BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 2 2007, pages 259–261
doi:10.1093/bioinformatics/btl496

Data and text mining

Sungear: interactive visualization and functional analysis of genomic datasets

Christopher S. Poultney[1], Rodrigo A. Gutiérrez[2,3], Manpreet S. Katari[2], Miriam L. Gifford[2], W. Bradford Paley[4], Gloria M. Coruzzi[2] and Dennis E. Shasha[1,*]

[1]Courant Institute of Mathematical Sciences, New York University, NY, USA, [2]Department of Biology, New York University, NY, USA, [3]Departamento de Genética Molecular y Microbiología, P. Universidad Católica de Chile, Santiago, Chile and [4]Digital Image Design Incorporated, NY, USA

Received on May 5, 2006; revised on August 8, 2006; accepted on September 20, 2006
Advance Access publication October 2, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Summary: Sungear is a software system that supports a rapid, visually interactive and biologist-driven comparison of large datasets. The datasets can come from microarray experiments (e.g. genes induced in each experiment), from comparative genomics (e.g. genes present in each genome) or even from non-biological applications (e.g. demographics or baseball statistics). Sungear represents multiple datasets as vertices in a polygon. Each possible intersection among the sets is represented as a circle inside the polygon. The position of the circle is determined by the position of the vertices represented in the intersection and the area of the circle is determined by the number of elements in the intersection. Sungear shows which Gene Ontology terms are over-represented in a subset of circles or anchors. The intuitive Sungear interface has enabled biologists to determine quickly which dataset or groups of datasets play a role in a biological function of interest.

Availability: A live online version of Sungear can be found at http://virtualplant-prod.bio.nyu.edu/cgi-bin/sungear/index.cgi

Contact: shasha@cs.nyu.edu

Supplementary information: Submitted—link TBD.

1 INTRODUCTION

The analysis of large datasets has become a necessary task for many researchers in the post-genome era. Genome sequences and microarray hybridizations are a common source of such data. Often, researchers want to see how different genome scale datasets relate to one another. For example, in comparative genomics, one might ask which genes are unique to a single species, which functionalities tend to be shared among related species, and which among remote species. Similarly, within an organism one might want to explore microarray data to determine which genes with which functionalities are expressed in subsets of multiple treatments and/or tissues.

There are several tools currently available to analyze and visualize genomic data (e.g. see Breitkreutz *et al*., 2003; Eisen *et al*., 1998; Shannon *et al*., 2003; Thimm *et al*., 2004). One of the most popular such methods combines clustering and heatmaps, (Eisen *et al*., 1998). Clustering typically is used to identify groups of genes that share expression profiles. A heatmap is a 2D grid in which each position in the grid is a colored box that represents the expression value of one gene in one experiment. Such a display conveys an intuitive impression of which sets of genes are similar to each other in expression. Sungear is complementary to such tools, because it facilitates the exploration of the overlapping relationships among different data lists. In this sense, Sungear is a generalization of the familiar Venn diagram (Venn, 1880). In most cases, a Venn diagram consists of a collection of three overlapping circles and can visualize the intersections among experiments. While asymmetric representations exist for four or more experiments (see Supplementary Figure 2D), they are difficult to interpret.

We have designed Sungear to support the analysis and visualization of an arbitrary number of datasets and Boolean combinations. These are presented within a framework that provides supporting annotation data [e.g. biological Gene Ontology annotations] and a statistical ranking of GO terms (based on Z-scores) which together enable the data to be understood in the context of biological functions in an interactive fashion. Further details of the scoring algorithm are given in Supplementary Material Appendix B.

2 DESCRIPTION

Figure 1 shows an example of the Sungear interface. The main Sungear window (bottom center) consists of a polygon whose vertices denote list names (in Fig. 1, the vertices represent species and the associated lists are lists of genes in each species). The circles inside the polygon denote intersections of the lists. The position of a circle is the average of the positions of the vertices whose intersection is denoted by that circle. We refer to the vertices as 'anchors' and the circles as 'vessels' since, in effect, each vessel is 'pulled' toward the anchors whose intersection the vessel represents. The size of each vessel is proportional to the number of genes in that intersection. The gene list window (left) shows the list of all genes in the Sungear plot and the controls window (top center) provides some navigation and optional export controls. The GO term window (right) gives hierarchical (top) and sorted (bottom) lists of GO terms.

*To whom correspondence should be addressed.

© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

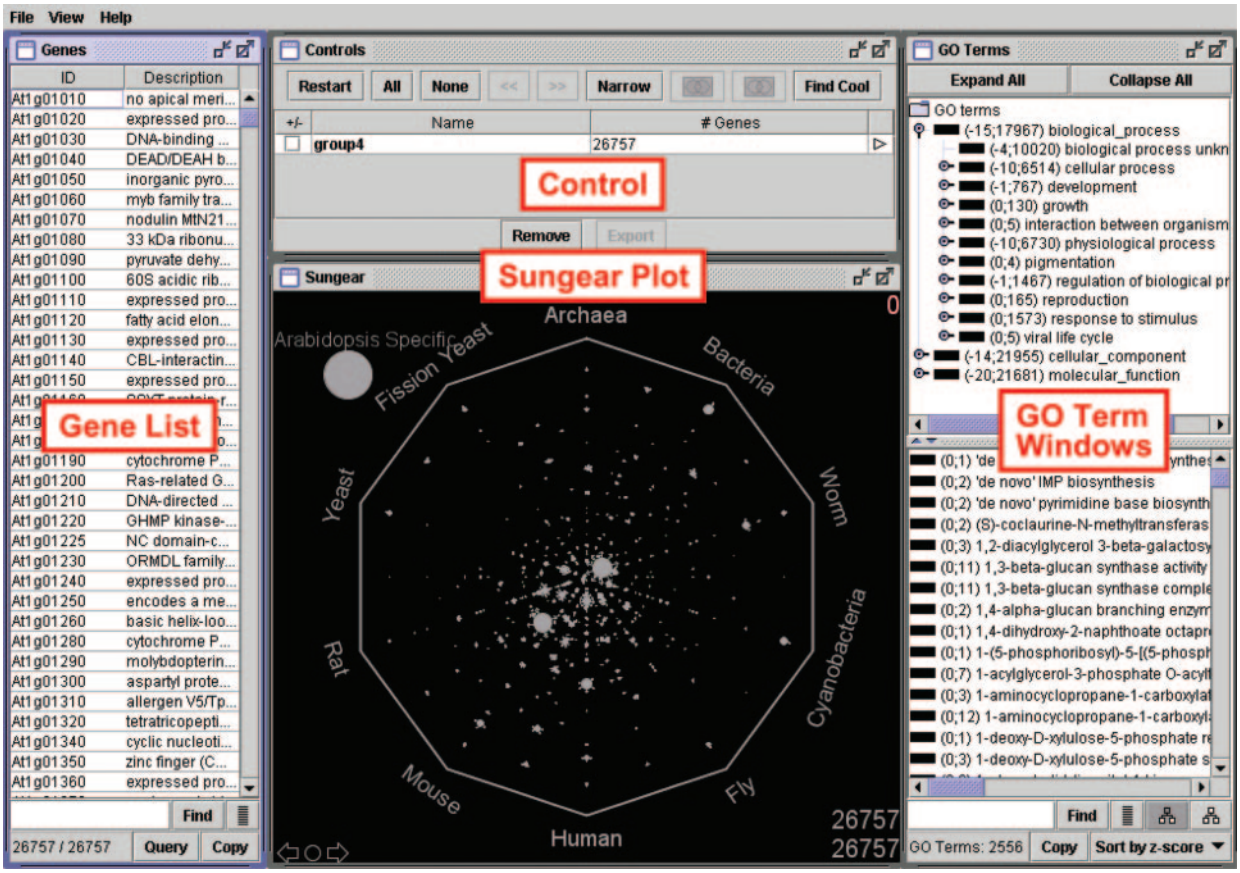

Fig. 1. A Sungear representation of data from a whole-genome comparison, showing the four Sungear windows. Within the Sungear plot (bottom center window), species 'anchors' on the vertices of the polygon contain lists of genes and encoded proteins shared between human, mouse, rat, fly, worm, fission yeast, yeast, Archaea, bacteria, or cyanobacteria.

Within each window, users can select genes, one or more vessels, anchors or GO terms. The selection can result in a Boolean intersection or union. The different windows are linked together, so that selections in one window will immediately be reflected in the other windows. This linking behavior is achieved by making genes the 'common currency' of Sungear: all windows provide a representation of a single master copy of the currently selected set of genes. Clicking on items in these multiple linked views drives the interactive visual exploration of list intersections, GO term associations and Z-scores, and individual genes. Sungear also provides a rapid way to identify the vessels containing genes that are most biased towards a specific biological functionality, implemented in the 'Find Cool' button of the navigation control window. 'Find Cool' ranks the vessels based on the number of functional GO terms that exceed a Z-score of 10, suggestive of statistical over-representation of genes in a biological function.

Sungear input data are provided by a simple text file giving anchor names, gene names and list membership. These files are easily generated from .csv or FASTA-like files using our conversion program, or can be created directly from microarray and other data using programs such as R. Sungear works well with large files, the primary limitations being available memory and the researcher's willingness to understand a plot with many vessels (Fig. 1 shows 10 anchors and over 26 000 genes). Gene and GO term annotations are provided by additional text files. Sungear can output lists of genes to other programs by standard copying and pasting, and can output directly to other web applications via the export facility in the controls window.

Although we have so far described Sungear in terms of genes and GO terms, Sungear is 'data agnostic': it can incorporate any species as well as data types other than genes (e.g. proteins or baseball players). Sungear also accommodates a wide range of users: it is written entirely in Java and can be run on nearly any computer via the web or as a stand-alone application.

2.1 Sungear case study: comparative genomics

Suppose one wants to investigate the degree to which particular biological processes are shared across the tree of life. The Sungear plot in Figure 1 shows the results of using BLASTP (E-value cutoff $\leq 1 \times 10^{-10}$) to compare genes encoding *Arabidopsis* proteins (hereinafter proteins) to genes encoding all the proteins in *Caenorhrabditis elegans* (worm), *Drosophila melanogaster* (fly), *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Saccharo myces cerevisiae* (yeast), *Schizosaccharomyces pombe* (fission yeast), and a collection of microbes including cyanobacteria, Archaea and bacterial proteomes (Gutierrez *et al*.,

2004). All proteins encoded in the *Arabidopsis* genome can be represented using Sungear. Circles inside the polygon in Figure 1 represent *Arabidopsis* proteins with various patterns of similarity to proteins in other organisms. The circle outside the polygon represents *Arabidopsis* proteins with no similarity to proteins in other organisms at the selected *E*-value cutoff.

We used Sungear to investigate the degree to which particular biological processes are shared across the tree of life (Supplementary Figure 1C). A cursory look at the data (Fig. 1) shows that the largest vessel in the Sungear window contains 13 562 proteins, and that these correspond to *Arabidopsis*–specific proteins, which include a high over-representation of 'unknown' proteins and proteins annotated to 'transcription factor activity' (Z-score = 14). This is consistent with the idea that divergence of factors that control gene expression is an important driver in plant evolution (Doebley *et al*., 1998) and also reflects the belief that transcription factor families have undergone a large expansion in plants (Shiu *et al*., 2005).

We next moved up the tree to analyze proteins common to all multicellular eukaryotes, but not in bacteria. For this, we used Sungear to select the vessel with 560 *Arabidopsis* proteins that are shared with human, mouse, rat, fly and worm. Among the top three z-scoring GO terms in this vessel are 'cyclic nucleotide binding' (Z-score = 25), and 'ion channel activity' (Z-score = 21) which play key roles in cell–cell signaling (Bridges *et al*., 2005). The observation that cell-cell signaling functions are linked to multicellularity shows that Sungear can uncover validated biological principles. Additional insight of how this case study enabled us to predict how biological functions evolved across the tree of life is detailed in the (Supplementary Material Appendix A).

2.2 Sungear case study: baseball scores

Sungear is indifferent to the data source. To illustrate this point, we used Sungear to visualize baseball statistics, data coming from a field far removed from biology. Changing the application domain of Sungear for this application is easy. In the baseball example, 'genes' are replaced by 'baseball players', and the 'GO hierarchy' is replaced by a 'league-team hierarchy'. We used publicly available team information and player statistics provided by The Baseball Archive (http://www.baseball1.com/statistics/). Using a very simple metric to determine 'remarkable' players during the 2004 season (small center vessel, Supplemental Figure 2A), we can quickly determine that these players are over-represented in the American League (Supplemental Figure 2B). A quick intersection between the two leagues shows that one of these players, Carlos Beltran, was active in both leagues during the season (Supplemental Figure 2C). Further exploration in the 'teams' (GO term) window shows that his respective teams for the American and National leagues were the Kansas City Royals and the Houston Astros.

3 CONCLUSIONS

Sungear is a tool that generalizes Venn diagrams to view multiple collections of genes, relates those collections to functional categories, and permits visual real-time, statistically-based data exploration. After minutes of training, users with few computer skills can comfortably navigate Sungear to explore and compare datasets. For the moderately sophisticated user, Sungear permits various data selection capabilities including 'and-functionality', 'or-functionality', and range selection. Sungear also provides support for the discovery of over-representation using any directed acyclic graph, such as the GO.

ACKNOWLEDGEMENTS

This work was funded by grants from the National Science Foundation (IIS-9988345, IIS-0414763, DBI-0445666 and IBN-0115586) to D.E.S.; grants from the National Science Foundation—N2010 (IBN-0115586) and (DBI-0445666) to G.M.C.; a grant from the National Science Foundation (DBI-0445666) to R.A.G.; and EMBO postdoctoral fellowship ALTF107-2005 to M.L.G. Funding to pay the Open Access publication charges for this article was provided by the U.S. National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Breitkreutz,B.J. *et al*. (2003) Osprey: a network visualization system. *Genome Biol*., **4**, r22.21–r22.24.

Bridges,D. *et al*. (2005) Cyclic nucleotide binding proteins in the *Arabidopsis thaliana* and *Oryza sativa* genomes. *BMC Bioinformatics*, **6**, 6.

Doebley,J. and Lukens,L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell*, **10**, 1075–1082.

Eisen,M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gutierrez,R.A. *et al*. (2004) Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? *Genome Biol*., **5**, R53.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Shiu,S.H. *et al*. (2005) Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol*., **139**, 18–26.

Thimm,O. *et al*. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J*., **37**, 914–939.

Venn,J. (1880) On the diagrammatic and mechanical representation of prepositions and reasonings introducing diagrams known today as Venn diagrams. *Philosophical Magazine J. Sci*., **5**, 1–18.