

Poster B-15

Patho-Gene and Pathos Databases – Support for Biodefense and Emerging Infectious Disease Research



Authors:

Elizabeth M. Glass (*Argonne National Laboratory*)

Mark D'Souza (*University of Chicago*)

Alexis Rodriguez (*Argonne National Laboratory*)

Dinanath Sulakhe (*University of Chicago*)

Mustafa Syed (*Argonne National Laboratory*)

Natalia Maltsev (*Argonne National Laboratory*)

Short Abstract: Presentation describes an interactive, integrated, bioinformatics environment to support biodefense and emerging infectious disease research. This environment has two components, PathoGene, a database of pathogenic factors that provides a knowledgebase for analysis of pathogenic organisms and Pathos. Pathos provides an integrated bioinformatics environment for analysis of genomes of pathogenic bacteria.

Long Abstract:

One of the core functional components for biodefense research is a computational infrastructure that support to scientific endeavors for vaccine and drug development and the overall understanding of pathogenic mechanisms.

The Bioinformatics group at Argonne National Laboratory has developed two complimentary databases, PathoGene and Pathos for support of the Region V "Great Lakes" Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Research consortium. Pathos is an integrated bioinformatics environment that provides access to the results of precomputed genetic sequence analysis representing over 417 pathogenic organisms available in public databases. Pathos also contains metabolic reconstructions for more than 180 complete genomes of pathogenic organisms and supports comparative analysis of sequence data and metabolic networks in the framework of taxonomic and phenotypic information. To facilitate identification and characterization of pathogenic factors Pathos also allows interactive analysis of sequence data and metabolic networks by over 30 bioinformatics tools, including tools developed by the Bioinformatics group at the Mathematics and Computer Science Division of Argonne National Laboratory (e.g., CHISEL (<http://compbio.mcs.anl.gov/CHISEL>) and PhyloBlocks (<http://compbio.mcs.anl.gov/ulrich/phyloblock>)), and tools for comparative analysis of metabolic networks). Curated datasets have been developed for species-specific projects in order to provide conjectures regarding proteins of interest to biodefense researchers. Some examples of protein datasets available in Pathos include potential vaccine candidates, possible antimicrobial drug targets, and identification of potential cases of molecular mimicry.

PathoGene is a database of pathogenic factors and corresponding pathogenic processes and provides a knowledge base for analysis of pathogenic organisms in Pathos. Data in

PathoGene is derived from primary literature articles and public resources such as the VFDB (Virulence Factor Database) (Chen et al., 2005) and the Toxin and Virulence Factor Database (<http://www.tvfac.lanl.gov/right.html>), which also have derived virulence factor information from the literature. The system currently describes known pathogenic factors with a primary focus on toxins, adhesion and secretion systems in both gram-negative and gram-positive organisms. The majority of pathogenic processes described in PathoGene are also presented in a graphical form. Each pathway or complex diagram depicts well-studied and documented processes and is annotated with literature references. Information in PathoGene provides an essential basis for comparative analysis and identification of analogous pathogenic factors and systems in organisms with poorly characterized virulence factors and mechanisms.

This bipartite system works integrally to support comparative analysis and discovery of pathogenic factors and features for biodefense research. Both of these resources leverage the PUMA2 system (<http://compbio.mcs.anl.gov/puma2>) which integrates information from over 20 sequence, metabolic, enzymatic and structural databases as well as the results of precomputed analysis of sequence data by variety of bioinformatics tools for publicly available genomes. This information provides the basis for efficient comparative analysis of pathogens by allowing for various types of data and information to be taken into consideration.

The Pathos database provides the bioinformatics environment for the identification and characterization of pathogenic factors, pathogenic processes, and essential genes.

a. Identification of Pathogenic Factors

Information regarding known pathogenic factors from PathoGene can be used for identification of similar factors in other organisms. For example, well annotated pili/fimbriae system components from *E. coli* strains were used to identify eight additional pilin clusters in *Y. pestis* CO92 and KIM. Information from PathoGene regarding all the components of the pili and the accessory proteins (e.g., chaperones and ushers) in *E. coli* strains were used to identify homologs of these proteins in *Y. pestis* using BLAST and analysis of domain composition. Analysis of the conserved chromosomal gene clusters relevant to the pili using the SEED (<http://theseed.uchicago.edu/FIG/index.cgi>) demonstrated that these newly identified pili components were members of the same chromosomal cluster with other pili/fimbrial proteins and in some cases, their accessory proteins. A number of putative and hypothetical proteins were also identified in the cluster suggesting their potential role in pili formation.

Identified sets of homologous pathogenic factors can also be further analyzed in Pathos by using PhyloBlocks. This tool developed by our group allows interactive identification of conserved protein regions and the development of HMM profiles from the set of homologous sequences. These HMM profiles can be utilized for automated identification of pathogenic factors and for the development of oligonucleotide primers to be used by experimentalists.

b. Identification of the Essential genes using Metabolic Reconstructions.

Metabolic reconstructions from the sequence data provide powerful tools for understanding of cellular functionality. They facilitate the detection of the essential genes, taxonomy and phenotype-specific metabolic pathways, as well as existence of alternative metabolic routes

and isozymes. Such information can contribute to the development of new anti-microbial drugs, vaccines and diagnostics. Metabolic reconstructions for over 180 pathogenic organisms are available for further interactive analysis and annotations by experts at the Pathos Web-site.

Pathos and PathoGene can be freely accessed at <http://compbio.mcs.anl.gov/pathos> and the content of these databases can be downloaded at <ftp://ftp.mcs.anl.gov/compbio/pathos>.