BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–1

Pathos and PathoGene Databases – Support for Biodefense and Emerging Infectious Disease Research

*Elizabeth M. Glass, Mark D'Souza,† Alexis Rodriguez, Dinanath Sulakhe, Mustafa Syed, and Natalia Maltsev**

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 60439, USA; †Computation Institute, University of Chicago, Chicago, IL 60637, USA

ABSTRACT

We have developed an interactive, integrated, bioinformatics environment to support biodefense and emerging infectious disease research. This environment has two components, PathoGene and Pathos. The PathoGene database is a comprehensive database of pathogenic pathways and related components containing information extracted from the literature. This resource contains graphical representation of the pathways of pathogenesis derived from the literature. PathoGene provides a knowledge base for the Pathos database. Pathos database is an integrated computational environment for genetic sequence analysis and metabolic reconstructions from sequence data. Pathos currently contains pre-computed analysis of over 170 complete genomes and over 135 automated metabolic reconstructions for completely sequence pathogenic organisms. It also allows interactive comparative analysis of genomes and metabolic networks in the framework of taxonomic and phenotypic information. Pathos and PathoGene databases are freely available at http://compbio.mcs.anl.gov/pathos.

1 INTRODUCTION

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

In response to the need for a computational environment for biodefense and emerging infectious disease research, we have developed two complimentary databases, PathoGene and Pathos, which are available at the following sites: http://compbio.mcs.anl.gov/pathos and http://compbio.mcs.anl.gov/pathogene. These bioinformatics systems were developed to support research at the Region V "Great Lakes" Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Consortium and are publicly available to the wider scientific community.

Pathos is an integrated bioinformatics environment that provides access to the results of precomputed genetic sequence analysis of over 170 genomes of pathogenic organisms available in public databases. Pathos also contains metabolic reconstructions for more than 135 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), PhyloBlocks (http://compbio.mcs.anl.gov/ulrich/phyloblock), and tools for comparative analysis of metabolic networks).

The PathoGene database contains information regarding pathogenic factors and processes extracted from the literature as well as pathway diagrams for some processes. PathoGene database provides a knowledge base for analysis of pathogenic organisms in Pathos.

These resources leverage the PUMA2 system (Maltsev et al., 2006) 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 (e.g., BLAST (Altschul et al., 1990), Blocks (Heinikoff et al., 1999), InterPro (Mulder et al., 2003)) for over 310 complete and over 500 incomplete genomes. This information provides the basis for efficient comparative analysis of pathogens by allowing for various sequence features, taxonomic infor-

* To whom correspondence should be addressed.

mation, pathogenic islands, and enzymatic and metabolic pathway information to be taken into consideration.

In the following sections we describe the content and capabilities of the Pathos and PathoGene systems.

2 DATABASE CONTENT AND CAPABILITIES

2.1 PathoGene

PathoGene is a comprehensive database of pathogenic factors and corresponding pathogenic processes. Data in PathoGene is derived from primary literature articles and describes known pathogenic factors with a primary focus on toxins, adhesion and secretion systems in both gram-negative and gram-positive organisms (Table 1). The majority of pathogenic processes described in PathoGene are also presented in a graphical form (Figure 1). Each pathway or complex diagram depicts well-studied and documented processes and is annotated with literature references.

Table 1. The representation of the known pathogenic factors found in Gram-negative and Gram-positive pathogens in the PathoGene database. Although PathoGene is not exhaustive, it represents current state of knowledge in the literature. Numbers reflect the number of genes recorded for a particular species for that particular factor and does not take into consideration accessory genes that may participate in the entire process. Note a disparity in the amount of information available for these two groups of bacteria in regard to virulence factors.

Pathogenic Factors	*Gram-positive organisms*[a]	*Gram-negative organisms*[a]
Adherence - Pili/Fimbrae	12	55
Adherence - Capsule	1	2
Invasion	7	20
Secretion	0	11
Toxins	14	16

[a] *Genes are only counted once per species even* if it known in several strains.

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.

2.2 Pathos

The Pathos database provides the bioinformatics environment for the identification and characterization of pathogenic factors, pathogenic processes, and essential genes. Examples below illustrate some of the capabilities of Pathos:

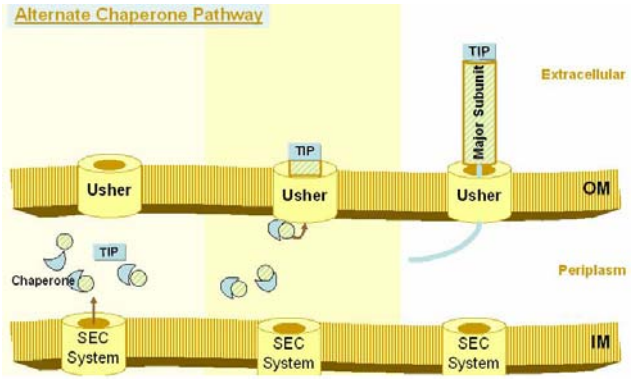

Fig. 1. Graphical representation of pili secreted via the alternate chaperone pathway. Such diagrams are available for pilin/fimbrae systems, secretion pathways, capsule biosynthesis and invasions.

2.2.1 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/fimbrae 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 (Overbeek et al., 2005) demonstrated that these newly identified pili components were members of the same chromosomal cluster with other pili/fimbiral proteins and in some cases, their accessory proteins (Figure 2). 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 tools for phylogenetic analysis and identification of conserved domains and motifs using PhyloBlocks. This tool developed by our group allows interactive extraction of conserved regions in the context of phylogenetic trees and the development of the subsequent HMM profiles (Hidden Markov Model) 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 (Figure 3).

Another example of this approach is the identification of gamma-glutamyltranspeptidase in pathogenic organisms. *Bacillus anthracis* produces a gamma-linked poly-D-glutamic acid capsule that is essential for virulence.

This capsule has a high molecular weight and encoded by the cap genes (capABC). In the presence of gamma-glutamyltranspeptidase (dep), it is degraded to the lower weight L-form. Low-molecular-weight, diffusible poly-glutamates produced through the action of the *dep* gene may act to inhibit host defense mechanisms (Uchida et al., 1993). Using *dep* from *Bacillus anthracis* "Ames Ancestor" as a query, homologous proteins were obtained using another tool developed by our group, BlocksBlast (http://compbio.mcs.anl.gov/blast/BloBla.html). These proteins were then submitted to PhyloBlocks to determine a set of orthologs of dep in other organisms. The consensus sequence and conserved motifs in the resulting set were analyzed by using PhyloBlocks. Protein clusters and consensus sequences for Gram-positive and Gram-negative variations of this enzyme were developed (Figure 4). Such analysis will allow the development of taxonomy-specific sequence signatures and corresponding degenerative primers.

Currently, information in the literature describing pathogenic factors is biased toward Gram-negative bacteria. PathoGene and Pathos provide an excellent venue for discovery of these factors in Gram-positive pathogens through comparative analysis. Such an approach was successfully used by Pallen in a 2002 study of ESAT-6 in *Mycobacterium tuberculosis*. ESAT-6 is a virulence factor that triggers cell-mediated immune responses and IFN-gamma production during tuberculosis. The ESAT-6 proteins found in *M. tuberculosis* were used for identification of ESAT-6 analogous proteins in *Staphylococcus aureus*. This observation led to studies to confirm that these analogs are involved in pathogenesis of *S. aureus* murine abscesses, suggesting that this may be involved in a general strategy of human bacterial pathogenesis (Burts et al., 2005).

2.2.2 Organism-Specific Analyses and Curated Datasets

Curated datasets have been developed for species specific projects in order to identify and characterize proteins of interest to biodefense researchers. Some examples of protein datasets available in Pathos are:

(1) *Vaccine candidates.* We have identified and characterized outer-membrane and extra-cellular proteins conserved in a number of species and annotated with information regarding signal peptide as well as with homology to known antigenic peptides (Olson, 2002). This category of proteins represents prospective protective antigens for vaccine development.

(2) *Antimicrobial drug targets.* We have identified 296 possible essential genes in *Bacillus anthracis.* To perform this analysis, confirmed essential genes from *B. subtilis* and their subsequent metabolic and pathways were compared with the sequences from *B. anthracis*. Candidate genes in *B. anthracis* with high scoring paralogs were eliminated.

(3) *Identification of cases of molecular mimicry*. Immune responses against pathogens can trigger autoimmune processes in the host or exacerbate inflammation via mechanism known as "molecular mimicry" (Bachmaier and Penninger, 2005) (D'Elios et al., 2005). Identification of such cases through rigorous comparative analysis using Pathos allows for prediction of new pathogenic mechanisms and potential immuno-modulating properties of microbial proteins. Identification of molecular mimicry is also essential for vaccine development. Development of a successful vaccine must elicit immunity against multiple serovars while at the same time curtailing damage caused by pro-inflammatory or autoimmune responses. Therefore identification of potential cases of molecular mimicry for vaccine candidates is essential. Pathos supports comparative analysis of microbial sequence data with human genomic data for identification of such cases.

2.2.3 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.

The developed metabolic reconstructions are based on pathway data from the EMP collection of enzymes and metabolic pathways (Selkov et al., 1996), being developed by the EMP Project Inc., containing enzymatic information and metabolic diagrams accumulated from the literature describing over 3000 metabolic pathways in a structured, indexed and searchable form. Metabolic reconstructions for over 135 pathogenic organisms are available for further interactive analysis and annotations by experts at the Pathos Web-site.

This bipartite system works integrally to support a comprehensive comparative analysis and discovery of pathogenic factors and features for biodefense research.

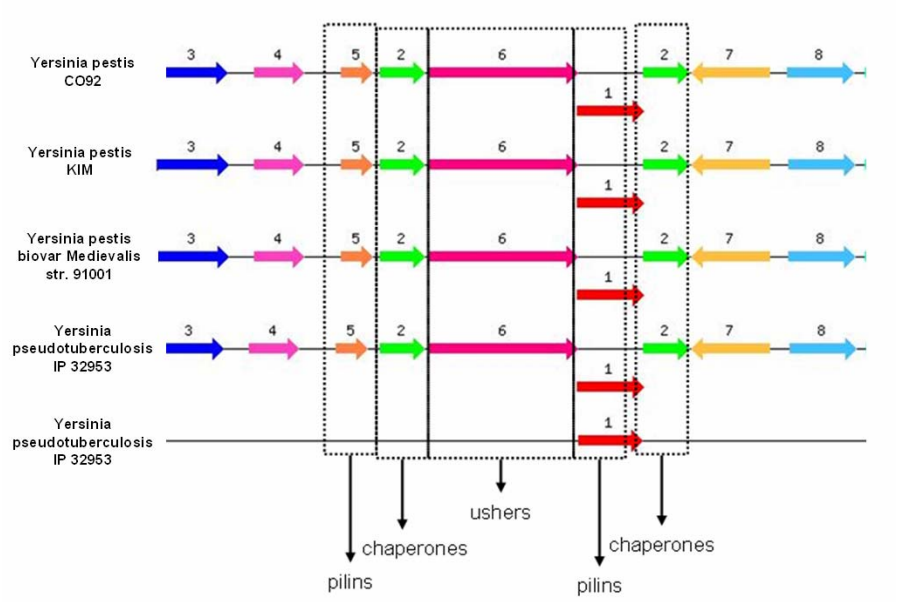

Fig. 2. Conserved chromosomal gene clusters of pili and cognate usher and chaperones of identified pili in Yersinia species. Pili and components were identified using known pili and secretion components from the PathoGene database

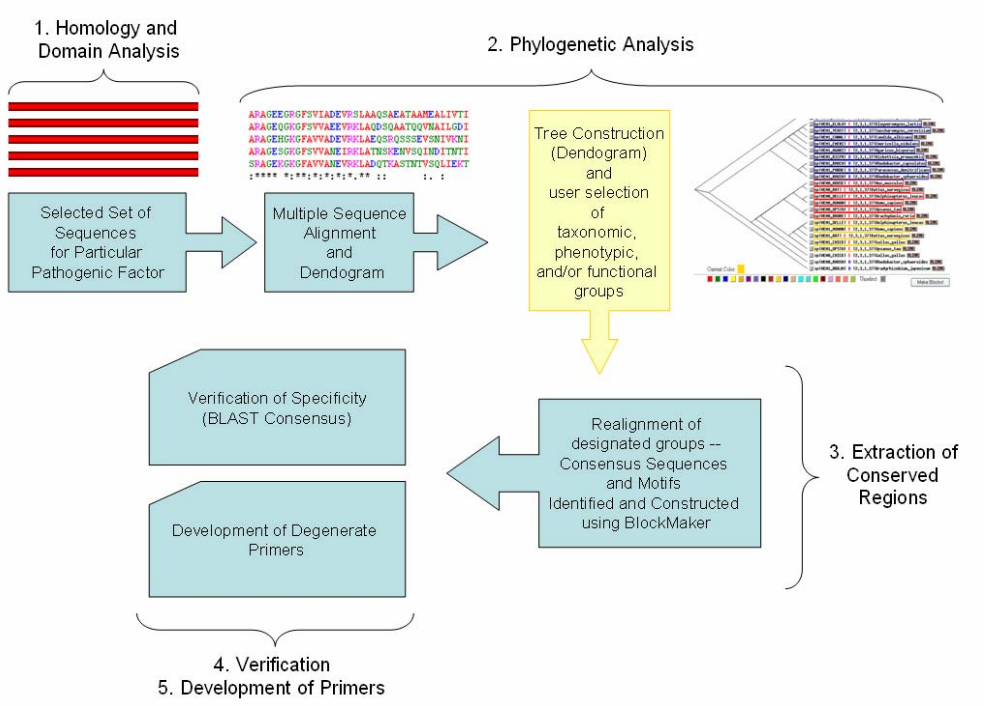

Fig. 3. Interactive extraction of conserved regions using Phylo-Blocks in Pathos. Consensus sequences and profiles are generated from a set of homologous sequences. These profiles can be used for automated identification of pathogenic factors and for the development of degenerative primers.

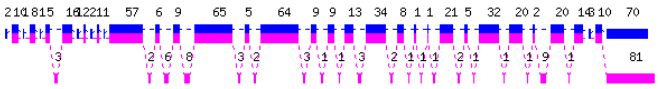

Fig. 4. POAVIZ (Grasso et al., 2003) alignment of gamma-glutamyltranspeptidase consensus sequences derived from Gram-positive and Gram-negative clusters using PhyloBlocks. Gram-positive sequence is designated in blue and Gram-negative in pink. The alignment demonstrates regions of homology and variability between Gram-positive and Gram-negative variations of glutamyltranspeptidase.

3 DATABASE ACCESS

Pathos and PathoGene can be freely accessed at http://compbio.mcs.anl.gov/pathos. Open user registration is required for the development of user-curated metabolic reconstructions and annotation of the data. Curated datasets are available upon request. Requests should be sent to mic@mcs.anl.gov.

ACKNOWLEDGMENTS

Special thanks go out to Luke Ulrich for the development of PhyloBlocks, John Peterson for the development of the initial user interface, and the members of the bioinformatics group at the Mathematics and Computer Science Division of Argonne National Laboratory for their support.

N. Maltsev, E. M. Glass, and M. Syed acknowledge membership within and support in part and D. Sulakhe in full from the Region V "Great Lakes" Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Consortium (GLRCE, National Institute of Allergy and Infectious Diseases Award 1-U54-AI-057153). M. D'Souza acknowledges membership and support to NMPDR Bioinformatics Resource Center NIH/NIAID – (Award NNSN 266200400042C). Alexis Rodriguez was supported by the Office of Biological and Environmental Research, US Department of Energy, under Contract W-31-109-Eng-38.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*., **215**(3), 403-410.

Bachmaier, K., Penninger, J.M. (2005) Chlamydia and antigenic mimicry. *Curr Top Microbiol Immunol*., **296**, 153-163.

Burts, M.L., Williams, W.A., DeBord, K., Missiakas, D.M. (2005) EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of Staphylococcus aureus infections. *Proc Natl Acad Sci U S A*., **102**(4), 1169-1174.

D'Elios, M.M., Amedei, A., Benagiano, M., Azzurri, A., Del Prete, G. (2005) Helicobacter pylori, T cells and cytokines: the "dangerous liaisons". *FEMS Immunol Med Microbiol*. **44**(2), 113-119.

Grasso, C., Quist, M., Ke, K., Lee, C. (2003) POAVIZ: A Partial Order Multiple Sequence Alignment Visualizer. Bioinformatics, 19, 1446-1448.

Henikoff, S., Henikoff, J.G., Pietrokovski, S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*. **15**, 471-479.

Maltsev,N., Glass,E., Sulakhe,D., Rodriguez,A., Syed,M., Bompada,T., Zhang,Y. (2006) PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**(Database issue), D369-72.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R., Zdobnov, E.M. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, **.31**, 315-318.

Olson, S.A. (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform*. **3**(1), 87-91.

Overbeek, R., Begley, T., Butler, R.M,, Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*., **33**(17), 5691-702.

Pallen, M.J. (2002) The ESAT-6/WXG100 superfamily -- and a new Gram-positive secretion system? *Trends Microbiol*., **10**(5), 209-212.

Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., et al. (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database *Nucleic Acids Res.,* **24**, 26–28

Uchida, I., Makino, S., Sasakawa, C., Yoshikawa, M., Sugimoto, C., Terakado, N. (1993) Identification of a novel gene, dep, associated with depolymerization of thecapsular polymer in Bacillus anthracis. *Mol Microbiol*., **9**(3), 487-496.