BEReX: Biomedical Entity-Relationship eXplorer

Minji Jeon[1], Sunwon Lee[2], Kyubum Lee[2], Aik-Choon Tan[3] and Jaewoo Kang[1,2,*]

[1]Interdisciplinary Graduate Program in Bioinformatics, [2]Department of Computer Science and Engineering, Korea University, Seoul, Korea and [3]Department of Medicine/Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Associate Editor: Mario Albrecht

ABSTRACT

Summary: Biomedical Entity-Relationship eXplorer (BEReX) is a new biomedical knowledge integration, search and exploration tool. BEReX integrates eight popular databases (STRING, DrugBank, KEGG, PhamGKB, BioGRID, GO, HPRD and MSigDB) and delineates an integrated network by combining the information available from these databases. Users search the integrated network by entering key words, and BEReX returns a sub-network matching the key words. The resulting graph can be explored interactively. BEReX allows users to find the shortest paths between two remote nodes, find the most relevant drugs, diseases, pathways and so on related to the current network, expand the network by particular types of entities and relations and modify the network by removing or adding selected nodes. BEReX is implemented as a standalone Java application.

Availability and implementation: BEReX and a detailed user guide are available for download at our project Web site (http://infos.korea.ac.kr/berex).

Contact: kangj@korea.ac.kr

Supplementary Information: Supplementary methods and user guide are available at *Bioinformatics* online.

1 INTRODUCTION

There is a good deal of biomedical databases, many of which become community resources. However, each database has its own limitations. First, each database covers only a small vertical domain. For example, STRING (Franceschini *et al*., 2013) covers only protein–protein interactions, DrugBank (Knox *et al*., 2011) and PhamGKB (Whirl-Carrillo *et al*., 2012) mainly focus on drugs and diseases, STITCH (Kuhn *et al*., 2012) covers only proteins and chemicals and KEGG (Kanehisa *et al*., 2012) covers pathways but the coverage is small. To comprehend a complex relation spanning over the different domains, users have to manually integrate the pieces of information obtained from each database. Second, their query capabilities are limited. For example, PharmGKB and DrugBank provide only a single result per query, and thus users have to manually combine the results from multiple queries to understand the interactions. Last, the existing databases produce a static query result, and thus users are not able to customize the result.

To address these problems, we developed a user-friendly database integration, search and exploration tool called BEReX (Biomedical Entity-Relationship eXplorer). BEReX has the following advantages:

(1) Reliable information: data in BEReX is reliable because BEReX integrates data from widely used databases such as STRING,

KEGG, PharmGKB, DrugBank, BioGRID (Chatr-Aryamontri *et al*., 2013), GO (Ashburner *et al*., 2000), HPRD (Prasad *et al*., 2009) and MSigDB (Liberzon *et al*., 2011), which are proven for their utility.

(2) Integrated network: various datasets are gathered to delineate an integrated network in BEReX. Users can see direct and indirect relations between various types of biomedical objects drawn from different databases.

(3) Interactive exploration: BEReX visualizes the results from a query as a graph. Users can interactively navigate through the relations and expand/modify the network. For example, users can find the shortest paths between two remote nodes; find the most relevant drugs, diseases, pathways and so on related to the current network; expand the network by particular types of entities and relations; and modify the network by removing or adding selected nodes.

(4) Generating new hypothesis: BEReX provides an environment for allowing a user-driven approach to automatically generate plausible and/or unexpected relationships from the derived integrated network. These plausible relationships describe potential interactions or mechanisms of action between query entities and may be used to generate a new testable hypothesis between query genes.

2 BEREX DESCRIPTION

2.1 Database integration

BEReX integrates human biomedical entity-relation information from the eight datasets to delineate the integrated network, which consists of 46 118 entities and 1 126 033 relations. An entity has 24.42 (standard deviation 239.12) relations on average. A detailed explanation of the datasets is given in the Supplementary Methods.

2.2 User interface

Figure 1 shows the BEReX user interface. Users can query BEReX for any combination of genes, drugs, diseases, pathways, microRNAs, transcription factors and gene ontology terms by using keywords. BEReX shows the relationships among the entities matching the query keywords. The tree view window at the top right shows the entity-relationship information of a selected node. The window below that shows the details of a selected node or edge and provides links to external sources such as PubMed articles that support the relation. The window at the bottom left allows users to filter information by particular entity types, relation types and data sources. The BEReX exploration window shows the query results in a graph. Users can interactively expand/modify the graph to construct custom pathway networks of their interest. Finally, users can save the current graph to a file for later use or export the graph as an image through the file menu at the top.

2.3 Ranking method

BEReX evaluates a user query by following three steps: (i) finding the matching (query) nodes, (ii) adding the shortest paths between the query
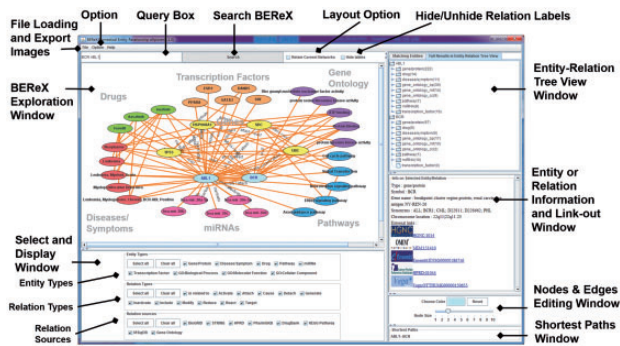
*To whom correspondence should be addressed.

Fig. 1. BEReX user interface

nodes and finally (iii) finding some extra nodes that are most relevant to the query and augmenting the current network with them. BEReX determines the degree of relevance based on Google's PageRank (Page *et al.*, 1999). In PageRank, a node's importance improves as the number of neighbors and the importance of the neighbors increase. Given a query, BEReX evaluates the PageRank scores of all neighboring nodes and shows only the top scoring ones in the result. A detailed explanation of the BEReX ranking algorithm is given in the Supplementary Methods.

2.4 Implementation

BEReX is implemented in Java, using JUNG (Java Universal Network/ Graph Framework) (http://jung.sourceforge.net) for graph visualization and Neo4j (http://www.neo4j.org) for graph storage and retrieval. BEReX runs on Windows, Mac and Linux operating systems with Java Runtime Environment 6 or later. A Cytoscape-plugin (Saito *et al.*, 2012) version of BEReX also will be available soon on our project Web site.

3 A USE CASE: BCR AND ABL1

The graph in Figure 1 is constructed by expanding the query result of 'BCR' and 'ABL1'. For the query 'BCR' and 'ABL1', BEReX returns a six node graph including the two query nodes (BCR and ABL1) and four other gene nodes (TP53, UBC, SRC and HSP90AA1). BCR and ABL1 have a direct interaction, as these genes are fused and drive the oncogenesis of chronic myeloid leukemia (CML). The relationship between CML and the two query genes can be found by expanding the query genes with the diseases/symptoms relationship option (see user guide). Next, by expanding the query with drugs, three components are revealed including imatinib, dasatinib and fasudil, which interact with BCR and ABL1. Imatinib is the first clinically approved drug to treat CML, as it inhibits the kinase activity of the BCR-ABL1 gene fusion (Druker *et al.*, 2001). Dasatinib has been approved to treat CML patients that are resistant to imatinib (Talpaz *et al.*, 2006).

The interactions of fasudil, a Rho-kinase inhibitor, with BCR and ABL1 were unexpected. By clicking the edge between fasudil and ABL1, one can see the information supporting this relationship. A recent study identified Rho-kinase as the key player in regulating the survival and transformation of BCR-ABL1 oncogenes, and showed that treating CML cell lines with fasudil inhibits the cancer cell proliferation (Mali *et al.*, 2011). Furthermore, another recent article also demonstrated synergistic effects in combining imatinib with fasudil in CML cell lines (Burthem *et al.*, 2007). This example suggests that the PageRank-based ranking algorithm implemented in BEReX could produce biologically relevant results and that the user-driven approach (same as aforementioned) could generate a testable hypothesis. The applicability of PageRank to biological network analysis also has been reported independently by Chen *et al.* (2009). They showed how PageRank can be used to prioritize disease candidate genes. Finally, by repeating the same process for different entity types, a custom graph highlighting the relationships between BCR and ABL1 with various biomedical entities would be generated as illustrated in Figure 1.

4 CONCLUSION

We developed BEReX, a user-friendly and interactive biomedical entity-relationship exploration tool, for navigating the biological knowledge integrated from eight well-known databases. BEReX extracts the entity-relationships from these databases and builds a comprehensive integrated network. By using the PageRank-based ranking algorithm, users can incrementally build their own custom networks through the interactive user interface. We believe that BEReX will be a useful bioinformatics tool for biologists in exploring the complex biomedical entity-relationship networks.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Burthem,J. *et al.* (2007) The ρ-kinase inhibitors Y-27632 and fasudil act synergistically with imatinib to inhibit the expansion of ex vivo CD34(+) CML progenitor cells. *Leukemia*, **21**, 1708–1714.

Chatr-Aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.

Chen,J. *et al.* (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 73.

Druker,B.J. *et al.* (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.*, **344**, 1031–1037.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D1035–D1041.

Kuhn,M. *et al.* (2012) STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res.*, **40**, D876–D880.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Mali,R.S. *et al.* (2011) Rho kinase regulates the survival and transformation of cells bearing oncogenic forms of KIT, FLT3, and BCR-ABL. *Cancer Cell*, **20**, 357–369.

Page,L. *et al.* (1999) The PageRank citation ranking: bringing order to the web#44# Technical Report, 1999-66, Stanford InfoLab, http://ilpubs.stanford.edu:8090/422/ (8 November 2013, date last accessed).

Prasad,T.K. *et al.* (2009) Human protein reference database - 2009 update. *Nucleic Acids Res.*, **37** (**Suppl. 1**), D767–D772.

Saito,R. *et al.* (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.

Talpaz,M. *et al.* (2006) Dasatinib in imatinib-resistant Philadelphia chromosome–positive leukemias. *N. Engl. J. Med.*, **354**, 2531–2541.

Whirl-Carrillo,M. *et al.* (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, **92**, 414–417.