Joslyn, Cliff A; Mniszewski, SM; Smith, SA; and Weber, PM: (2006) "SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology", in: Joint BioLINK and 9th Bio-Ontologies Meeting (JBB 06), pp. 109-125

SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology

CA Joslyn*[1], SM Mniszewski[1], SA Smith[2] and PM Weber[2]

[1]Computer Science Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545, USA
[2]Decision Applications Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545, USA

Email: CA Joslyn - joslyn@lanl.gov; SM Mniszewski - smm@lanl.gov; SA Smith - sas@lanl.gov; PM Weber - pmw@lanl.gov;

*Corresponding author

Abstract

We introduce SpindleViz, a three dimensional visualization environment for the Gene Ontology (GO) and other taxonomically structured bio-ontologies. SpindleViz exploits the mathematical properties of the partially ordered sets (posets) implied by identified portions of the GO to properly represent the interval-valued vertical levels of GO nodes, and then orient them around a central "spindle" with respect to their centrality in the structure. A novel force-directed layout algorithm allows a natural distribution of the GO nodes in 3-D space, while still respecting the underlying mathematical constraints.

1 Introduction

Currently available tools for visualizing and interacting with portions of large bio-ontologies such as the Gene Ontology (GO) [3] (identified from e.g. gene expression experiments, functional annotation applications, etc.) are inadequate in various ways:

- Classical tools, including early versions of Amigo[1], and DAG-Edit, represent the GO as an indented list, similar to file browsers. While appropriate for tree structures, GO's Directed Acyclic Graph (DAG) structure results in multiply inheriting nodes being "flattened" by being listed multiple times.
- Other systems also provide an actual visualization of the GO structure, but still rely on a flattening approach to transform the inherent DAG structure of the GO into a tree [1].
- Amigo now provides a graphical viewer, and others (e.g. GeneInfoViz[2]) display the DAG structure of the GO correctly, but in a manner which doesn't accurately reflect the underlying mathematical properties such as vertical distance, and doesn't include such basic tasks as showing all the descendants of a node. Resulting layouts can be difficult to read and maneuver.
- Finally, in our opinion, the size and complexity even of the portions of the GO typically encountered by researchers requires embedding visualizations in three dimensions. While limited 3-D lattice visualizers are available from the mathematics community (see Freese's work in particular [2][3]), we are not aware of any 3-D GO visualizers currently available.

We introduce SpindleViz, a three dimensional visualization environment for the GO. SpindleViz exploits the mathematical properties of the partially ordered sets (posets) implied by GO portions to properly represent the interval-valued vertical levels of GO nodes, and then orients them around a central "spindle" with respect to their centrality in the structure. A force-directed layout algorithm allows a natural distribution of the GO nodes in 3-D space, while still respecting the underlying mathematical constraints.

[1] http://godatabase.org
[2] http://genenet.org/geneinfoviz
[3] http://www.math.hawaii.edu/~ralph/LatDraw

2 Method

Consider the example diagram of a model portion of the GO DAG shown in the left side of Fig. 1. While this example is very small, it illustrates some of the key structural features of the GO, including multiple inheritance and a "bushy" structure widening towards the bottom, with many leaves.

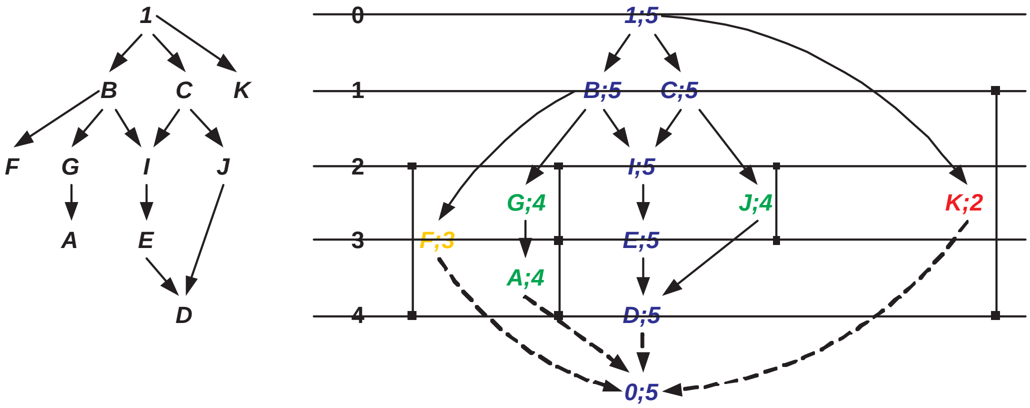

Figure 1: (Left) A model portion of the GO DAG. (Right) The same portion shown with: centrality on each node, including color coding; the interval rank of each node; and laid out vertically by interval rank midpoint and horizontally by centrality.

Following our prior work [6, 8], we represent the GO DAG as a discrete mathematical structure called a partially ordered set (poset) $\mathcal{P}$ on a set of GO nodes P. The right side of Fig. 1 shows the same structure with mathematical quantities related to the poset structure identified, and laid out accordingly. While space precludes a detailed technical description of the poset mathematics here (see [5–8]), we will note:

- We have added a **virtual bottom** node $0 \in P$, useful for completing the mathematical analysis.
- Each node $a \in P$ can be characterized by the set of **complete chains** it sits on connecting the top node $1 \in P$ to the virtual bottom node $0 \in P$ through a. For example, B sits on the three chains $1 \to B \to F \to 0, 1 \to B \to G \to A \to 0$, and $1 \to B \to I \to E \to D \to 0$.
- Consider depth in the structure as vertical level or rank. The node $C \in P$ is at the "first level", in that it is one down from the top, which we call "top rank" $r^t(C) = 1$. But while K is also one down from the top, unlike C, it is also a leaf, and thus *also* close to the *bottom*. In fact, K is both one down from the top and one up from the bottom, while C is four up from the bottom. In this way, each node also has a bottom rank, so that $r^b(C) = 1, r^b(K) = 4$, and thus an **interval rank** $R(a) = [r^t(a), r^b(a)]$. So $R(K) = [1, 4]$, existing at levels one through four *simultaneously*, as shown in the figure. Similar concepts have been used by Freese [2].
- The **centrality** $s(a)$ of a node $a \in P$ is the length of the largest complete chain it sits on. For example $s(K) = 2$, while $s(C) = 5$. All the blue nodes have maximum centrality 5, and are oriented horizontally at the center along a (possibly complex) central "spindle". Nodes are also colored from cold to warm with decreasing centrality.

For our 3-D layout algorithm, a set of initial 3-D cylindrical coordinates $\langle \rho, z, \theta \rangle$ are established where z is vertical placement at the interval rank midpoint $r^m(a) = (r^t(a) + r^b(a))/2$; ρ is horizontal placement proceeding outwards from the spindle with decreasing centrality; and θ is the angular position out of the plane around the spindle randomly assigned based on the number of nodes with the same initial z. Then a force-directed layout algorithm allows the nodes to relax in three dimensions: z is constrained between the rank interval midpoint and bottom $[r^m(a), r^b(a)]$, limiting "vertical violations" due to overlapping interval ranks; ρ is constrained to limit "horizontal violations" where a more central node is positioned outside of a less central node; and θ is free to rotate out of the plane around the spindle.

Since vertical position z is a direct function of the interval rank $R(a)$ reflecting the inheritance relations, no vertical violations are tolerated. However, horizontal position ρ is a function of centrality, whose semantics may be somewhat arbitrary based on the particular chain lengths present in a part of the GO; thus limited horizontal violations can be tolerated to improve the layout. Finally, reflecting its status as a mechanism both for the mathematics and for the layout, the virtual bottom is "locked" into place below the top node, and then hidden from the user.

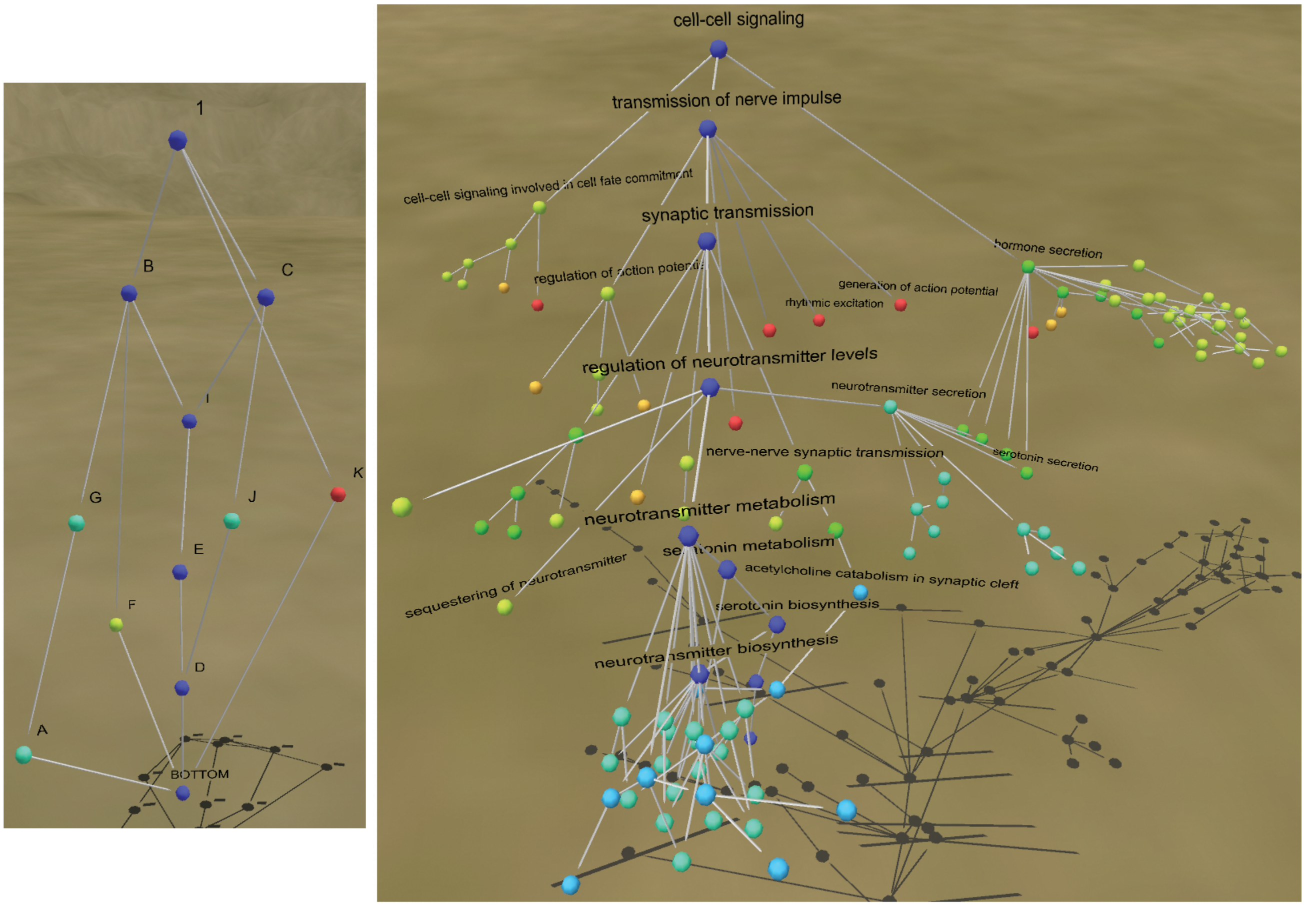
1
B
C
I
G
J
K
E
F
D
A
BOTTOM
cell-cell signaling
transmission of nerve impulse
cell-cell signaling involved in cell fate commitment
synaptic transmission
hormone secretion
regulation of action potential
generation of action potential
rhythmic excitation
regulation of neurotransmitter levels
neurotransmitter secretion
nerve-nerve synaptic transmission
serotonin secretion
neurotransmitter metabolism
serotonin metabolism
acetylcholine catabolism in synaptic cleft
sequestering of neurotransmitter
serotonin biosynthesis
neurotransmitter biosynthesis

Figure 2: (Left) Example from Fig. 1 in SpindleViz. (Right) Cell-Cell Signalling and its descendants in BP.

3 Implementation and Example

SpindleViz is implemented in Flatland[4], a 3-D visualization environment developed by the University of New Mexico. While designed to run in a fully immersive environment such as a CAVE or Head Mounted display, Flatland also runs on a workstation or even a modern laptop. In flat, fixed images such as presented in this paper, depth cues like perspective scaling, occlusion, distance attenuation (fog), lighting, shadows, and textures are all that can be relied upon to perceive 3-D effectively.

SpindleViz uses a novel generalized force-directed layout algorithm developed by us for use in Flatland. As in standard force-directed approaches (which are widely used in general bioinformatics applications, e.g. [4]), nodes are assigned mass and charge, and edges length and strength. Nodes experience a force proportional to the spring strength K divided by the difference between the length of the edge and it's specified length. We enhance the standard inverse-square Coulomb-like repulsion force $F = A/R^2$, where R is the distance between nodes and A is a scaling factor, by freeing the exponent and introducing an

[4] http://www.hpc.unm.edu/research/scientific_visualization/homunculus_project.htm

attractive force inspired by the Lennard-Jones potential used in molecular modeling, resulting in a two-term expression $F = A/R^{\alpha} - B/R^{\beta}$. Continuous adjustment of α, β, and the ratio A/B, allows a wide variety of results to be displayed, and exploration and exposure of different features of graphs.

For poset display, central nodes (blue) are connected by short, strong edges while peripheral nodes (red) are connected by longer, weaker edges. In particular, spring length and the spring constant for each edge is mapped to the square of the minimum of the two centralities of the nodes it connects. We also adjust α and β to encourage both tight packing of highly connected components (clustering) and broad separation of loosely disconnected branches. The left side of Fig. 2 shows the example from Fig. 1 in SpindleViz, while the right side shows the GO node "Cell-Cell Signalling" in the Biological Process branch, GO:0007267, and all its descendants. Both use $\alpha = .5, \beta = 1.0, A = 50$, and $B = 80$, yielding $A/B = .625$. Space considerations will limit us to showing and discussing this single specific real example.

We note a number of features drawn out by the layout, and reflecting the underlying mathematical structure. First, the central spindle in blue descends a relatively deeper portion of the GO, down to "neurotransmitter biosynthesis". Leaf nodes, which indicate the most specific functions and would tend to accumulate the most protein annotations, are readily visible, for example "serotonin secretion". However, not all leaves are the same, and in particular red nodes such as "rhythmic excitation" indicate functions which are specific in the sense of accumulating annotations, but are also shallow, indicating a region of the GO which is not as fully developed. Areas of multiple inheritance are also evident, for example "serotonin secretion" drawing from both "neurotransmitter secretion" and "hormone secretion". Most significant is the clustering effect achieved, with two clear groups under each of those nodes and "neurotransmitter metabolism" (on the spindle), but with "hormone secretion" the largest and deepest.

4 Conclusions and Future Work

SpindleViz is a 3-D capability which is difficult to translate to a flat image in this paper. The ability to rotate, orbit, and zoom, with text becoming more apparent as nodes come closer, is essential to a full apprecation of SpindleViz'es capability to reval the underlying structure of arbitrary GO portions. Indeed, visualization of such large discrete structures as the GO with tools such as SpindleViz and Flatland is best done within immersive, interactive virtual environments such as CAVEs and RAVEs.

However, we also intend to deploy SpindleViz within a web environment coupled to a simple query system allowing the specification not only of arbitrary GO nodes to visualize, but also expressions such as all the ancestors $\uparrow a$ or descendants $\downarrow a$ of a node $a \in P$, both ancestors and descendants $\Xi(a) = \uparrow a \cup \downarrow a$, the most specific common ancestor (least common subsumer) $a \vee b$ of a pair of nodes, and combinations (unions and intersections) of these. The resulting system will provide a high degree of utility in bioinformatics tasks such as microarray analysis and protein inference.

In this prototype implimentation, to aid in debugging and design, color codes for centrality, but this is redundant with the radius ρ. We intend to explore color mappings and other visual cues which show additional information, for example density of protein annotation, inheritance structure, user selections, etc.

References

1. Eric H Baehrecke, Niem Dang, Ketan Babaria and Ben Shneiderman: (2004) "Visualization and analysis of microarray and gene ontology data with treemaps", *BMC Bioinformatics*, 5:84 doi:10.1186/1471-2105-5-84
2. R Freese: (2004) "Automated Lattice Drawing", in: *Concept Lattices (ICFCA 04)*, LNAI v. **2961**, pp. 112-127
3. Gene Ontology Consortium: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. **25**:1, pp. 25-29
4. K Han and Y Byun: (2004) "Three-dimensional visualization of protein interaction networks", *Computers in Biology and Medicine*, 34 127-139
5. Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work* LNAI v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin
6. CA Joslyn, SM Mniszewski, A Fulmer and G Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177
7. Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston
8. KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2006) "A Categorization Approach to Automated Ontological Function Annotation", *Protein Science*, in press