BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 10 2013, pages 1357–1358
doi:10.1093/bioinformatics/btt137

Databases and ontologies

Advance Access publication March 20, 2013

PHISTO: pathogen–host interaction search tool

Saliha Durmuş Tekir[1,*], Tunahan Çakır[2], Emre Ardıç[3], Ali Semih Sayılırbaş[4], Gökhan Konuk[3], Mithat Konuk[3], Hasret Sarıyer[3], Azat Uğurlu[3], İlknur Karadeniz[5], Arzucan Özgür[5], Fatih Erdoğan Sevilgen[3] and Kutlu Ö. Ülgen[1]

[1]Department of Chemical Engineering, Boğaziçi University, 34342 Bebek, Sarıyer, İstanbul, Turkey, [2]Department of Bioengineering and [3]Department of Computer Engineering, Gebze Institute of Technology, 41400 Gebze, Kocaeli, Turkey, [4]Department of Biotechnology, Hamburg University of Applied Sciences, 21033 Bergedorf, Hamburg, Germany and [5]Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Sarıyer, İstanbul, Turkey

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Knowledge of pathogen–host protein interactions is required to better understand infection mechanisms. The pathogen–host interaction search tool (PHISTO) is a web-accessible platform that provides relevant information about pathogen–host interactions (PHIs). It enables access to the most up-to-date PHI data for all pathogen types for which experimentally verified protein interactions with human are available. The platform also offers integrated tools for visualization of PHI networks, graph–theoretical analysis of targeted human proteins, BLAST search and text mining for detecting missing experimental methods. PHISTO will facilitate PHI studies that provide potential therapeutic targets for infectious diseases.

Availability: http://www.phisto.org.

Contact: saliha.durmus@boun.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 27, 2012; revised on March 14, 2013; accepted on March 15, 2013

1 INTRODUCTION

The interactions between the proteins of infectious microorganisms, pathogens and their human hosts allow the microorganisms to manipulate human cellular mechanisms to their own advantage, resulting in infection in the host organism. The recent advances in high-throughput protein interaction detection methods have led to the production of large-scale interspecies protein–protein interaction (PPI) data of pathogen–human systems. Currently, there are a number of pathogen–host interaction (PHI) resources that are specific to some pathogens. The only available resource to access all PHI data in a single database (Kumar and Nanduri, 2010) does not offer any additional functionality to analyze PHI networks. We have developed pathogen–host interaction search tool (PHISTO) to serve as an up-to-date and functionally enhanced source of PHI data through a user-friendly interface. PHIs in PHISTO are imported from several PPI databases using the PSICQUIC tool (Aranda *et al.*, 2011). Text mining is used to label PHIs extracted without any information on interaction detection method. Tools for visualization of small PHI networks and graph–theoretical analysis of targeted human proteins may enable users to gain crucial insights on infection mechanisms. The BLAST interface offers to search for orthologous PHIs for pathogens lacking experimental data.

*To whom correspondence should be addressed.

2 ARCHITECTURE

PHISTO is designed as a Web-accessible platform with two-tier architecture. The back tier is a MySQL-based database. The front tier is a PHP- and Javascript-based user interface that runs on an Apache Web server. Figure 1 illustrates the architecture of PHISTO.

2.1 Database

The PSICQUIC tool is used to extract PHI data from nine databases (see Section A in Supplementary Material). To present protein interaction data in a consistent format, Uniprot IDs and names of interacting proteins, taxonomy IDs and names of pathogens, experimental methods and PubMed IDs of literature references are collected; the data are stored in separate database tables, with one PHI table containing the core data (UniProt IDs, Taxonomy ID, Experimental Method, PubMed ID) and the others being mostly ID-related tables (see Section B in Supplementary Material). Thanks to the implementation of ID-based core data, standardized data formats and relational data tables, the PHISTO database can easily be maintained and enhanced. It is automatically updated by our Java-based offline application on a monthly basis. Currently, PHISTO stores data on 23 661 PHIs between human and 300 pathogen strains (247 viral, 45 bacterial, 3 fungal and 5 protozoan). Among the PHI data extracted from the PPI databases, there were 12 751 PHI data that were not labeled with the experimental methods used to detect these interactions. PHISTO contains a text mining module for experimental method extraction for such data. A dictionary of interaction detection methods is compiled from the PSI-MI ontology. The abstracts of the articles that contain PHIs without experimental method information are obtained from PubMed. An exact string matching-based approach was used to assign 2952 experimental method names to 2109 unique PHIs. The experimental method detection module was evaluated by using the PHIs with experimental method information in PHISTO. The module achieved a promising precision of 74%. The recall and the F-score of the module are 34 and 47%, respectively (see Section C in Supplementary Material).

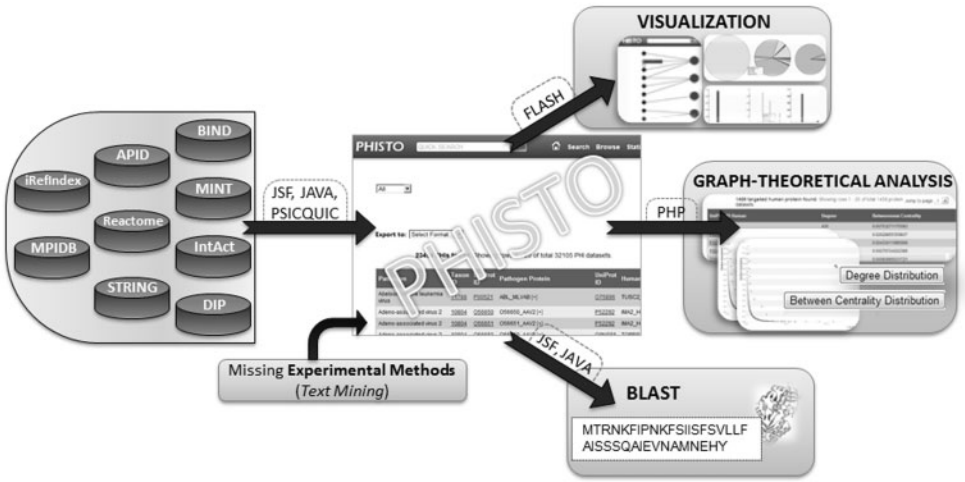

Fig. 1. The architecture of PHISTO (See Supplementary Material for details)

2.2 User interface

The functional and easy-to-use Web-based user interface provides various search, browse and data analysis options. The 'Quick Search' option is for performing a query without a specified identifier, whereas the 'Advanced Search' option can be used to search based on any selected subset of identifiers (i.e. taxonomy ID and name of pathogen, Uniprot ID and name of pathogen protein, Uniprot ID and name of human protein, experimental method and literature reference). The 'Browse' option provides users with easy access to the entire PHI data of any specified pathogen within the taxonomical classification. The PHI search results are presented in a clear and consistent ID-based formatted table, which includes information about the eight identifiers mentioned previously. Proteins, pathogens and publications listed in the results are linked to external databases UniProt, NCBI Taxonomy and PubMed, respectively, offering users quick navigation in these informative databases. The search results can be exported to a variety of file formats. The visualization option allows the resulting PHI network to be represented as a bipartite graph. Statistics of search results can be visualized through pie or bar charts, describing the distribution of PHIs over the types, families, species and strains of pathogens. For a protein of interest, one can search for homolog pathogen proteins in PHISTO using the 'BLAST' interface (see Section D in Supplementary Material). Finally, the 'Graph Analysis' interface provides the degree and betweenness centrality distributions of the targeted human proteins compared with all proteins within the human intranetwork (see Section E in Supplementary Material).

3 DISCUSSION

With regular monthly data updates and complete coverage of all data available for each pathogen type, PHISTO will always provide unified access to up-to-date PHI data. Otherwise, gathering such information would require careful and tedious data integration. To our knowledge, PHISTO is the first PHI tool that uses text mining to identify the interaction detection methods of PHIs that have not been originally annotated with such information. Additionally, PHI results visualized as a bipartite graph allow users to capture PHI mechanisms (i.e. most connected pathogen/host proteins can be easily observed at a glance). The sequence search option, BLAST, makes it possible to retrieve homologous PHIs. After homolog pathogen proteins in PHISTO are found, interacting human protein partners of the homolog proteins may be assigned as putative interactors for the protein under investigation. Finally, the distributions of degree and betweenness centrality values in the graph analysis tool give crucial insights on attacking strategies of pathogens during infections (Dyer *et al.*, 2008; Durmuş Tekir *et al.*, 2012). Our goal with PHISTO is to provide a centralized and up-to-date platform for studying pathogen–host protein interaction systems with future curation of PHIs from literature by text mining and additional advanced analysis tools for PHI networks.

Funding: Boğaziçi University Research Funds Project 5554D, Marie Curie FP7-Reintegration-Grants, Ekin Kimya Tic. Ltd. Şti. and Elginkan Foundation.

Conflict of Interest: none declared.

REFERENCES

Aranda,B. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.

Durmuş Tekir,S. *et al.* (2012) Infection strategies of bacterial and viral pathogens through pathogen-human protein-protein interactions. *Front. Microbiol.*, **3**, 46.

Dyer,M.D. *et al.* (2008) Landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.*, **4**, e32.

Kumar,R. and Nanduri,B. (2010) HPIDB – a unified resource for host-pathogen interactions. *BMC Bioinf.*, **11**, S16.