Published online 4 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D647–D650***
doi:10.1093/nar/gkn799

PIG—the pathogen interaction gateway

Tim Driscoll[1], Matthew D. Dyer[1], T. M. Murali[2] and Bruno W. Sobral[1,*]

[1]Virginia Bioinformatics Institute and [2]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Received August 15, 2008; Revised September 15, 2008; Accepted October 10, 2008

ABSTRACT

Protein–protein interactions (PPIs) play a vital role in initiating infection in a number of pathogens. Identifying which interactions allow a pathogen to infect its host can help us to understand methods of pathogenesis and provide potential targets for therapeutics. Public resources for studying host–pathogen systems, in particular PPIs, are scarce. To facilitate the study of host–pathogen PPIs, we have collected and integrated host–pathogen PPI (HP–PPI) data from a number of public resources to create the Pathogen Interaction Gateway (PIG). PIG provides a text based search and a BLAST interface for searching the HP–PPI data. Each entry in PIG includes information such as the functional annotations and the domains present in the interacting proteins. PIG provides links to external databases to allow for easy navigation among the various websites. Additionally, PIG includes a tool for visualizing a single HP–PPI network or two HP–PPI networks. PIG can be accessed at http://pig.vbi.vt.edu.

INTRODUCTION

Protein–protein interactions (PPIs) play a vital role in initiating infection in a number of pathogens. For example, HIV uses host surface proteins to gain entrance to the host cell. HIV protein ENV attaches to the host human glycoprotein CD4 and subsequently to host chemokine receptors CCR5 and CXCR4. These binding events cause conformational changes to viral proteins that allow the membrane of the virus to fuse to the host cell membrane and enable the virus to enter the host cell. Knowing which PPIs allow a pathogen to infect its host provides critical insights into methods of pathogenesis and potential targets for therapeutics. Unfortunately, resources for studying host–pathogen PPIs require the navigation of several websites. To this end we have created the Pathogen Interaction Gateway (PIG), an integrated platform of experimentally verified and manually curated host–pathogen PPIs (HP–PPIs) and associated computational tools.

Currently there are a number of public databases (1–4) and other resources (e.g. http://www.proteomicsresource.org) that store data for experimentally verified and manually curated host–pathogen PPIs. PIG is designed to integrate data from these various public resources and primary literature into a single data warehouse. Currently PIG only contains data on human–pathogen PPIs; data for other host–pathogen systems will be included as they become available.

Our goals for PIG are: (i) to create a centralized location for experimentally verified and manually curated HP–PPIs that is integrated with other public resources; (ii) to provide an easy-to-use web interface for accessing and using data that would otherwise require the navigation of several websites; (iii) to provide a platform upon which various tools can be developed for identifying potential targets for therapeutics and (iv) to set the stage for developing methods for predicting host–pathogen PPIs.

CONSTRUCTION AND CONTENT

We designed a PostgreSQL relation database to store the data in PIG. From each interaction database, we gather important information for each HP–PPI including protein IDs, organism information, literature references and the experimental method used to identify the interaction. Next, we map all protein ids to UniProt (5) entry IDs to allow for easy integration of other external resources and genomic information. We ignore any interaction in which a protein does not have a mapping to the UniProt system. We focus on HP–PPIs corresponding to host–pathogen systems of interest (currently only human pathogens). We then create a non-redundant set of HP–PPIs and insert them into PIG. In addition to HP–PPIs we also gather and integrate intra-species PPIs from public resources (1,3,4,6–9). We obtain protein sequence

*To whom correspondence should be addressed. Tel: +1 540 231 2582; Fax: +1 540 231 2606; Email: sobral@vt.edu
Correspondence may also be addressed to T. M. Murali. Tel: +1 540 231 8534; Fax: +1 540 231 6075; Email: murali@cs.vt.edu
Present address:
Matthew D. Dyer, Department of Microbiology, University of Washington, Seattle, WA 98195, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Summary of the host–pathogen PPI data currently stored in PIG

Group	Number of PPIs	Number of strains	Number of references
HIV	9095	49	1035
Yersinia	4111	3	5
Bacillus	3077	3	11
Francisella	1383	1	1
Hepatitis	1244	16	95
Influenza	287	4	5
Papillomavirus	229	12	177
Epstein Bar virus	206	2	36
Adenovirus	82	9	61
Herpesvirus	63	20	56
Sarcoma virus	51	6	57
Clostridium	45	4	5
T-lymphotrophic virus	25	2	12
Escherichia coli	22	2	7
Chlamydia	20	2	3
Neisseria	16	1	3
Streptococcus	14	5	7
Vaccinia virus	13	4	8
Staphylococcus	12	3	17
Pseudomonas	11	1	4
Measles virus	10	3	5
Polyomavirus	8	3	11
Leukemia virus	7	1	1
Shigella	6	1	7
Plasmodium	6	2	2
Anemia virus	4	4	8
Hantaan virus	4	1	1
SARS	4	1	5
Listeria	4	1	3
Salmonella	3	1	4
Dengue virus	3	3	3
Seoul virus	3	1	1
Echovirus	3	2	2
Helicobacter	3	2	5
Rotavirus	3	3	3
Foamy virus	2	1	1
SIV	2	2	1
Dictyostelium	2	1	2
Puumala virus	2	1	1
Orf virus	2	2	1
Aeromonas	2	1	1
Stomatitis virus	2	1	1
Mycoplasma	2	1	1
Bothrops	2	1	1
Campylobacter	2	1	1
Vipera	1	1	1
Sendal virus	1	1	1
Pneumocystis	1	1	1
Corynephage	1	1	1
Nucleopolyhedrovirus	1	1	1
Candida	1	1	1
Rabies virus	1	1	1
Toxoplasma	1	1	1
Poliovirus	1	1	1
Nipah virus	1	1	1
Klebsiella	1	1	1
Enterobacteria	1	1	1
Mokola virus	1	1	1
West Nile virus	1	1	1
Tula virus	1	1	1
Ebolavirus	1	1	1
Total	20 113	206	1322

For each host–pathogen system, we list the number of known PPIs, the number of strains and the number of literature references.

information from UniProt (5), functional annotations for protein entries from the Gene Ontology (10) and functional domain data from InterProScan (11).

PIG currently focuses on known human–pathogen interactions. It contains 20 113 host–pathogen PPIs for 206 different pathogen strains collated from 1322 literature sources. A summary of the PPIs included in PIG can be seen in Table 1. Users can download the data stored in PIG as either a PostgreSQL dump or as tab-delimited files.

USING PIG OR THE PIG WEB INTERFACE

PIG contains a number of tools for identifying data of interest. First, PIG has two search tools: a simple text search (Figure 1a) and a BLASTP (12) interface (Figure 1b). The text interface allows users to search the database for key terms of interest (e.g. protein identifier or protein name). This search is executed against the protein IDs and descriptions stored in PIG. Valid results return links to protein-specific pages that contain information about the protein itself, such as functional annotations and domains, along with a list of interactions in which the protein participates (Figure 1c). Protein attributes such as ID, functional annotations and domains are hyperlinked to their respective external websites to allow easy navigation among different resources (Figure 1d). The BLASTP interface allows users to search entries in PIG that have sequence similarity to a protein of interest. Users can adjust BLAST search parameters. The search is executed using the BLASTP algorithm against all protein entries in PIG. Results are displayed in a standard fashion and each significant hit links back to a protein-specific page in PIG.

In addition to search tools, PIG also provides a platform for visualizing HP–PPI networks. Users can either visualize the network for a single host–pathogen system or for two host–pathogen systems. In both cases the user first selects a host system and a group of pathogens (Figure 1e). Each group entry expands to list-specific strains, thus allowing a user to compare closely related strains of a particular pathogen or pathogens from two different groups. Networks can be visualized directly within a Web browser using a custom applet (Java required), or downloaded as GraphML files for offline use (Figure 1f). Our custom applet provides full interactivity to the user, including zooming and panning, click-and-drag, and rich contextual information. The visualization of two HP–PPI networks allows users to identify human proteins that interact with both pathogens. These conserved interactors can provide critical insights into general strategies employed by pathogens during pathogenesis. The visualization tool also allows users to quickly identify functional annotations and domains for a selected protein in the network. For each of the protein's interactors, the supporting literature references are also displayed along with the technology used to identify the interaction. These features can be used to visualize subsets of the HP–PPI network (e.g. those interactions identified using a yeast two-hybrid approach) (Figure 1g–i). Each of these features also links to the

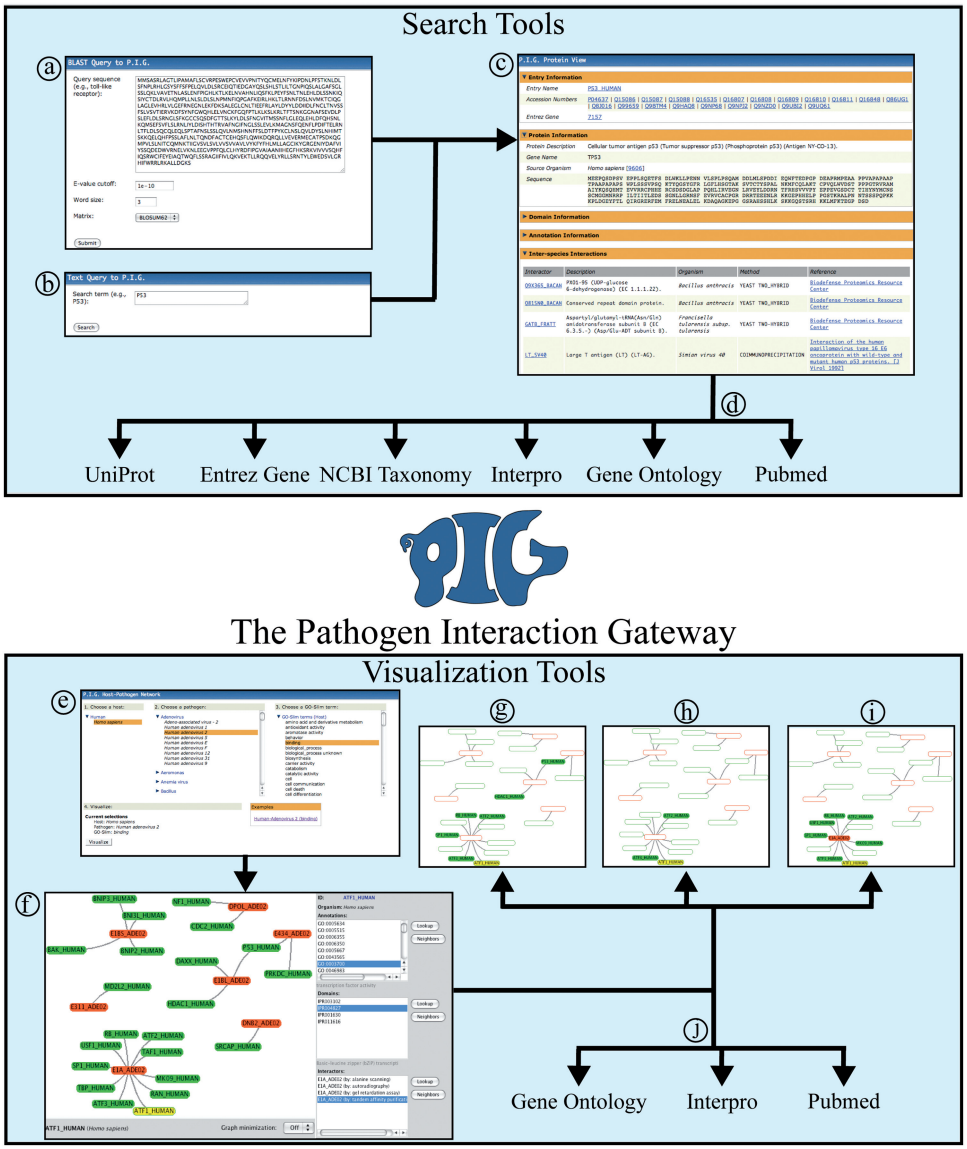

Figure 1. Summary of tools available on PIG. Users can search data within PIG using either (**a**) a simple text search or (**b**) a BLAST interface. Search results provide users with links to (**c**) individual protein pages. The individual protein pages contain information functional annotations, domains, and known inter-species and intra-species PPIs. (**d**) Each piece of information contains a direct link to the corresponding external database. (**e**) PIG also contains visualization tools. Users can select a host–pathogen system of interest and (**f**) view the corresponding network. Users can use the genomic information in PIG to view subsets of the networks using (**g**) domain, (**h**) annotation and (**i**) experimental method data. (**j**) Users can follow links from the visualization page to corresponding external databases.

corresponding external website (Figure 1j). Finally, users can restrict a network using GO-Slim terms (10) to focus on specific functional sub-networks.

CONCLUSIONS

PIG is an integrated resource that acts a centralized location for public information on known HP–PPI data. Each protein entry in PIG is hyperlinked to its corresponding entry in the UniProt database (5), functional annotations to the Gene Ontology (10), functional domains to InterProScan (11) and interactions to PubMed entries. These links allow for easy navigating among the various websites. PIG includes a number of tools for accessing available data including a simple text search, a blast interface and a tool for visualizing and comparing HP–PPI networks. Data stored in PIG are available for download on the website.

Future improvements of PIG will include the integration of additional data sources such as known virulence factors and toxins (9), tools for identification of therapeutic targets and computational methods to allow users to predict PPIs between any host and pathogen system of their choice. We believe that PIG will be a valuable resource for researchers of host–pathogen systems and possibly aid in the identification of potential targets for therapeutics.

ACKNOWLEDGEMENTS

The authors thank Clint Torres for the design of the PIG logo.

FUNDING

National Institute of Allergy and Infectious Diseases grant HHSN26620040035C. Funding for open access charge: NIH grant.

Conflict of interest statement. None declared.

REFERENCES

1. Gilbert,D. (2005) Biomolecular Interaction Network Database. *Brief. Bioinform.*, **6**, 194–198.
2. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
3. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
4. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
5. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
6. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
7. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
8. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
9. Zhou,C.E., Smith,J., Lam,M., Zemla,A., Dyer,M.D. and Slezak,T. (2007) MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.*, **35**, D391–D394.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
11. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.