A Diagram Editor for Efficient Biomedical Knowledge Capture and Integration

Bohua Yu, M.S.¹, Elvis Jakupovic, B.S.¹, Justin Wilson, B.S.¹, Manhong Dai, B.S.¹, Weijian Xuan, M.S.¹, Barbara Mirel, PhD², Brian Athey, PhD¹, Stanley Watson, MD, PhD¹ and Fan Meng, PhD¹*

¹Psychiatry Department and Molecular and Behavioral Neuroscience Institute, ²School of Education, University of Michigan, Ann Arbor, MI 48109

Abstract
Understanding the molecular mechanisms underlying complex disorders requires the integration of data and knowledge from different sources including free text literature and various biomedical databases. To facilitate this process, we created the Biomedical Concept Diagram Editor (BCDE) to help researchers distill knowledge from data and literature and aid the process of hypothesis development. A key feature of BCDE is the ability to capture information with a simple drag-and-drop. This is a vast improvement over manual methods of knowledge and data recording and greatly increases the efficiency of the biomedical researcher. BCDE also provides a unique concept matching function to enforce consistent terminology, which enables conceptual relationships deposited by different researchers in the BCDE database to be mined and integrated for intelligible and useful results. We hope BCDE will promote the sharing and integration of knowledge from different researchers for effective hypothesis development.

Introduction
An important goal of translational bioinformatics is the development of methods to “optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health” ¹. However, understanding how genomic analysis results are related to molecular mechanisms underlying complex disorders is a major challenge. Under most situations, genes or SNP alleles identified in genomic studies have no known relationship with the targeted pathophysiological process. Researchers have to combine information from different sources together with their background knowledge in order to develop hypotheses about the potential molecular mechanisms underlying the disease in question.

Since no database contains all the potential functional interpretations of genomic analysis results, developing hypotheses that link genomic events to diseases is an almost exclusively manual process involving multiple databases as well as free text literature. We believe hypothesis development tools that can help researchers efficiently capture, annotate, share, integrate and mine molecular level changes and their relationships to diseases need to be developed.

Solutions that match the above requirements the closest are biological pathway drawing programs such as: BioDraw ², Protein Lounge ³, PathWiz ⁴ and PATIKA ⁵,⁶. Researchers can use these tools to produce presentation quality pathway diagrams containing beautiful, pre-made icons and stencils together with extensive manual annotations.

There are several major problems with existing solutions. The first problem is that it is time-consuming to convert information from database or web search results into a conceptual diagram relating molecular events to diseases. The process of selecting, naming and annotating objects in a diagram is tedious and must be performed by hand. Consequently, most researchers still use word processors or note-taking programs such as OneNote due to their better recording efficiency. However, these programs lack built-in support for biomedical concepts. The second problem is that it is not easy to share and integrate diagrams because of inconsistent terminology among different researchers or even the same researcher at different time. None of the existing solutions have the capability of mapping different expressions of the same concept to each other or identifying semantic relationships between different concepts. Another problem is that existing diagram editors are designed to generate diagrams for publication and presentation purposes. They lack functions for searching, comparing and integrating related diagrams from different researchers.

In this paper, we describe a novel biomedical concept diagram editor (BCDE) that was designed for the generation of molecular level hypotheses about diseases etiology by integrating functions for the retrieval of functional information about SNPs, genes

* Corresponding author
and biological processes from various data sources, sharing diagrams with other researchers using a consistent concept matching scheme and eventually searching existing diagrams for latent conceptual relationships.

**BCDE Overview**

The BCDE application is a Java program based on the open source JHotDraw v5.3 framework. We choose Java because it is platform independent.

**Standard diagramming functions:** The BCDE application is similar to other diagram/flowchart programs in terms of standard diagramming functions such as Undo, Redo, Copy, Cut and Paste. It has a stencil of 16 icons, each of which represents a family of biological concepts. Users can rotate, scale and color icons as well as add annotation and attachments through a property dialog box. There are also different types of arrows for linking nodes and text tools for nodes, links and the canvas.

Objects and diagrams in BCDE have a variety of associated properties and they are stored in the BCDE XML format. This format is a superset of the BioPAX Level 2 format. We plan to support BioPAX Level 3 when it is released.

**BCDE data capture function:** Efficient data capture is realized by dragging concepts and related data (text, image or a mixture) displayed in Internet Explorer and dropping in the BCDE diagramming canvas. Internet Explorer is capable of communicating with BCDE through a local port. BCDE runs an internal Jetty web server, which is able to receive input from a local port accessible only from the same computer. Such a configuration minimizes security risks. The ability to drag-and-drop sets BCDE apart from other applications. We will show an example of data capture later.

**BCDE concept match and spell check web service:** BCDE’s annotation function is integrated with our concept match web service to provide consistent terminology and remove spelling errors.

The concept match service uses a radix tree-based algorithm to identify similar concept strings, as the pronunciation-based algorithms are not suitable for chemical and biochemical names that include unpronounceable elements such as digits, dashes and parentheses. Nearly 30 million text string variations corresponding to about 18 million concepts from our gene/protein synonyms list, dbSNP, PubChem, UMLS and WordNet are behind this service.

A web demo of this function is available at http://arrayanalysis.mbni.med.umich.edu/concept/. It will map a user input string to a list of similar strings in our concept database and list unique concepts that correspond to each string. A user can then select the correct concept based on the detailed description and the source for each concept. Spelling errors can also be corrected when choosing the correct concept.

Since a fully automatic concept mapping method is not yet possible due to context–dependent semantics and limitations in current algorithms and databases that deal with lexical variations, we resort to the practical solution of user selection in the mapping of various biomedical terms to unique biomedical concepts in major databases. This interactive approach solves most of the problems of context-dependency since a researcher’s knowledge is utilized in the concept mapping process. For example, a user can decide that “REG” means “Reg receptor” rather than “islet cell regeneration factor” or “regression model” during the diagramming process instead of having an algorithm guessing the researcher’s intentions.

**BCDE object window:** To increase the efficiency of annotating diagrams, we built a BCDE object window that is always on top of all the canvases. The BCDE object window is used to display all BCDE objects on the current active canvas and their annotation status (“complete”, “incomplete” and “satisfactory”). Figure 1 is a screenshot of the BCDE object window (the small window on the right side) containing the

Figure 1. BCDE annotation through object window

term “Opioid receptor” (highlighted). The larger window on the left side displays matched concepts, including the correct concept “Opioid receptor”, from our concept match and spell check service.

Utilizing the BCDE object window for batch annotation after a diagram is generated is the preferred way for annotating BCDE objects. While
individual object annotation through the object property box could enforce a consistent vocabulary at any point in the diagramming process, the related operations are usually disruptive to the researcher’s thinking process.

**BCDE import/export functions:** BCDE can import diagrams in the BioPAX Level 2 format for display, modification and storage in the BCDE XML database described in the next section. We use a tool called JDOM to parse the BioPax OWL or XML file. Once the BioPAX file is parsed, the application arranges the physical entities and interaction nodes by applying a force-directed placement algorithm. The end result is a diagram that occupies the whole canvas with minimum crossing connections.

BCDE currently supports the export of diagrams into three formats: JPEG image, Microsoft Visio and PowerPoint. BCDE automatically converts the related annotations and attachments to properties and links in Visio and PowerPoint. As a result, researchers can use BCDE exported PowerPoint slides with full links and annotation.

**BCDE XML database:** The BCDE XML database is an Oracle XML DB with two main functions: to provide conceptual relationships for the RDF database and to be central storage for all the diagrams and their display information (location, color, size, icon, etc.) created using the BCDE application or imported from external databases. The XML database is also the place for researchers to share their knowledge and ideas through BCDE diagrams since it offers concept-based search functionality for diagram retrieval. Knowledge integration and conceptual relationship-based searches are the functions of the BCDE RDF database.

**BCDE RDF database:** The lynchpin that links all of our tools and data together is the BCDE RDF database. It is an Oracle 10g R2 database using the Oracle Spatial Network Data Model (NDM) on top of a relational database (Murray, 2005). Our RDF database contains triples representing biological interactions in the form of “element-interaction-product”. These triples are automatically extracted from diagrams stored in the BCDE XML database. All RDF triples in the database must be verified by the concept matching web service. Each element in the triple must contain our unique URI and end with a valid BCDE ID from the concept matching web service. By using BCDE IDs as identifiers for unique concepts, we can essentially create a directed graph of all the information in our database; thus forming something we call the “pathway-interaction continuum”. With it, more complex query algorithms can be used to explore conceptual relationships. For example, one can query all the concepts in a triple with SNP rs958247. By looking at those adjacent concepts, a researcher might be able to form new hypothesis linking rs958247 to a disease.

By combining the XML and the RDF databases, we can make hypothesis development an interactive experience. A researcher can search the XML database for a concept and link back to the RDF database to get a bird’s eye view of how that concept fits into the pathway-interaction continuum. On the other hand, he/she may also query the RDF database and examine the context of specific interactions using diagrams stored in the XML database.

Using a Java Server Page (JSP), a user can upload his or her BCDE diagrams in BCDE XML format into the database. Each diagram will be validated against the BCDE XML schema during the upload process to ensure that it is a valid BCDE diagram before accepting it into the database.

All interactions between the user and the database happen through a series of JSP web pages. Therefore, the user is never directly connected to the database. This setup eliminates several security issues such as direct database hacking and unauthorized manipulations of the database.

**BCDE account management:** We use the Lightweight Directory Access Protocol (LDAP) for all BCDE functions and resources that require access control. LDAP offers a centralized and flexible method for administering literally millions of user accounts. The security model of LDAP is flexible for both maintenance and machine access. Furthermore, LDAP-enabled directories are a useful place to store information about users' preferences which allows them to personalize their web environment.

**BCDE availability:** To download the BCDE application, users can go to our web site at [http://brainarray.mbnimed.umich.edu/Brainarray/default.asp](http://brainarray.mbnimed.umich.edu/Brainarray/default.asp) and then select “Graphic Editor” under the “Data Mining” menu. There are several Flash demos showing various BCDE functions on the same page.

**Result and Example**

As described, the current version of BCDE already incorporates a number of unique functions that greatly increase the efficiency of creating well-annotated biomedical concept diagrams. We will illustrate some of these unique functions here.

**Capture of an image from a webpage:** The BCDE installation package downloaded from our site will
install the BCDE capture function as a plug-in for the Internet Explorer. A user can simply drag an image from a web page and drop it on the BCDEBar in the Internet Explorer toolbar.

Figure 2. The BCDE capture function

Figure 2 is a screenshot showing the drag-and-drop capture function and the automatic generation of a generic icon on the BCDE canvas. The drag-and-drop area on the BCDEBar plug-in is labeled “Drop text, images, or files, create new BCDE node” and is indicated by the cursor. In Figure 2, when the image of a tissue sample has been dropped on the BCDEBar plug-in from a webpage, a generic icon (the rectangle icon in Figure 2) appears instantly on the BCDE canvas and it is listed in the BCDE object window. Useful metadata such as the URL of the source webpage and time of capture are automatically transferred to the proper annotation fields of the newly created icon. The image itself is stored as an attachment, which can then be opened at a later time using an appropriate program such as Microsoft Paint (the partly shown image A at the upper right corner).

Query SNP Function Portal in BCDE: Figure 3 shows how SNP function properties can be queried while drawing a diagram. In this example, a user selected the icon representing the SNP “rs1067”. By selecting “Query External Databases” => “SNP Function Portal” in the popup menu, the SNP ID “rs1067” is sent to the SNP Function Portal. The user could then set other query parameters (e.g., linkage disequilibrium criteria) to query the potential function implication of this SNP in the SNP Function Portal. A user can then drag-and-drop a relevant concept, such as an altered transcription factor binding site, to BCDE canvas for further exploration.

RDF Database Query: The RDF database query can be performed on any BCDE icon by selecting the icon and then right clicking to bring up the popup menu. Currently, only the immediate neighbor query is enabled. In the popup menu, the user needs to select “Query Interactions”. If the BCDE icon does not have a BCDE ID assigned to it, the BCDE application will first activate the concept matching web service to request that the user assign a BCDE ID to the selected icon. Once a BCDE ID has been assigned to the icon, BCDE queries the RDF database for the nearest neighbors and imports them into the canvas to facilitate hypothesis development.

Use Case: To generate a diagram that summarizes results from genomic studies for bipolar disorder in order to select candidate genes and SNP alleles for follow-up investigations, a user can start with a Medline search for gene expression and genotyping studies related to bipolar disorder (BPD). He may then drag-and-drop genes, SNPs and other concepts such as molecular processes that are found to be related to BPD from different abstracts and papers to a BCDE canvas. With a populated canvas, he may then link some concepts to each other using different types of arrows based on the literature or his background knowledge. He can also use the aforementioned search functions to fill conceptual gaps and create new nodes, relationships and annotations using information from other databases. If there is already a pathway diagram generated by other researchers in the BioPAX format, he can use
the existing diagram as his starting point by importing it into BCDE. The diagram he created can be saved and later modified by his collaborators with modification privileges. BCDE will track the data sources, authors and time for each icon and relationship in the diagram. The concept match service ensures consistent annotation. Such a diagram can also be exported to PowerPoint or Visio with full data source links for inclusion in a presentation. The use of BCDE in this process leads to more efficient capture, integration and sharing of information.

**Discussion**

Although still under development, BCDE and its unique functions stand out among other pathway editors in terms of ease of data capture and terminology consistency.

The main focus during the next phase of development will be functions for comparing, merging and mining diagrams in the BCDE database. For example, without a function that can identify the similarity among different diagrams, users may spend a considerable amount of time trying to understand the differences between diagrams, since diagrams that are conceptually equivalent may look very different due to differences in layout, icon selection and color scheme. Similarly, combining information from different diagrams will also be difficult without an automatic diagram merging function. We plan to start with the relative simple “bag of words” approach widely used in document retrieval systems. More powerful but computationally intensive graph-based diagram comparison and retrieval methods, such as the SAGA algorithm will be included later.

When the BCDE database is well populated with conceptual relationships from existing databases as well as those captured by individual researchers, it will be possible to link together disconnected pieces of information stored in many different diagrams for novel discoveries using the ArrowSmith algorithm (Swanson, 1988; Smalheiser & Swanson, 1998). The discovery of new conceptual relationships in the BCDE database will be easier than using free text literature alone because the conceptual relationships stored in BCDE are conscientiously extracted by researchers and presented using consistent terminology.

BCDE is a new effort combining drag-and-drop data capture with consistent terminology to facilitate the integration and sharing of conceptual relationships derived from heterogeneous sources. With additional improvements in its functionality by programmers and contributions to its data stores by biomedical researchers, we hope BCDE can become a useful tool for linking genomic level analysis results to pathophysiological processes through efficient data and knowledge integration.

**Acknowledgement:** Bohua Yu, Elvis Jakupovic, Justin Wilson, Manhong Dai, Weijian Xuan, Fan Meng and Stanley Watson are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through NIH grant 1U54DA021519-01A1 to U of Michigan.

**References**

1. AMIA. (2007) AMIA Strategic Plan. [http://www.amia.org/inside/stratplan/](http://www.amia.org/inside/stratplan/)
2. CambridgeSoft. (2004) BioDraw. [http://www.cambridgesoft.com/products/family.cfm?FID=55](http://www.cambridgesoft.com/products/family.cfm?FID=55)
3. ProteinLounge. (2005) Pathway Builder 1.0. [http://www.proteinline.com/pathwaybuilder.asp](http://www.proteinline.com/pathwaybuilder.asp)
4. Visimation. (2005) PathViz 2.0. [http://www.shapesource.com/scripts/itdviewit.asp?idproduct=70](http://www.shapesource.com/scripts/itdviewit.asp?idproduct=70)
5. Demir, E. et al. (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* 18 (7), 996-1003
6. Demir, E. et al. (2004) An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics* 20 (3), 349-356
7. Gamma, E. and Eggenschwiler, T. (2005) JHotDraw as Open-Source Project. [http://www.jhotdraw.org/](http://www.jhotdraw.org/)
8. Xuan, W. et al. (2003) Identifying Gene and Protein Names from Biological Texts. Proceedings of the 2003 IEEE Bioinformatics Conference (CSB2003), 639-643
9. FRUCHTERMAN, T.M.J. and REINGOLD, E.M. (1991) Graph Drawing by Force-directed Placement. SOFTWARE—PRACTICE AND EXPERIENCE 21 (11), 1129-1164
10. Donley, C. (2002) LDAP Programming, Management, and Integration, Manning Publications Company
11. Wang, P. et al. (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics* 22 (14), e523-529
12. Tian, Y. et al. (2007) SAGA: a subgraph matching tool for biological graphs. *Bioinformatics* 23 (2), 232-239