Visualizing Multivariate Hierarchic Data Using Enhanced Radial Space-Filling Layout

Ming Jia[1], Ling Li[2], Erin Boggess[1], Eve Syrkin Wurtele[2], and Julie A. Dickerson[1]

[1] Dept. of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA
[2] Dept. of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA
jiaming@iastate.edu, liling@iastate.edu, eboggess@iastate.edu, mash@iastate.edu, julied@iastate.edu

Abstract. Currently, visualization tools for large ontologies (e.g., pathway and gene ontologies) result in a very flat wide tree that is difficult to fit on a single display. This paper develops the concept of using an enhanced radial space-filling (ERSF) layout to show biological ontologies efficiently. The ERSF technique represents ontology terms as circular regions in 3D. Orbital connections in a third dimension correspond to non-tree edges in the ontology that exist when an ontology term belongs to multiple categories. Biologists can use the ERSF layout to identify highly activated pathway or gene ontology categories by mapping experimental statistics such as coefficient of variation and overrepresentation values onto the visualization. This paper illustrates the use of the ERSF layout to explore pathway and gene ontologies using a gene expression dataset from *E. coli*.

1 Introduction

Linking high-throughput experimental data with hierarchical ontologies that relate biological concepts is a key step for understanding complex biological systems. Biologists need an overview of broader functional categories and their performance under different experimental conditions to be able to ask questions such as whether degradation pathways have many highly expressed genes, or which biological process categories are overrepresented in the data. These needs pose many unique requirements on the visualization of biological ontologies, such as being able to visualize an overview of an ontology mapped with experimental data and clearly show the non-tree connections in ontology.

Current tools which visualize biological ontologies normally employ the traditional Windows™ Explorer-like indented list, as are found in EcoCyc[1] and AmiGO[2], or node-link based layouts, e.g., OBOEdit[3] and BinGO[4]. These kinds of layouts are well suited for tens of nodes, however quickly become cluttered if hundreds of nodes are shown simultaneously. As a result, users often collapse the ontology to reduce its visual complexity, and only expand small portions when needed. The tradeoff of this abstraction is the loss of context of the overall ontology structure. Moreover, biological ontologies are not pure tree structures, but are directed acyclic graphs (DAG), i.e.,

G. Bebis et al. (Eds.): ISVC 2010, Part I, LNCS 6453, pp. 350–360, 2010.

they contain non-tree edges where many child nodes having multiple parents. Current software tools are not suitable for tracing such connections.

To address these problems, we propose the enhanced radial space-filling (ERSF) algorithm that uses an intuitive orbit metaphor to explicitly visualize non-tree edges, and make them appear differently than the major hierarchic structure. The ERSF is implemented in a software package based on the Google Earth API. To the best of our knowledge, this is the first application to use 3D RSF in biology and the first algorithm to visualize non-tree edges on a RSF plot.

Some preliminary results that demonstrate the use of ERSF for a single ontology dataset have been published in a workshop proceeding [5]. This work focuses on the visualization benefits of the ERSF layout on multiple datasets in terms of user requirements. The platform is also extended to visualize general ontology structures and multivariate data. Moreover, we here conducted an initial user test and summarize this feedback in Section 4.

The contributions of our ERSF-based software to information visualization area are:

- Applying the radial space-filling layout to a common but challenging visualization task in biology field.
- Enhancing the radial space-filling technique with orbits metaphor for visualizing non-tree edges in hierarchic dataset.
- Mapping key summary statistics from experimental high-throughput data on the hierarchical visualization and links with traditional parallel coordinate views.

This paper is organized as follows: Section 2 describes the properties of the biological datasets, lists the requirements for the visualization and assesses the related work. Section 3 d the ERSF layout. Section 4 reveals some interesting findings from the initial user testing.

2 Background

2.1 Ontology Data Description

An ontology is a formal explicit description of concepts, or classes in a domain of discourse [6]. Biologists use ontologies to organize biological concepts. The Gene Ontology (GO) [2] is a controlled vocabulary of gene and gene products across all species. The Pathway Ontology (PO) [7] is a recent concept that provides a controlled vocabulary for biological pathways and their functions. PO, like many other ontologies, is hierarchical data, but it is not a pure tree structure because several pathways may have multiple parents. Both ontologies are actually directed acyclic graphs. To facilitate the visualization, we first construct a spanning tree in the ontology, and then define the connections in the spanning tree as tree edges and all remaining edges as non-tree edges or cross links. The non-tree edges are of particular interest since they represent pathways that perform multiple functions.

We illustrate this application with the *E. coli* Pathway Ontology from EcoCyc [1]. The EcoCyc PO contains 442 nodes, where 289 of them are pathways or leaves. It also contains 508 edges, where 67 (13.2%) are non-tree edges. PO's for other species are slightly different, however, they are of similar scale. Another feature typical of a

PO is that the height of the hierarchy is normally low, e.g., 6 for *E. coli*, which results in a very large width/height ratio (289/6=48.1).

Another dataset we focus on is the Gene Ontology (GO) Slim [8], which are important subsets of GO that contain around 100 terms.

Besides studying PO structure, in day-to-day research, biologists need to make sense of system-wide experimental data and wish to understand how the experimental conditions affect the underlying biology. One typical type of experimental data is transcriptomics (often referred to as gene expression data), which describes the abundance of gene transcripts during an experiment. Other experimental data types include metabolomics and proteomics. For gene expression data, the original data is typically a data matrix where each row describes a gene, and each column records the expression level of genes under a certain condition, e.g., a point, treatment, or replicate.

2.2 Visualization Requirements

Based on the data described above and tasks biologists perform, the basic requirements for visualization of PO are to:

- View the whole ontology on a single screen to gain a global feeling for the data and the main hierarchical structure.
- View ontology details by navigation and/or interaction (zoom, pan, rotation).
- Map attributes on the ontology so that they are easily visible.
- Clearly show non-tree connections.

2.3 Related Work

The most widely-used representation for ontology structure is the Windows™ Explorer-like tree view, or indented list. One implementation of indented list (Class Browser) is evaluated in [9] with three other methods (Zoomable interface, Focus + Context, and Node-link/tree). The indented list lacks the ability to show non-tree edges. Users presented with indented list naturally think the underlining data is a pure tree structure.

Node-link graph and treemaps [10] are also widely used to visualize ontology. OBO-Edit [3] combines an indented tree browser (Tree Editor) and a graphical tree drawing (Graph Editor) which uses the node-link based layout from GraphViz[11]. BinGO[4], a Cytoscape plug-in for analyzing Gene Ontology, uses the default 2D hierarchic layout from Cytoscape. The node-link based layout is very good at showing simple hierarchical structures (e.g. contain less than 50 nodes). However when the number of entities increases, those layouts become very cluttered and incomprehensible. Fig. 1 shows the result of our PO dataset using these layout methods. We can see that the whole hierarchic structure and non-tree edges are not obvious in these views. Due to the cluttered layout for a large number of entities, researchers normally confine their view to a limited subset of the whole structure, and are thus unable to gain the global knowledge.

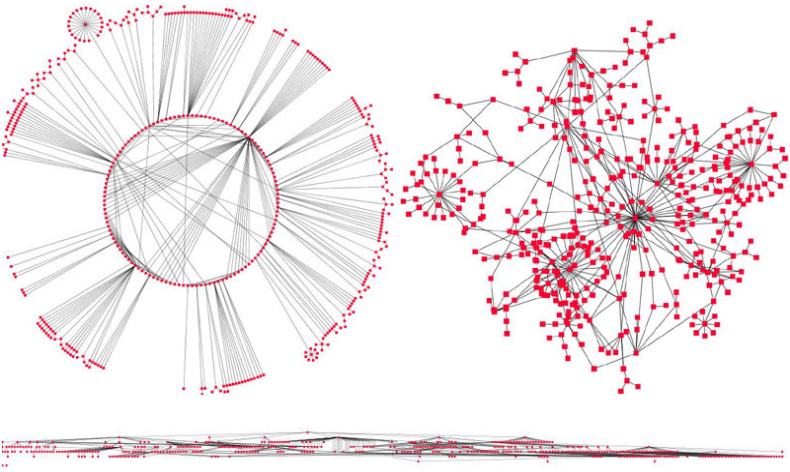

Fig. 1. The Pathway Ontology of *E.coli* from EcoCyc is shown in Cytoscape using circular layout (top left), organic layout (top right), and hierarchic layout (bottom). The ontology contains 442 nodes and 508 edges. The hierarchical structure can hardly be seen.

Treemap based systems [10] are able to visualize the whole GO with mapped data in one screen, and are suitable for identifying regions of interest. However, the hierarchical structure is hard to see in a treemap since it is a nesting-based layout which overplots the parent nodes with their children nodes [12]. Another limitation of treemap is that it lacks a meaningful representation of non-tree edges, a key requirement. As observed in [6], treemaps and other space-filling layouts normally duplicate nodes which have multiple parents. If the node being duplicated is a non-leaf node, the whole substructure rooted at this node will be duplicated as well. Thus duplicating nodes in hierarchic dataset may greatly increase a graph's visual complexity. Duplication also causes confusion for the user. For example, when user finds two regions have similar visual patterns in a treemap, they may think that they have discovered two groups of genes functioning similarly. Unfortunately, they often turn out to be the identical GO terms being drawn twice.

Besides the visualization methods mentioned above, Katifori et al. [6] have also presented many tools and layout algorithms to visualize ontologies and graphs in general. For example, a hyperbolic tree [13] can handle thousands of nodes. However, in a hyperbolic tree visualization, it is difficult to distinguish between tree and non-tree edges among hundreds of edges since they are all represented as links. Another disadvantage is that hyperbolic trees are not space efficient, and normally only a couple of pixels are used for each node. Therefore attributes (like gene expression data) mapped on nodes become hard to distinguish and interpret.

Space-filling methods are considered very space-efficient and are good for mapping attributes on node regions. Despite the disadvantages of rectangular space-filling (such as treemap), evaluations [14] find that radial space-filling (RSF) methods [15] are quite effective at preserving hierarchical relations.

3 Enhanced RSF Design

3.1 Visualizing an Ontology as a Tree

For explanatory purpose, we first assume the ontology as a pure tree structure that does not have any non-tree edges, and explain how the traditional RSF technique can visualize this simplified data.

In the RSF drawing, each circular region represents one node in the tree. The node represents an ontology term, and can either be a pathway (leaf) or a category (non-leaf). For the ease of explanation, we will interchangeably use the words ontology term, node, and region.

The height of each region is set as proportional to the height of the subtree rooted at that node. For color, initially we used structure-based coloring [15] to convey additional hierarchical information, where the leaf node regions are colored according to the color wheel and the non-leaf node regions are colored as the weighted average of its children's color. However, during the initial user testing, several users pointed out that the drawing is full of saturated color, and it is hard to distinguish orbits from the main drawing. To solve this problem, we propose and provide an option to use orbit-based coloring where every category regions are white and highly transparent.

Fig. 2a shows a small tree with eight leaf nodes and five non-leaf nodes, labeled as graph G1. Fig. 2b shows the result of using RSF in 3D on graph G1. Non-leaf nodes correspond to pathway categories.

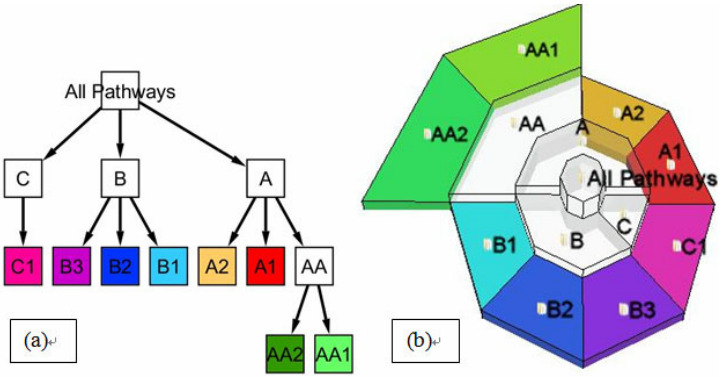

Fig. 2. Graph G1 shows hierarchical relationships among leaf nodes (pathways) and non-leaf nodes (pathway categories), drawn in hierarchic layout in Cytoscape (a) and the radial space-filling layout (b)

3.2 Visualizing Pseudo Ontology with Non-tree Edges

As noted earlier, RSF cannot support non-tree edges. To better meet the visualization requirements, we proposed the enhanced RSF layout, or ERSF, which uses orbits to represent non-tree edges. Fig. 3a shows graph G2, which adds four non-tree edges to G1. The ERSF drawing of G2 is shown in Fig. 3b where the spanning tree is drawn

using traditional RSF. The metaphor of "satellite orbits" represents non-tree edges as circular links. For each child node, which has at least two parents, one orbit circle is drawn on the layer of that node. The parent that connects the node in the spanning tree is the major parent and other parents are minor parents. The region of each node is placed under the region of its major parent as in RSF. For every minor parent, a green edge from the center of its region to the orbit of the child is called the 'downlink'. The intersections between downlinks and orbits are called access points, which are represented by red dots.

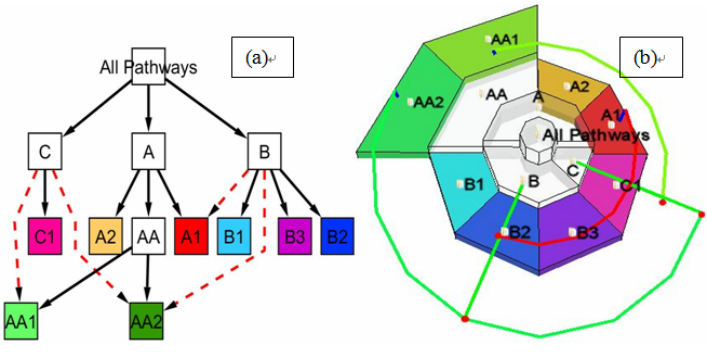

Fig. 3. Graph G2 is drawn using hierarchic layout in Cytoscape (a) and the Enhanced Radial Space-Filling (ERSF) layout using orbit-based coloring (b). In the hierarchic layout, red dashed lines represent non-tree edges. In the ERSF layout, orbits with blue and green radial links represent non-tree relations. For example, the green line extruded from B contains two red-dots: the inner one intersects with red orbit of A1 and the outer one intersects with green orbit of AA2. These orbits mean that B is the minor parent of both A1 and AA2.

To help viewers find and visually trace interesting non-tree edges, the orbits need to be distinguishable from one another. In order to do this, the orbits are first restricted to span in the middle area of each layer, thus leaving a visually apparent gap between orbits in adjacent layers. To distinguish orbits in the same layer, we make the orbit the same color as the child region originating the orbit.

3.3 Visualizing Pathway Ontology Dataset

The PO data from *E.coli* is presented with the ERSF view in Fig. 4. Compared to Fig. 1 where the same dataset is shown by node-link based layout, it is clear that ERSF can show some patterns on the overview. For example, the most orbits are concentrated on the third layer, and one category (*methylglyoxal detoxification*) contains many children in other categories because its green uplink intersects many light blue orbits.

The orbit-based coloring allows users to visually trace the orbits. For example, the category *amino acids degradation* (on left) intersects with one orange orbit. To find the child of this orbit one can visually trace the orbit along the circular curve or directly glance at the orange regions on the right side, and find the child region which originates the outer-most orbit. The red dot serves as a "shortcut" for this specific task. For instance, users can click on the red dot on the intersection of category *amino*

acids degradation and the orbit and then a pop-up dialog will indicate that it connects to child *superpathway of threonine metabolism*.

It is also clear that three pathways in the category *cell structures biosynthesis* are also the children of another category *fatty acids and lipids biosynthesis*. When a user wants more information about those non-tree edges, he can rotate and zoom the view.

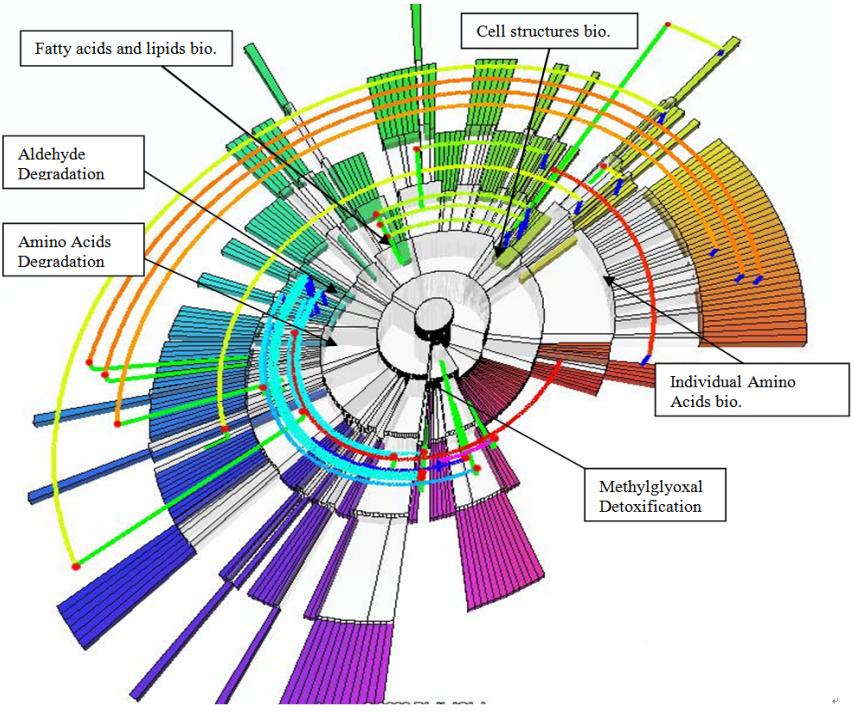

Fig. 4. The hierarchical structure of the ontology is clearly shown in ERSF method using orbit-based coloring. There are many pathways that belong to at least two categories. This kind of multiple inheritance information is difficult to see in other visualization methods.

3.4 Mapping Experimental Values on Ontology

The strategy of a biological scientist evaluating experimental data is to look for which parts of the network show significantly different measurements across different conditions. Questions such as 'Which pathways or categories are most changed under *anaerobic stress*?' can be addressed by mapping the values onto the whole Pathway Ontology.

We use animation to show the values of a series of experiments. For instance, one time-series experiment with 4 time points is presented as animation of 4 frames.

To analyze the gene expression data, we initially map the average expression value on color, and map the variation on height. The visual results of two frames are shown in Fig. 5a and Fig. 5b. The first frame shows the value of one replicate in controlled condition, while the second frame shows that of the treatment condition.

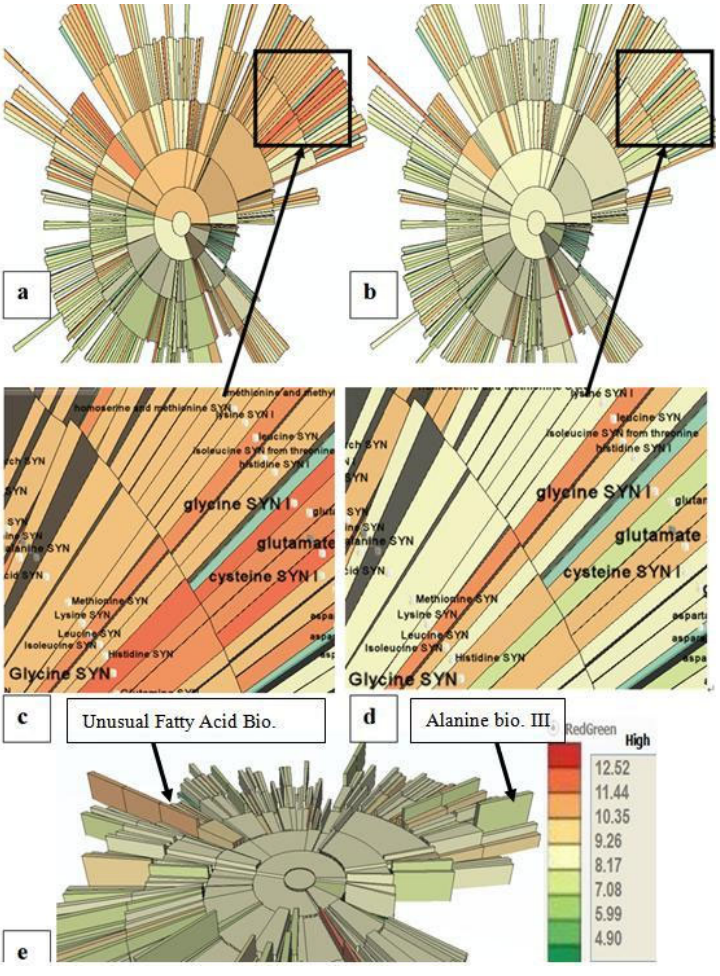

Fig. 5. Average expression values and coefficient of variation are shown for each condition. Color gradient represents values for gene expression, from green (low) to red (high). Two conditions are compared. The orange and red colors in condition 1 show that these categories have much higher expression values in condition 1 (a) than in condition 2 (b). The differences between these two conditions are more obvious when using animation. (c), (d) show the details in the *glycine biosynthesis* categories. When the view is tilted (e), the categories with high variation are shown by their higher height.

Users can tilt the view to see the height of each region (Fig. 5e). In this view, one category (*unusual fatty acid biosynthesis*) stands out, because its and its descendents have very high variation and expression values. This discovery demonstrates the benefit of using 3D to show these two attributes together. Another similar interesting discovery is pathway *alanine biosynthesis III*, which also has very high variation but very low expression values.

By switching between these two conditions, we notice that most of the pathways and categories have a greenish color under the treatment, which indicates lower

expression values in the treatment condition than in the controlled condition. This is an interesting trend, since in most experiments the treatments normally have greater values. To confirm this trend, we can map the difference between these two conditions directly on the ontology. We can also map many other attributes onto the color and height of the ERSF drawing, e.g. statistical significance p-value.

4 Initial User Testing

To get some initial feedback from the users, we conducted a pilot user testing involving four users who are PhD students or postdocs in biology field. The goal is to better understand the needs of the biologist-users and to test the effectiveness of the ERSF.

Users were presented with several tasks in two categories: understand the ontology and the gene expression data. One typical user task is to find the pathways which belong to multiple categories. In order for the users to provide the most realistic and valuable feedback, they worked in a relaxed setting where the tasks were not timed.

All users who participated in the pilot user test preferred the ERSF solution to the traditional indented list and node-link based layout. They think the ability to show the whole ontology structure is an important feature, and is especially useful for the system scale experimental dataset. The reason is that knowing which parts of the whole system the experiment affected is an important goal in their research. However, this is hard to do if they are only presented with a small subset of the system. Moreover, users generally gave up on some time-consuming tasks. For example, finding the pathways that belong to at least two categories is extremely difficult using indented lists and node-link based layouts.

Another interesting phenomenon is that although ERSF provides a 3D view of the ontology, users mostly view it from the top down orientation, which is essentially a 2D ERSF layout. Therefore, when users were given the choice to map an attribute to either color or height, all of them prefer mapping the most important attribute to color. Some possible reasons include: biologists are used to traditional 2D tools, and height is hard to interpret precisely due to foreshortening [16]. Nevertheless, the 3D view provides the benefit of mapping two variables simultaneously (color and height). This ability is important for some tasks that may lead to interesting discoveries, e.g. finding pathways that both have high variation and high expression value.

5 Discussion

Fig. 4 shows that visualizing the ontology using ERSF has several advantages. First, this design clearly distinguishes between spanning tree relationships and non-tree edges. Second, compared to treemaps with a crosslink overlay [17], there are much fewer edge-crossings and the drawing is neater since orbits and links are circular and radial respectively. Third, all downlinks of a parent share only one link edge, thus the total length of those edges is the same as the length of the longest link. This property reduces the graph's visual complexity, especially when one node is the minor parent of many other child nodes.

Another benefit of using ERSF is that it does not duplicate nodes, which reduces the visual complexity compared to normal RSF. For the *EColi* PO dataset, ERSF

reduced 67 duplicated nodes out of 442 nodes (15.2%). For the GO Slim dataset, since many nodes that have multiple parents are categories, RSF duplicates the whole subtree rooted at those nodes. On the contrary, ERSF reduced 38 duplicated nodes out of 112 nodes (33.9%).

When mapping the node experimental data onto regions' height, e.g. in Fig. 5, it is cumbersome to render the orbits because the orbits may be occluded by higher regions. It is also difficult to follow the orbits when the regions color is mapped by experimental values. As a result, the orbits are not shown by default when mapping attributes onto regions.

The Pathway Ontology dataset shown here contains around 500 nodes, and can be gracefully drawn in one screen. Our suggestion is to limit the data size to 1000 nodes since otherwise the peripheral regions will become as thin as one pixel in width and are difficult to be distinguished. We also noticed that in our dataset, the percentage of non-tree edges is relatively low (from 10% to 20%). The edge bundling method proposed in [18] may be helpful for dataset with high percentage (e.g. above 40%) of non-tree edges. As a result, we suggest that the ERSF method is best suited for visualizing medium-sized multivariate hierarchic data (contains 100 to 1000 nodes) and with medium-to-low percentage of multiple inheritances.

6 Conclusion and Future Work

This work focuses on effective visualization of hierarchic ontologies in biological research. To satisfy the visualization requirements, we propose the enhanced radial-space filling (ERSF) method which arranges ontology regions circularly in a 3D space and uses orbits to represent the non-tree edges. To facilitate the study of large-scale, system-level experimental data, we provide various and customizable ways to map data and statistical results on the ERSF visualization.

The proposed ERSF algorithm has two major advantages over traditional methods in biological data visualization. First, it provides easy visual identification and navigation of non-tree edges in ontology without duplicating nodes. Second, it allows large scale experimental data to be mapped and navigated on the context of the hierarchical structure of the ontology, which may lead to discoveries on a system level.

Initial testing by users has shown the tool to be preferable to their current working solutions, which have been based on indented lists and node-link layouts. A larger quantitative user study is planned in the near future.

The proposed ERSF method can also be adapted to visualize other types of hierarchic data, e.g., company hierarchy and software inheritance diagrams.

References

1. Keseler, I.M., et al.: EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic Acids Res. 37(Database issue), D464-D470 (2009)
2. Carbon, S., et al.: AmiGO: online access to ontology and annotation data. Bioinformatics 25(2), 288–289 (2009)
3. Day-Richter, J., et al.: OBO-Edit–an ontology editor for biologists. Bioinformatics 23(16), 2198–2200 (2007)

4. Maere, S., Heymans, K., Kuiper, M.: BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21(16), 3448–3449 (2005)
5. Jia, M., et al.: MetNetGE: Visualizing biological networks in hierarchical views and 3D tiered layouts. In: IEEE International Conference on Bioinformatics and Biomedicine Workshop, BIBMW 2009 (2009)
6. Katifori, A., et al.: Ontology visualization methods a survey. ACM Computing Surveys 39(4), 10 (2007)
7. Green, M.L., Karp, P.D.: The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Res. 34(13), 3687–3697 (2006)
8. Consortium, G.O.: GO Slim and Subset Guide (2009), `http://www.geneontology.org/GO.slims.shtml`
9. Katifori, A., et al.: Selected results of a comparative study of four ontology visualization methods for information retrieval tasks. In: Second International Conference on Research Challenges in Information Science, RCIS 2008 (2008)
10. Baehrecke, E.H., et al.: Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinformatics 5, 84 (2004)
11. Ellson, J., Gansner, E.R., Koutsofios, E.: Graphviz and dynagraph static and dynamic graph drawing tools. Technical report, AT&T Labs - Research (2003)
12. Tekusova, T., Schreck, T.: Visualizing Time-Dependent Data in Multivariate Hierarchic Plots -Design and Evaluation of an Economic Application. In: Information Visualisation, IV 2008, Columbus, OHIO, USA (2008)
13. Munzner, T.: Exploring Large Graphs in 3D Hyperbolic Space. IEEE Computer Graphics and Applications 18(4), 18–23 (1998)
14. John, S.: An evaluation of space-filling information visualizations for depicting hierarchical structures, pp. 663–694. Academic Press, Inc., London (2000)
15. Yang, J., et al.: InterRing: a visual interface for navigating and manipulating hierarchies, pp. 16–30. Palgrave Macmillan, Basingstoke (2003)
16. Munzner, T.: Process and Pitfalls in Writing Information Visualization Research Papers. In: Information Visualization: Human-Centered Issues and Perspectives, pp. 134–153. Springer, Heidelberg (2008)
17. Fekete, J., Wang, D.: Overlaying Graph Links on Treemaps. In: Information Visualization 2003 Symposium Poster Compendium. IEEE, Los Alamitos (2003)
18. Holten, D.: Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. IEEE Transactions on Visualization and Computer Graphics 12(5), 741–748 (2006)