected components of the anatomy. Incidence data, such as information on specific isolates and the geographic location of outbreaks, is also stored in MRS. The initial data curation effort in collaboration with Ibis has been focused on defining threat systems and cataloging incidence data for the NIAID Category A, B, and C pathogens.

Each element of a threat system is linked to a controlled vocabulary. The controlled vocabularies have been developed at TIGR or adapted from MeSH vocabularies or other sources and integrated into Gemina via a Chado database (also used to store annotation data). The controlled vocabularies are publicly available in Open Biomedical Ontologies (OBO) format.

The Panda database (Protein and Nucleotide Data Archive), developed at TIGR, is used to store a complete, up-to-date set of sequence data. Panda obtains regular updates from NCBI's GenBank, RefSeq, and WGS divisions (among others) and can store sequence datasets not yet submitted to a public repository.

Panda also stores information on the source and completeness of each sequence. This is used to implement a tiered system of sequence quality. When querying for genomic sequence for a taxon, we are able to retrieve the most-complete, non-redundant set of genomic sequence data. The designation and implementation of the tiers is an ongoing process. The current tiers, in order of preference are:

1. The completed Refseq genome if available, otherwise
2. WGS sequence from a genome sequencing project, otherwise

3. Any available sequence from a genome sequence project.

Gemina's public web interface, currently under development, utilizes the controlled vocabularies for searching. Users may select multiple search criteria including pathogen, host, disease, symptom, transmission method, anatomical component, geographic location, isolate date, host age, and host gender. Selections are made from list boxes or tree views of the ontologies and the set of organisms exhibiting those characteristics will be returned. We are also developing an interface for open-ended browsing of the databases' contents to support undirected exploration of the data.

Results

The utility of Gemina can be demonstrated through a sample use case. With Gemina, a researcher can query the database for all strains of E. coli causing diseases with symptoms of vomiting. After obtaining the set of organisms meeting those criteria, he or she can then obtain the larger set of all organisms found in the human gastrointestinal system. Next, the genomic sequence data for each group can be submitted to the Insignia pipeline as the target and background sets respectively. The output of Insignia will be signatures identifying those subsequences of DNA unique to all members of the E. coli target set. These signatures can then be cross referenced against Chado to obtain annotation data, facilitating further scientific inquiry.

Conclusions

Gemina successfully integrates a comprehensive data model for epidemiological data with the annotated genomic sequence of the associated microbial pathogens. This system is a valuable tool for identifying pathogens meeting highly-specific disease criteria and for investigating the complex relationships between disease characteristics and genomic sequence for an organism or group of organisms. It will also support the development of nucleotide and protein signature-based assays for the identification of pathogens or sets of pathogens.