

**Abstract**

We have developed two web accessible database systems to provide bioinformatics resources for the biodefense research. They are: 1. The Pathosystems Resource Integration Center (PATRIC) to provide annotation and analysis of a select group of pathogens 2. The Proteomics Data Center (VBIPDC) to disseminate the host and pathogen data from 7 Proteomics Research Centers (PRCs). Through the integration of the PATRIC and VBIPDC databases and the implementation of data-mining and querying tools, VBI will serve as a Bioinformatics PathoSystem portal for biodefense research and facilitate the discovery of drug, vaccine and diagnostic targets.

**Introduction**

Advances in Genomics and Computer technologies are fueling the development of knowledgebase from large data sets created by high-throughput genomic technologies using digital computation, data, information, and networks. The integration knowledge from various fields such as computer science, chemistry and biology has created a vast opportunity by creating new research environments. Through two NIAID-funded projects, we are developing bioinformatics resources for several Biodefense Pathosystems with three main objectives to:

- A) integrate all genomic sequence data and develop quality annotations of the genes
- B) integrate the transcriptomics and proteomic data and develop the functional characterization of the genomes
- C) develop the software system for integration and interoperability of diverse data sets.

**Methods**

A new automated genome annotation pipeline, a new web-based curation interface, and a user-friendly web portal have been developed at VBI. High-throughput and high quality annotation is facilitated by applying Reference Genome (RG) annotations in a controlled way to their Associated Genomes (AG). Curation means reannotation of published genomes and annotation of novel genome sequences. Results are available at <http://patric.vbi.vt.edu> through query, visualization and analysis tools. PATRIC works closely with scientists in each organism community. PATRIC provides previously available annotations on all its pathogens (315 genomes). We have also updated annotations for RGs *Brucella melitensis* 16M and *Rickettsia prowazekii*. In Rp 15 genes (1.8 %) were revised and 33 new genes were added. By June 2006 updated annotations for all RGs will be available.

Tools to analyze genomic sequence are being developed in parallel with annotation efforts:

1. A system to generate universal primers, useful as pathogen diagnostics, has been developed to design primers against rabies sequences and is being automated.
2. A web-based phyotyping tool, whereby researchers can confidentially submit sequences for alignment and phylogenetic estimation, is being developed. This will create highly refined phylogenetic trees for the user which can then be downloaded.
3. An automated pipeline for the selection of potential membrane surface-attached vaccine targets.

The VBI Proteomics Data Center is accessible at <http://proteinbank.vbi.vt.edu/vbpc>. UML language has been adopted for modeling proteomic area domains. An Oracle hosted proteomics data repository has been developed. Materialized view and stored procedure are employed in the database logic layer. A user-interactive web application has been developed using the J2EE technologies with N-tier architecture. Security mechanisms are provided by role-based Java Authentication and Authorization service and Oracle-based database roles.

**PATRIC Curation Workflow**

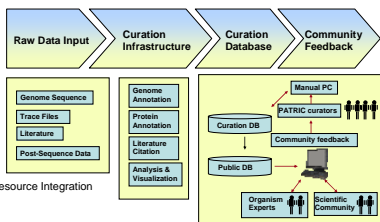


Figure 1: Pathosystem Resource Integration Center (PATRIC) Curation Workflow

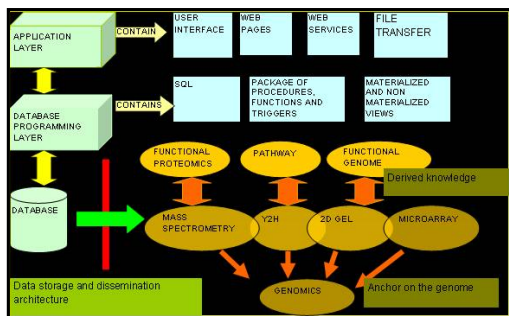


Figure 2: VBI Proteomics Data Center System architecture. The Database system was designed in three layers: Genomics database which serves the basis for all types of biological data; Experimental databases which store all types experimental data; Knowledge databases which are derived from the experimental data integration and curation.

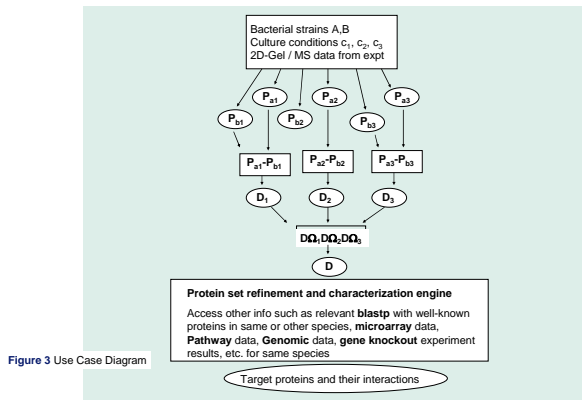


Figure 3 Use Case Diagram

**Given:** We have protein abundance (2D-Gel) data for 2 different strains A, B of a bacterial pathogen in 3 different culture/growth conditions C1, C2, and C3. Strain A is known to be more pathogenic than strain B. The 3 conditions may be normal or free-living, acidic culture and a culture medium deficient in certain nutrient. Proteomic abundance data from the experimental design described offers a unique opportunity to compare and contrast observations between these closely related bacterial strains. One protein family (X) along with other proteins was drastically more abundant in strain A (the more virulent strain) than A. This may be a contributor to the increased pathogenicity in strain A.

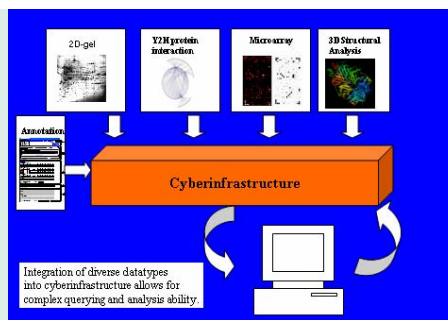
To further investigate and analyze this data, bioinformatics approach is desired. How does one short-list the protein set or gene set of interest starting from this data? What pathway or proteomic analysis would be appropriate to perform on these short-listed proteins and genes? Are these proteins involved in any known metabolic or pathogenic pathways? Are they the targets of any pharmacological studies, and if not, is there a way to identify which proteins are likely to be of interest as drug targets?

**Solution:** Suppose we have run the appropriate 2D gel analysis. We can extract the following information from the analysis:

-Pa1-3 and Pb1-3 represent the sets of proteins that are specifically expressed in strains a and b under the 3 different culture conditions.

-By querying the database where each of the 2D gel or MS data entries for the experiments are catalogued, we can extract the difference in the expressed proteins for the two strains in each of the three conditions. This is represented by Pa1-Pa3, Pb1-Pb3 and Pa3-Pb3.

-The unique set of proteins from all three experimental conditions is determined in the last step resulting in protein subset D. These proteins represent potential virulence proteins which may be candidates for subsequent study. Upon running the query, the detailed annotation information from the annotation database for each of the proteins in D will be instantly available as well as additional bioinformatics tools for their analysis.



**Results**

PATRIC provides previously available annotations on all its pathogens (315 genomes). We have also updated annotations for RGs *Brucella melitensis* 16M and *Rickettsia prowazekii*. In Rp 15 genes (1.8 %) were revised and 33 new genes were added. By June 2006 updated annotations for all RGs will be available. Tools to analyze genomic sequence are being developed in parallel with annotation efforts. A system to generate universal primers, useful as pathogen diagnostics, has been developed to design primers against rabies sequences and is being automated. Additional tools are in development to perform phylogenetic analysis and visualization of selected organisms and to automate the selection of potential membrane surface-attached vaccine targets.

VBIPDC currently supports 2D gel, mass spectrometry, and yeast-two-hybrid data. The system serves browsing, searching, and retrieving data/information for individual project or experiment. XML files are made available to the users through http-based downloading. All genes and proteins in the VBIPDC database are integrated based on the accession numbers. These accession numbers are cross linked with PIR and NCBI websites. The accession numbers also provide a direct link to the VBI pathogen annotation database.

The integration of these two datatypes into a cohesive cyberinfrastructure greatly facilitates the analysis of the proteomes of microbial pathogens. For example, a researcher can query for all proteins which have been experimentally demonstrated to interact with secretion system chaperones and further refine that list by choosing those proteins that have been annotated as having signal peptide characteristics and are conserved among a list of pathogens.

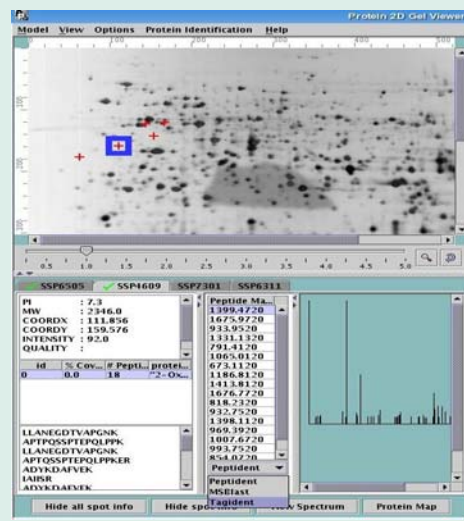


Figure 4: Web visualization. An integrated seamless data visualizing system was developed. Two-dimensional gel and coupled mass analysis spectrometry and microarray analysis datatypes are linked.

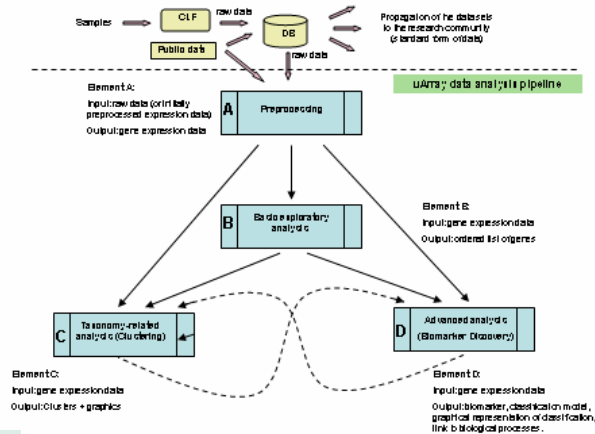


Figure 4: Microarray Data Analysis Pipeline

**Conclusions**

VBI provides a much-needed genomic information resource for important biodefense pathogenic organisms. As more data are integrated into the resource we hope it will become an indispensable tool to researchers worldwide. VBIPDC as part of AC is expected to upload and organize proteomics data for a large number of biodefense organisms and provide access to researchers working on discovery of diagnostics/drug targets/vaccines.

Funding was provided by NIAID contract HHSN266200400035C to B. Sobral, and NIAID contract HHSN266200400061C to SSS, PI - M. Moore, subcontract number: DM1D1-VBI to B. Sobral.