BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 14 2011, pages 2015–2017
doi:10.1093/bioinformatics/btr328

Data and text mining

Advance Access publication June 2, 2011

LinkinPath: from sequence to interconnected pathway

Supawadee Ingsriswang*, Sunai Yokwai and Duangdao Wichadakul

Information Systems Laboratory (ISL), Bioresources Technology Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), Klong Luang, Pathumthani 12120, Thailand

Associate Editor: Jonathan Wren

ABSTRACT

Summary: LinkinPath is a pathway mapping and analysis tool that enables users to explore and visualize the list of gene/protein sequences through various Flash-driven interactive web interfaces including KEGG pathway maps, functional composition maps (TreeMaps), molecular interaction/reaction networks and pathway-to-pathway networks. Users can submit single or multiple datasets of gene/protein sequences to LinkinPath to (i) determine the co-occurrence and co-absence of genes/proteins on animated KEGG pathway maps; (ii) compare functional compositions within and among the datasets using TreeMaps; (iii) analyze the statistically enriched pathways across the datasets; (iv) build the pathway-to-pathway networks for each dataset; (v) explore potential interaction/reaction paths between pathways; and (vi) identify common pathway-to-pathway networks across the datasets.

Availability: LinkinPath is freely available to all interested users at http://www.biotec.or.th/isl/linkinpath/.

Contact: supawadee@biotec.or.th

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

Received on March 21, 2011; revised on May 9, 2011; accepted on May 25, 2011

1 INTRODUCTION

Pathways are functional units resulted from the interplay of interacting genes, RNAs, proteins and small molecules. Mapping genes or proteins into the context of pathways can help gain more insights into their functions and interactions in an organism. Although sequence similarity-based methods [e.g. NCBI BLAST (Altschul *et al.*, 1990)] have been commonly used for identification of pathways to genes/proteins based on their orthologous genes/proteins annotated in the well-characterized pathways, these methods have limitations such that some best hits (such as those annotated as ‘hypothetical’ or ‘unknown’ genes/proteins) may not necessarily be annotated in any pathway. The incorporation of additional data such as protein–protein interactions and enzymatic reactions can help infer the pathways and their interconnection and uncover the biological function of genes/proteins. However, the information regarding the connections between pathways through molecular interactions and reactions are not included and adequately represented to support the exploratory analysis in most pathway mapping tools (reviewed in Gehlenborg *et al.*, 2010), for example, GenMAPP (Salomonis *et al.*, 2007), Pathway Explorer (Mlecnik *et al.*, 2005) and Pathway Projector (Kono *et al.*, 2009). To overcome these limitations, a web-based interactive tool, so-called LinkinPath, was developed to analyze, map and visualize the gene/protein lists in the context of interconnected pathways, which provides a valuable resource for not only comprehensive studies of gene–gene interaction, but also functional genomics in virtually all organisms.

2 METHOD

LinkinPath has been developed as a web-based interactive exploration tool for pathway analysis. Users can upload the datasets of DNA or protein sequences in FASTA format and submit to the LinkinPath’s job queue (Supplementary Fig. S1), which could support the analyses of genome-scale inputs. Although, many jobs may be submitted at the same time, the LinkinPath web server accepts a total maximum of 5000 sequences for each job. When jobs are finished, users can retrieve the result using the bookmarked URL during the submission or the web link from the notification email. LinkinPath processes each job in five main steps: (i) sequence and protein domain search; (ii) pathway mapping and annotation; (iii) identification and comparison of enriched pathways; (iv) construction of interconnected pathway networks and (v) identification of common pathway subnetworks (Supplementary Fig. S2). First, input sequences are searched against KEGG (Kanehisa and Goto, 2000), NR (Benson *et al.*, 2007), PFAM (Finn *et al.*, 2008) and RFAM (Griffiths-Jones *et al.*, 2003) databases. Second, if significant similarities are found to match with an enzyme class (EC) and/or KEGG Orthology (KO), genes/proteins will be annotated with the corresponding pathways. Third, to support comparison of functional composition, Treemaps and statistical methods are employed to examine the pathways enriched in the dataset. Fourth, information of molecular interactions and reactions from BIND (Bader *et al.,* 2003) and KEGG will be used for inference of pathways and networks. In case that an input sequence cannot be annotated in any pathway, LinkinPath will use its interacting partners to infer its related pathways. The pathways will thus be linked together via interaction or reaction paths to form the network in this step. Lastly, to identify the commonalities across multiple datasets, LinkinPath includes an algorithm to extract the frequent subnetworks from the pathway networks built in previous step.

3 RESULT

Using the method described above, LinkinPath automatically maps and annotates genes/proteins in the datasets into the context of interconnected pathways and presents the results to users via a Flash-driven interactive web interfaces. The results are organized into five analysis steps (see following subsections), which can easily be browsed, searched and downloaded in any of supported formats. Summary charts and an interactive Venn diagram are also provided to illustrate of how a dataset differs from others according to their annotated sequences with EC numbers, interactions and pathways (Supplementary Fig. S3). The annotation results of input sequences can be interactively explored in different visualization perspectives

*To whom correspondence should be addressed.

including KEGG-pathway maps, TreeMaps, pathway-to-pathway networks and interaction and reaction paths.

3.1 Detecting changes in animated pathway maps

In most tools, the KEGG pathway image is statically displayed with positioned genes/proteins in the pathway. To enable the dynamicity, LinkinPath allows users to run the pathway map animation to depict changes of expressed genes/proteins annotated in each pathway over multiple and time-series datasets (Supplementary Figs 4 and 5). Detecting changes between the same pathway maps taken from different times/experiments/organisms can help reveal the functional characterization of genes/proteins expressed under different conditions. In addition, the co-occurrence and co-absence of genes/proteins in a series of pathways can suggest the functional relationships between these molecules. Alternatively, users can interactively explore the annotated pathways across datasets on KEGG Atlas, in which connected paths are colored according to the datasets.

3.2 Visualizing functional composition via TreeMaps

Many classification schemes such as enzyme classification, KO and KEGG pathways have tree-based structures that are difficult to simultaneously display and compare the information contained in the trees. To circumvent this problem, Treemaps are included in LinkinPath to facilitate the visualization and comparison of those hierarchical data via a set of nested rectangular maps. A base rectangle represents the root of the hierarchy and is divided into rectangular subareas proportional to data size and colored by data type. As a result, Treemaps enable users to compare sizes of nodes and subtrees, and are helpful in revealing patterns. To disclose the functional composition among datasets, LinkinPath provides users with three different Treemaps: (i) the enzyme compositions using top-level EC numbers; (ii) the pathway compositions using KEGG pathway classification; and (iii) and KO compositions using KO numbers (Supplementary Fig. S6). The Treemap of enzyme composition, for example, helps users visually examine enzyme enrichment in lists of genes/proteins, where each rectangle represents the top-level EC numbers with different colors and indicates the proportion of genes/proteins annotated with the EC group. Alternatively, LinkinPath allows users to compare and examine the enrichments of the EC and KO annotations of entire input data across pathways.

3.3 Identification of statistically enriched pathways

Since pathways that are highly enriched with the list of annotated genes/proteins are more likely to be biologically relevant, LinkinPath employs KOBAS (Wu *et al.*, 2006) to help identify significantly enriched pathways in a dataset. KOBAS uses KO number to link genes/proteins to KEGG pathways and calculates the statistical significance of each pathway in a queried dataset against all pathways in the referenced datasets. There are three statistical tests available including binominal, chi-square and hyper-geometric distribution tests to assess the enrichment of the found pathways. For each pathway found in the input datasets, LinkinPath calculates these statistics by comparing the number of sequences involved in the pathway for each dataset with the total number of sequences involved in the same pathway for all datasets.

3.4 Inferring pathway-to-pathway interconnections

Despite their complexity, the interconnections between pathways can help unravel novel regulation mechanisms including metabolism and signaling in organisms. LinkinPath infers the pathway-to-pathway connections using molecular interactions and reactions. Pathways will be connected to each other and represented in a form of interactive networks. Each dataset could have several pathway networks with different sizes. Each node represents either a pathway or an input sequence. The pathway node contains the information on the number of input sequences annotated in that pathway. An edge between nodes indicates the pathway connection types with different colors and line styles. Two pathway nodes are connected with a solid line if a path between them exists in KEGG database. The dash line indicates an inferred path from an input sequence or a query node to a known pathway. With its interactive network browser, LinkinPath allows users to browse and access the network characterization and the associated information such as the number of input sequences and interactions involved in a pathway node and the list of input sequences that appear in single or multiple pathways.

Exploring putative functions of genes/proteins: LinkinPath utilizes molecular interactions to infer the pathway and putative function of an un-annotated sequence on the basis of its interacting partner's function. The shortest interaction paths connecting from a query node to a pathway node in the network are identified using all paths breadth first search. In addition, the putative function of an input sequence might be inferred by its proximity to functionally annotated genes/proteins within the context of interconnected pathways. In the network browser, users can traverse the interaction paths to the inferred pathways of the nodes with dashed edge (Supplementary Fig. S7).

Exploring reactions between metabolic pathways: LinkinPath searches the reaction paths from an input sequence mapped in a metabolic pathway to the compounds linking to other metabolic pathways. On a solid blue edge in the pathway-to-pathway network, if the reaction paths exist, users can explore what enzymes and compounds are essential to a metabolic process via the reaction paths between two pathways (Supplementary Fig. S8).

3.5 Discovering the common pathway subnetworks

LinkinPath extracts the frequently occurred subgraphs /subnetworks from the pathway networks to discover the commonalities across the datasets (Supplementary Fig. S9). Users can navigate and visualize frequent subnetworks of varied sizes that occur in a number of different datasets.

4 CONCLUSION

LinkinPath is a web-based tool that was applied the state of the art visualization techniques for original aspects of pathway mapping and analyses. Its novel contributions such as functional composition using Treemaps, pathway-to-pathway interconnections and the global metabolic map with highlighting mechanisms add value over comparable existing tools.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Anan Jongkaewwattana for critically reading the manuscript and a former ISL staff, Eakasit

Pachawongsakda, who helped in programming during the initial stage of the project.

Funding: National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bader,G.D. *et al.* (2003) BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.

Benson,D.A. *et al.* (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.

Finn,R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Gehlenborg,N. *et al.* (2010) Visualization of omics data for systems biology. *Nature Methods*, **7**, S56–S68.

Griffiths-Jones,S. *et al.* (2003) An RNA family database. *Nucleic Acids Res.*, **31** 439–441.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kono,N. *et al.* (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One*, **4**, e7710.

Mlecnik,B. *et al.* (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, **33**, W633–W637.

Salomonis,N. *et al.* (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.

Wu,J. *et al.* (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.