GoSurfer: A graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space

Sheng Zhong[1], Kai-Florian Storch[2,], Ovidiu Lipan[1,4], Ming-Chih J. Kao[1], Charles J. Weitz[2] & Wing H. Wong[1,3,*]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA

[2]Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA

[3]Department of Statistics, Harvard University, Science Center, Cambridge, Massachusetts 02138, USA

[4]Present address: Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta, GA 30912, USA

* To whom correspondence should be addressed. Email: wwong@hsph.harvard.edu

Abstract:

The analysis of complex patterns of gene regulation is central to understanding the biology of cells, tissues, and organisms. Patterns of gene regulation pertaining to specific biological processes can be revealed by a variety of experimental strategies, particularly by microarrays and other highly parallel methods, which generate large datasets linking many genes. While methods for detecting gene expression have improved substantially in recent years, understanding the physiological implications of complex patterns in gene expression data is a major challenge. Here we present GoSurfer, an easy-to-use, graphical exploration tool with built-in statistical features that allows a rapid assessment of biological functions represented in large gene sets.

GoSurfer takes one or two list(s) of gene identifiers (Affymetrix probe set ID, LocusLink ID or Unigene ID) as input, and retrieves all the Gene Ontology (GO) terms that are associated with the input genes. GoSurfer visualizes these GO terms in a hierarchical tree format. With GoSurfer, users can perform statistical tests to search for the GO terms that are enriched in the annotations of the input genes. These GO terms can be highlighted on the GO tree. Users can manipulate the GO tree in various ways and interactively query the genes that are associated with any GO terms. The user generated graphics can be saved as graphic files, and all the GO information related to the input genes can be exported as test files.

GoSurfer is freely available at **http://www.gosurfer.org**

Keywords: GoSurfer, Gene Ontology, Data mining, Visualization, Comparative analysis

1. Introduction

GoSurfer employs the Gene Ontology (GO) resource[1], which dynamically structures biological knowledge using a controlled vocabulary consisting of GO terms. GO terms are organized in three general categories, "biological process, "molecular function," and "cellular component," and the terms within each category are linked in defined parent-child relationships that reflect current biological knowledge (see below). On the basis of accumulated information, individual genes from all organisms are systematically associated to GO terms, and these associations continue to grow in complexity and detail as sequence databases and experimental knowledge grow.

GoSurfer is based upon the GO annotation tables linking genes and GO terms (for current annotations, see www.geneontology.org). For analysis, a list of genes (from any kind of experiment) tagged with typical identifiers (Affymetrix probe set ID, Unigene ID, LocusLink ID) is uploaded, and the genes are then matched to their associated GO-terms in the GoSurfer database. As output, GoSurfer constructs a hierarchical tree display (for any of the three categories of GO) in which each branchpoint or node is a GO term that made a match to a gene in the uploaded gene set. The user can control the sensitivity and complexity of the output by setting a threshold for the number of genes that must match a given GO term in order for that node to be displayed.

2. GoSurfer in Action

2.1. Mapping genes onto the GO space

An example of the graphical output of GoSurfer is shown in Fig. 1. The tree display represents biological processes that matched to a set of 575 genes exhibiting circadian rhythms of expression in mouse liver[2]. This tree constitutes only a subset of the complete GO biological process category.

The total biological information represented in the tree structure can be quickly explored simply by using the mouse of a personal computer. When the cursor is directed to a node, the node and the path in the tree leading to it are highlighted, and the GO terms corresponding to the selected node and its parents appear in the status line (Fig. 1, arrows, "intracellular signaling cascade"). Clicking on a node opens a pop-up window displaying all of the genes in the uploaded list that matched to this node (Fig. 1). The window lists both official gene names and LocusLink (LL) identifiers, which are hyperlinked to the LocusLink database (ncbi.nih.gov/LocusLink). Clicking on a given LL identifier opens a browser window displaying the entire LocusLink entry, which provides comprehensive information on the selected gene. Horizontally sweeping the cursor across the tree leads to a succession of GO terms appearing in the status line, quickly revealing the diversity of biological processes associated with the genes in a dataset. Moving the cursor from the top to the bottom of the tree reveals the current extent of knowledge about the genes in the dataset, with longer paths corresponding to more detailed information.

Fig. 1 also reveals a minor drawback of displaying the relationships among GO terms as a simple hierarchy. To incorporate the complexity of biological knowledge, each GO vocabulary is structured as a directed acyclic graph (DAG), which allows a given child term to be linked to multiple parent terms. Thus unfolding a GO DAG into a hierarchical tree display inevitably results in certain GO terms appearing more than once. For instance, the GO term "steroid biosynthesis" appears three times in Fig. 1 (magenta nodes). However, the path of every GO term upward towards its top-level parent is unique, so there is only a single biological context for each node. We feel that the benefits of a simple tree display clearly outweigh this limitation.

GoSurfer has built-in statistical tools for the comparative analysis of large gene sets. Fig. 2 shows the biological processes found by GoSurfer to be significantly associated with genes showing altered expression in prostate cancer in comparison with normal prostate. We used the raw data from a microarray study of gene expression in 52 prostate tumor specimens and 50 normal prostates[3]. Using the entire 102-chip dataset, we identified 1261 genes that were significantly up-regulated and 1808 that were significantly down-regulated in cancerous compared to normal prostate (t-test, $P < 0.02$; Supplementary Note 1 online). The two lists corresponding to up- and down-regulated genes, respectively, were uploaded into GoSurfer, and a biological process tree was constructed from GO terms making a match to at least one gene in either set. This tree, derived from the union of the two gene sets, provides a framework for comparative analysis (Fig. 2, gray tree structure).

2.2 Comparing gene groups

By means of a Chi-square test, GoSurfer then examines the entire tree structure for nodes that are significantly over- or under-occupied by genes from one set compared to the other (see Supplementary Note 2 online for detailed statistical methods). Nodes meeting a user-defined P-value threshold are color-coded. The example in Fig. 2 highlights biological processes that are significantly associated ($P < 0.01$) with genes induced (magenta) or repressed (blue) in prostate cancer compared to normal prostate. The predominance of blue in the tree display suggests that repression of gene expression could be a major factor contributing to tumor phenotype. Interestingly, a different microarray study on prostate cancer reported that metastatic tumors were distinguished from nonmetastatic by a larger number of down-regulated genes[4,5].

Biological processes that were significantly associated with genes up-regulated in prostate cancer included "protein metabolism" and "protein biosynthesis" (Fig. 2, nodes 8, 9). Of the 64 structural components of the ribosome represented on the array (node 9), 59 were up-regulated in prostate tumors (see Supplementary Table 1 online), perhaps reflecting aberrant proliferative control or energy metabolism. Biological processes that were associated with genes down-regulated in prostate cancer included "regulation of cell proliferation" (node 13), "muscle contraction" and "muscle development" (nodes 14, 15), and a pathway representing surface cell surface receptor signal transduction" (nodes 2, 10, 11, 12**)**. Down-regulation of genes involved in the regulation of cell proliferation points to the expected defect of mitotic control in cancer cells. The down-regulated genes in this node (Supplementary Table 2 online) include several well-known tumor suppressors, but also many positive regulators of cell proliferation, implying that the regulatory circuits for mitotic control are generally perturbed in prostate tumors. Down-regulation of genes associated with muscle contraction and muscle development (Supplementary Tables 3 and 4 online) perhaps reflects the de-differentiated phenotype of tumor cells, since the normal prostate gland is a contractile organ containing smooth muscle cells. Suppression of genes associated with cell communication (node 2) and cell surface receptor signaling (nodes 11, 12) in prostate cancer (Supplementary Table 5 online) is in agreement with the reported insensitivity of cancer cells to exogenous anti-growth signals[6].

3. Conclusion

As illustrated in the prostate cancer example, GoSurfer's intuitive and interactive tree visualizes the entire biological content of a dataset literally at a glance, without the need for scrolling or browsing through multiple windows. GoSurfer's built-in statistical tools allow the functions represented in

large gene sets to be compared and analyzed rigorously, providing the means to extract genetic networks embedded in gene expression data or other experimentally-derived lists of genes. GoSurfer is designed to use transparent databases that can be updated at will by the user[7], so that the sophistication of its analyses will grow as knowledge about gene function grows. GoSurfer encourages a broad, systems level understanding of the biology and permits links to be made between what might otherwise be considered isolated, independent processes.

4. Accessibility

GoSurfer is a windows-based program freely available for non-commercial use and can be downloaded at http://www.gosurfer.org Supplementary materials to this paper are available at: http://www.gosurfer.org/Supplementary materials.htm. Data sets used to construct the trees in Fig. 1 and 2 are available at http://www.gosurfer.org/download/GoSurfer.zip.

5. Acknowledgements

We thank Dr. Cheng Li for valuable discussions and suggestions. We thank Editor Dr. Allen Rodrigo for valuable suggestions on manuscript revision. This work was supported by grants of the National Cancer Institute (W.H.W).

References

1. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet.* **25,** 25-29 (2000)

2. Storch, K.F. *et al.* Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**, 78-83 (2002)

3. Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. **1**, 203-209 (2002)

4. Dhanasekaran, S.M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412,** 822-826 (2001)

5. Varambally, S *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629 (2002)

6. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100,** 57-70 (2000)

7. Zhong, S., Li, C. & Wong, W.H. ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Research* **31**, 3483-6 (2003).

Fig. 1 GoSurfer screen shot displaying a GO tree of the biological process category, where each node represents an individual GO term. All GO terms at display are associated with at least one out of 575 genes which have been selected for circadian expression in mouse liver[2]. For clarity, only terms (nodes) that are associated with at least 4 genes from the data set are shown. The path to the GO term "intracellular signaling cascade" is highlighted in red and corresponds to the terms displayed in the status line (arrows**).** Inset: pop-up window displaying all genes in the data set that are associated with the GO term "intracellular signaling cascade." Redundant nodes corresponding to the GO term "steroid biosynthesis" are highlighted in magenta (see text). Selected nodes are marked with numbers, and the corresponding GO terms are listed underneath the tree structure.

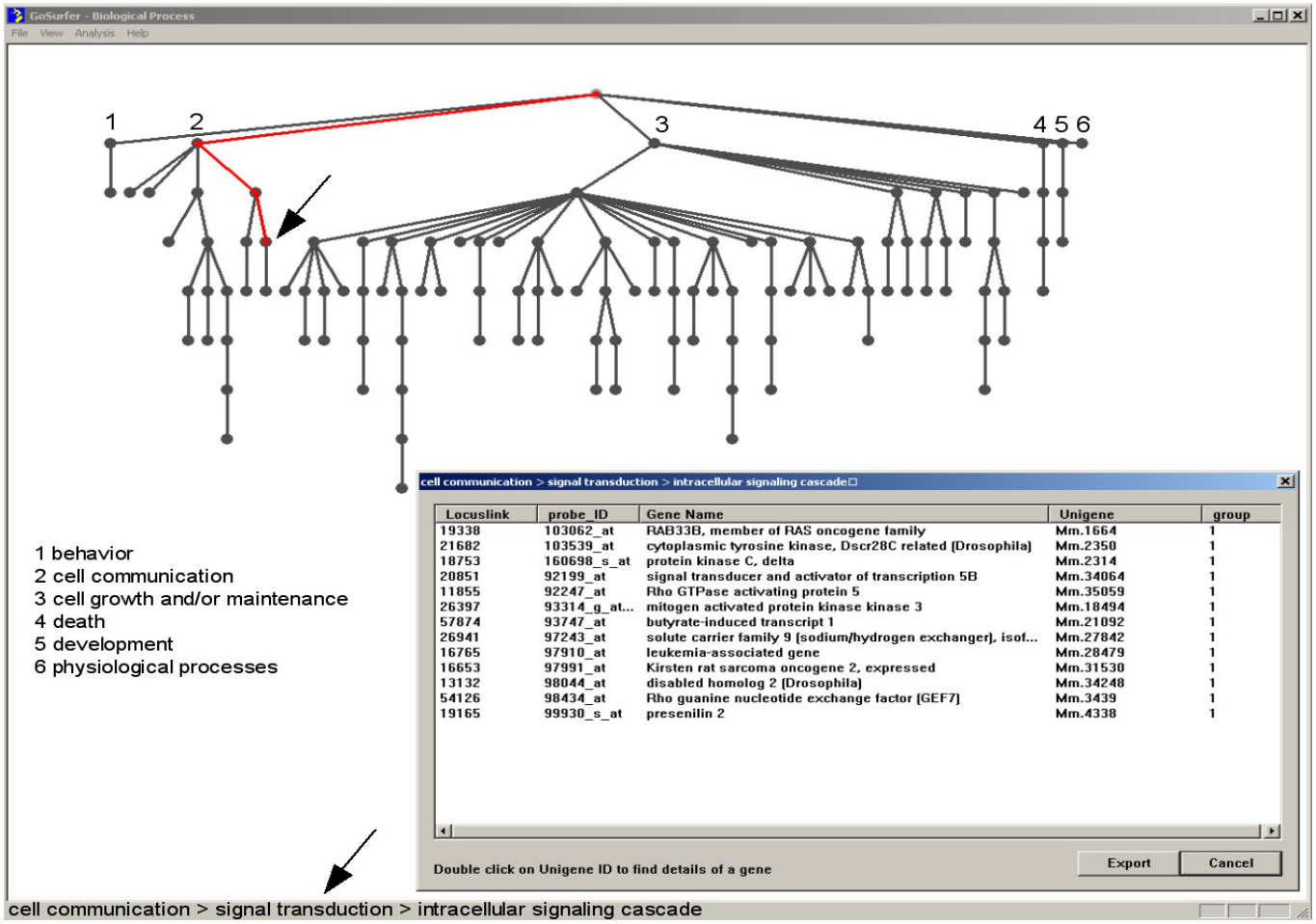
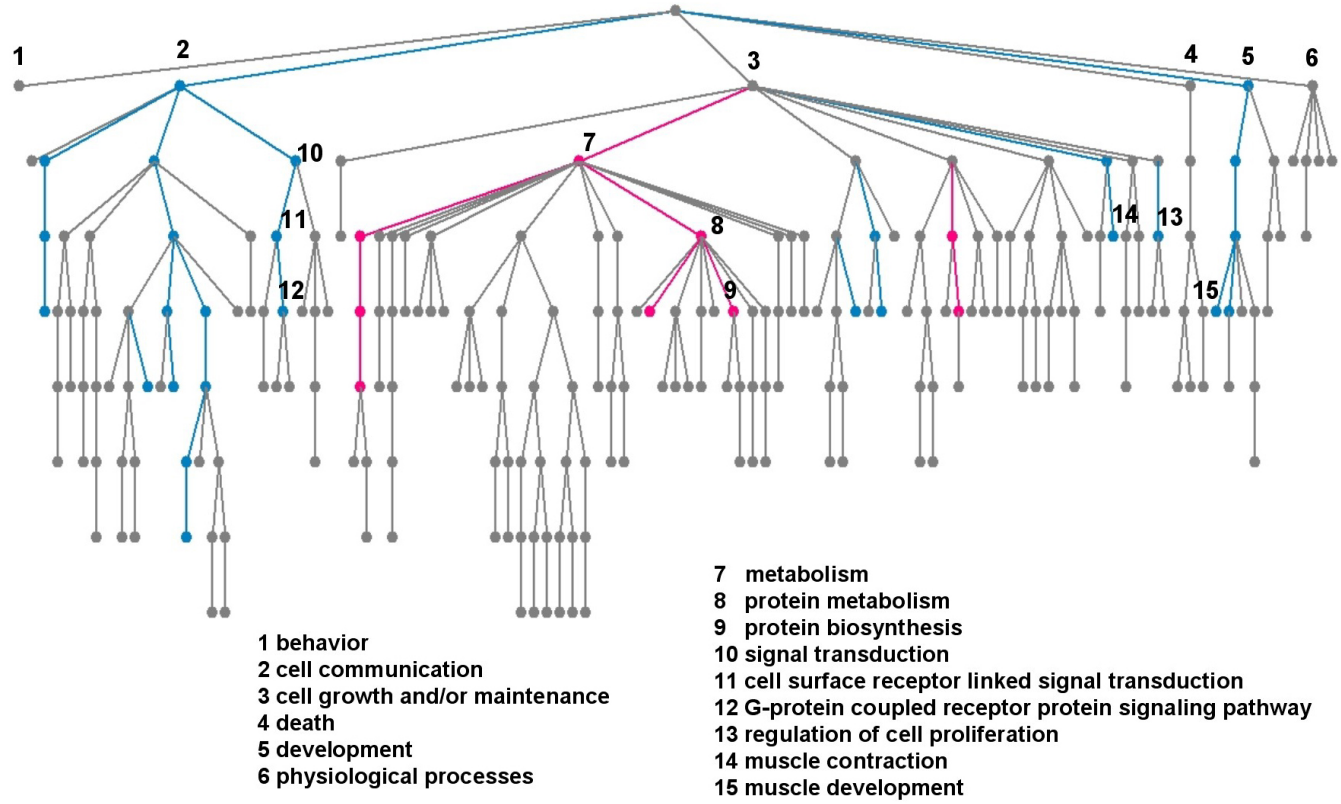
Fig. 2 GoSurfer comparison of biological processes significantly (P <0.01) associated with genes up-regulated (magenta) or down-regulated (blue) in prostate cancers compared with normal prostate. For clarity, only terms that are associated with at least 10 genes of the induced or repressed gene sets are shown. Selected nodes are marked with numbers, and the corresponding GO terms are listed underneath the tree structure.