Dynamic visualization of multi-level molecular data

Icay, Katherine

2018-01


http://hdl.handle.net/10138/298158
https://doi.org/10.1016/j.cmpb.2017.10.013

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
Dynamic visualization of multi-level molecular data: The Director package in R

Katherine Icay, Chengyu Liu, Sampsa Hautaniemi

Research Programs Unit, Genome-Scale Biology, Faculty of Medicine, University of Helsinki, Helsinki, POB 63, 00014, Finland

ARTICLE INFO

Article history:
Received 6 July 2016
Revised 23 February 2017
Accepted 10 October 2017

ABSTRACT

Background and objective: High-throughput measurement technologies have triggered a rise in large-scale cancer studies containing multiple levels of molecular data. While there are a number of efficient methods to analyze individual data types, there are far less that enhance data interpretation after analysis. We present the R package Director, a dynamic visualization approach to linking and interrogating multiple levels of molecular data after analysis for clinically meaningful, actionable insights.

Methods: Sankey diagrams are traditionally used to represent quantitative flows through multiple, distinct events. Regulation can be interpreted as a flow of biological information through a series of molecular interactions. Functions in Director introduce novel drawing capabilities to make Sankey diagrams robust to a wide range of quantitative measures and to depict molecular interactions as regulatory cascades. The package streamlines creation of diagrams using as input quantitative measurements identifying nodes as molecules of interest and paths as the interaction strength between two molecules.

Results: Director’s utility is demonstrated with quantitative measurements of candidate microRNA-gene networks identified in an ovarian cancer dataset. A recent study reported eight miRNAs as master regulators of signature genes in epithelial-mesenchymal transition (EMT). The Sankey diagrams generated with data from this study furthers interpretation of the miRNAs’ roles by revealing potential co-regulatory behavior in the extracellular matrix (ECM). An additional analysis identified 32 genes differentially expressed between good and poor prognosis patients in four significant pathways (FDR < 0.1), three of which support a complementary role of the ECM in ovarian cancer. The resulting diagram created with Director suggest elevated levels of COL11A1, INHBA, and THBS2 — a signature feature of metastasis [1] — and decreased levels of their targeting miRNAs define poor prognosis.

Conclusion: We have demonstrated a visualization approach suitable for implementation in an analysis workflow, linking multiple levels of molecular data to gain novel perspective on candidate biomarkers in a complex disease. The diagrams are dynamic, easily replicable, and rendered locally as HTML files to facilitate sharing. The R package Director is simple to use and widely available on all operating systems through Bioconductor (http://bioconductor.org/packages/Director) and GitHub (http://kzouchka.github.io/Director).

© 2017 Published by Elsevier Ireland Ltd.

1. Introduction

Large research projects such as The Cancer Genome Atlas (http://cancergenome.nih.gov/) are built on the premise that understanding the molecular basis of a cancer can be improved with the sequencing and comprehensive analysis of multiple, related levels of molecular data [2–4]. Such multidimensional data makes possible analyses that, for example, identify a set of microRNAs (miRNAs) and their target genes as ‘master regulators’ of a tumor subtype associated to poor overall survival [5]. Explaining the context in which a list of molecules are functionally contributing to a disease phenotype, however, is not straightforward. The major challenge thus lies in the effective interpretation of multidimensional data into biological and clinically relevant implications.

Data visualization offers an effective means to summarize results from multiple data levels, to reveal overall trends, and to relate key findings. Modern visualization technology enables automatic generation and manipulation of diagrams efficiently from data [6]. While visualizations succeed in making data more appealing and approachable to a general audience, the addition of dynamic elements further engages users to explore and interact with the data. Users thus gain a more intuitive understanding of
the data and, consequently, improved discovery of biomedical insights [7–9].

The R programming language and software environment [10] provides a robust and reliable medium for data processing, analysis, and visualization — including integration with state-of-the-art graphical techniques. The Bioconductor project, a repository of R packages, has become a popular data-analysis framework facilitating and enriching biomedical research [11].

We introduce a novel implementation of a directional flow diagram in the R package Director to depict downstream molecular effect and potential regulatory cascades from result data. The remaining article is organized in the following sections: a rationale for the package implementation and design, a summary of package features, a case study in ovarian cancer to demonstrate usage and results, discussion of data-driven visualization tools currently available and the practical role Director fulfills.

2. Rationale

Much effort has gone into the development of computational algorithms to identify potentially interesting molecules and molecular networks in disease [12,13]. The interpretation of biologically and clinically relevant interactions from these molecules, however, is often slow and error-prone manual work. This does not have to be the case, and the search for biological meaning can be sped up significantly by harnessing the power of dynamic, data-driven visualization in the analysis process. Equipping researchers with the ability to create and explore links between molecules in a familiar software like R would not only improve understanding of the data but accelerate the discovery of biomedical insights.

Molecular interactions are generally depicted as network graphs [12]. However, regulation is a form of molecular interaction that, with consideration of downstream effects, could be better represented as a directional flow. This is the heart of Director’s primary function: to visualize molecular interaction as a flow of biological information. The D3 library [6] contains a plugin to create a type of directional flow diagram called a Sankey diagram. These diagrams comprise of paths and nodes rendered to represent flow measurements within a system of distinct events. Larger downstream contributions are emphasized by making path widths proportional to input measurements [14]. This inherent ability to depict causal, downstream effect makes Sankey diagrams exceptionally good at representing quantitative flows through multiple, distinct events.

3. Package description

Director is an exploratory visualization and interpretation tool. To automatically generate dynamic diagrams from R tables and data frames, Director ‘directs’ the use of multiple web technologies: HTML for page content, CSS for user-defined aesthetics, JavaScript for dynamics and use of SVG for graphic elements. The package, however, only requires users to know R to create the diagrams.

Director is implemented as an R package available through the Bioconductor project and Github. It makes use of the D3 JavaScript library [6] to generate dynamic graphical elements from input data to an HTML document. D3 renders more specialized plots such as Sankey diagrams as an additional plugin (https://github.com/d3/d3-sankey).

Director introduces novel drawing capabilities to D3’s Sankey plugin to improve depiction of molecular interactions in regulatory cascades. By enabling input for both path and node values, colors are meaningfully assigned according to a gradient scale rather than being assigned randomly or uniformly. Furthermore, traditional Sankey diagrams determine node and path sizes based on a single, minimum to maximum scale of positive input values. However, in molecular biology both positive and negative measurement values can be interesting, so Director was developed (1) to allow for negative input values and (2) to automatically scale the sizes of and assign colors to nodes and paths emphasizing both extremes.

In Director, nodes represent molecules and a path represents the interaction strength between two molecules. Nodes can be represented by quantitative values such as expression, fold-change, and methylation score. Similarly, paths can be represented by predicted binding affinity scores [15,16] or expression correlation values. In such cases where a connection is interesting but no dataset-specific quantitative value is provided, Director can supply a value based on the average downstream contribution of preceding paths. This is useful, for example, in visualizing drug treatment effect.

Quantitative values input as R tables or data frames determine color assignment, size, and organization of nodes and paths in a diagram. Molecules with strong quantitative values are emphasized with bold colors and strong molecular interactions emphasized as thick paths. As a result, molecules with multiple and strong interactions become large nodes. A connected series of such molecules would, in turn, identify potentially interesting regulatory cascades — visual information not as easily nor intuitively conveyed with standard network approaches.

By default, Director uses a red-blue palette for assigning colors to nodes and paths based on quantitative measures. However, it can use any color-scheme defined with color names or hexadec-imal codes. The color range for paths is also independent of that for nodes. The package assumes larger absolute values are stronger than smaller values by default. For the reverse case (e.g. statistically significant p-values), values can be defined for paths and nodes such that stronger colors are given to smaller values and weaker colors to larger values. These features make the creation of diagrams with Director robust to a wide range of quantitative values used in biomedical research, with diagrams able to match almost any color theme.

The diagrams produced are dynamic, which engages users to directly explore quantitative data. Using existing features of the D3 library and D3 Sankey plugin, nodes created with Director can be moved around the space for optimal presentation. Quantitative and qualitative information are directly accessible from nodes and paths via mouse-over (Fig. 1). Additionally, entire pathways for molecule(s) of interest can also be highlighted with a simple mouse-click (Fig. 2A and B). Altogether, these features facilitate active hypothesis formation. For example, a common result in biomedical research is a set of differentially expressed molecules in disease. With Director, expression values and known interactions of these molecules can visually identify the potential effect of expression change in disease.

Diagrams are loaded and saved as HTML documents to preserve information dynamics and publication-quality vector graphics. HTML also enables basic formatting such as diagram dimensions, fonts, and the addition of extra descriptive content through Director. The R package htmltools [17] facilitates user-customization and loading of HTML content quickly and effectively. The resulting visual documents are compact and viewable on any computer with an internet browser. Additionally, the diagrams are rendered and made available entirely within the local R working directory, making it simple to share the results. Multiple diagrams containing different information content can also be saved to the same directory to conveniently share JavaScript and style sheet resources.

Data to render the diagrams are stored as R data frames consisting of interacting molecule pairs. These data frames are normally saved and loaded in R, but supply the necessary values determining how information dynamics are rendered. Supporting func-
tions facilitating user-modification of input data frames and diagram properties are documented in the package manual.

Director has few limitations. First, Sankey diagrams are strictly directional flow diagrams and are not suitable for depicting feedback loops (i.e. a path from one node that leads back to itself). Paths that loop also trigger infinite cycles with the Javascript plugin, a known issue that prevents diagram rendering. Director checks for such loops in the data-frame and notifies the user. It then renders the target of a loop as a new node. Second, as with all data-driven diagrams, large sets of molecules can be difficult to visualize and still be able to distinguish individual nodes and paths. Director can easily render up to 200 connected molecules in a single diagram. As the number of molecules grows, so does the diagram. While this package does not restrict the amount of paths and nodes defined, it will take proportionally longer amounts of time to perform color assignments. It will also warn users that the diagram may not be rendered correctly. In extreme cases, less-connected paths and small nodes will not be visible. In such a situation, users have two options: (1) filter the data manually by using the quantitative and qualitative filters of the package, and/or (2) manually define the diagram height and width to be large enough to render all nodes and paths. Instructions on how to perform these actions are described in the manual.

4. Application of Director to high-grade serous ovarian cancer data

We demonstrate the utility of Director with results from two analyses of high-grade serous ovarian cancer (HGSOC) obtained from The Cancer Genome Atlas [2]. Epithelial ovarian cancer is the sixth leading cause of female cancer deaths in Europe and HGSOC is the largest and most lethal epithelial ovarian cancer subtype. Though a majority of patients respond well to first-line therapy, cancer usually relapses resulting a 5-year survival rate of less than 50% [18]. Thus, novel treatment options are needed to improve HGSOC patients’ survival.

Yang et al. [5] identified a master miRNA and gene regulatory network associated to poor prognosis in the mesenchymal subtype of serous ovarian cancer. The study highlighted eight miRNAs as master regulators of signature genes in epithelial-mesenchymal transition (EMT), a process linked to tumor invasiveness and progression, and demonstrated their potential as therapeutic tools against this more aggressive ovarian cancer subtype. We re-annotated the 219 mesenchymal-associated genes predicted to have targeting miRNAs to known KEGG pathways (details in supplementary).

Mesenchymal-associated genes annotated to five significantly enriched pathways (FDR < 0.1) and putatively regulated by the eight key miRNAs were visualized with Director (Fig. 1). Node colors represent expression fold-change, with upregulation in the mesenchymal subtype in blue and downregulation in red. Nodes with bold shades of red and blue indicate large absolute fold-change expression of the molecules they represent. Path color represents Pearson correlation of microarray expression across 459 cases, with darker paths indicating stronger correlation between miRNAs and their putative targets. Paths connecting genes to the pathways represent the average correlation of gene expression to targeting miRNAs.

While Yang et al. [5] focused on EMT inhibition by miR-506, our analysis suggests a functional role also in the related extracellular matrix (ECM) pathways. Furthermore, co-regulation of the same or similar genes by the other key miRNAs suggests that regulation of cell-matrix interactions is a complementing process to EMT that the core miRNA-gene network modulates in poor prognosis mesenchymal subtype.

We hypothesized that comparison of patients with the longest survival time after diagnosis (good prognosis) to those who succumbed early (poor prognosis), irrespective of transcriptional subtype, could reveal important genes and pathways characterizing differences between long and short survival rates. Analysis of good versus poor prognosis ovarian cancer data is detailed as supplementary and the significant miRNA-gene network visualized in Fig. 2. Our findings further support the idea of a miRNA-gene net-
Fig. 2. Genes differentially expressed between good and poor prognosis tumor samples, their predicted targeting miRNAs with anti-correlated expression, and the pathways in which the genes are enriched (FDR < 10%). The diagrams are produced with (A) default and (B) custom color and font settings. Node colors represent expression fold-change with upregulation in the good prognosis group in red (purple) and in the poor prognosis group in blue (orange). For biological pathways, a fixed value of –2 was given if upregulated in poor prognosis samples. Path color represents Pearson correlation of normalized microarray expression, with darker paths indicating stronger correlation between miRNAs and their targets. Paths connecting genes to the pathways present average expression correlation as the effect of targeting miRNAs. Paths connecting to Focal Adhesion are highlighted with mouseclick. A simple network graph (C) rendered with the same node information is provided for comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
work modulating poor prognosis HGSOC through cell-matrix interactions (FDR < 0.1). Notably, the diagram produced with Director indicated elevated levels of COL11A1, INHBA, and THBS2 — a signature feature of metastasis [1] — and decreased levels of their targeting miRNAs in poor prognosis.

While miRNA regulation of the ECM is well-documented in some cancers [19], the role in HGSOC is still developing [20]. To our knowledge, the putative connection between miRNAs and collagen genes have not yet been validated. Cheon et al. [21] has shown that collagen genes contribute to poor overall survival in HGSOC. Ren et al. [22] has shown that ectopic expression of miR-509-5p inhibits cervical cancer cell proliferation and invasion. Visualizations generated with Director thus identified genes and their putative regulatory miRNAs with active roles in ECM and ECM-related pathways as candidate prognosis predictors of ovarian cancer treatment outcome and warrant further examination.

The major aims of visualizing multidimensional molecular data are (1) to combine key analysis results in such a way that allows users to intuitively understand how the data is linked and (2) to glean insights from the flow of biological information. In other words, to create ‘a picture worth a thousand words’. Fast and efficient high-throughput analysis of molecular data is facilitated with R projects like Bioconductor, but the ability to conceptualize biological relevance from key results remains an exclusively human skill [23–25]. The latter is greatly improved when visualizations also allow exploratory analysis of the data represented [26]. Co-regulation by miRNAs, for example, is a biological feature that is simpler to understand visually than statistically by lay people and experts alike.

Several visualization techniques can be applied to a set of data, but the choice of technique affects how ideas and the overall interpretation of key results is conveyed [14]. The visual advantage of Sankey diagrams is the inherent ability to depict causal downstream effect. It is commonly used in industry to depict energy and material flow [14]. More recently, it has been used in healthcare to represent disease progression [8] and patient cohorts [9] over time to help reveal previously undetected outcome associations in the data. To the best of our knowledge, Director is the first application of Sankey diagrams with molecular-level data to depict biological regulatory cascades. Novel application to gene expression data in the case study has revealed a broader picture of how a miRNA-gene network potentially contributes to complex disease.

Director was developed as a post-analysis visualization tool to supplement statistical results. In the case study, Director was used to explore why a set of molecules with statistically significant differential expression could also be biologically relevant through their known interactions. The package uses color assignment to depict two types of expression-based values: fold-change for nodes and correlation for paths. Fold-change represents the amount of expression change found statistically significant for a molecule in the original study, and correlation the interaction strength between miRNAs and their predicted target genes. Molecules were also filtered for membership in significantly enriched biological pathways before rendering. This is different from global first line analyses such as clustering and principal component analyses that identify sets of interesting molecules based on similar data measurements; Director is used to visually explore the values of these molecules for why these sets could be potentially interesting in a biological context. In the case study, the diagram revealed miRNAs regulating collagen genes as plausible, candidate biomarkers of poor prognosis HGSOC. We believe any effort that furthers understanding of

**Box 1.** R script to reproduce Fig. 1.

```r
# Import data
data(Director)
data(ovca)
initSankey() # Generates necessary javascript and CSS templates.
Level1 <- createList(mesenchymal$Level1) # miRNA-gene data
Level2 <- createList(mesenchymal$Level2) # gene-pathway data.
List <- append2List(Level1, Level2) # append lists.

# Make Sankey diagram such that the effect value on
# pathways is the average (correlation) effect on all
# genes connected.
Sankey <- makeSankey(List, averagePath = TRUE)

drawSankey(Sankey,
caption="Top SPIA pathways with pGFdr < 0.1.<font size=-1>Nodes are
blue if upregulated in mesenchymal subtype.</font>",
width=1200, height=600,
legendfont="lato, helvetica, sans-serif",
legendsize=22, nodeValue="Fold change")

# Write to working directory as 'mesenchymal.html'
writeSankey(name="mesenchymal")
```
# Import data
library(Director)
data(ovca)
Level1 <- createList(poorprog$Level1) # miRNA-gene data
Level2 <- createList(poorprog$Level2) # gene-pathway data.
List <- append2List(Level1, Level2) # append lists.

# Figure 2A with default settings
initSankey() # Generates necessary javascript and CSS templates.
template1 <- makeSankey(List, averagePath = TRUE)
Sankey1 <- drawSankey(template1,
caption="Top SPIA pathways with pGFdr < 0.1.<br><font size=-1>Nodes are red if upregulated in good diagnostic samples.</font>",
width=1100, height=500,
legendfont="lato, helvetica, sans-serif",
legendsize=20, nodeValue="Fold change")

# Figure 2B with custom settings.
# Information dynamics are fixed but the look of a figure is dramatically altered with a few extra parameters. See manual for details on parameter use.
initSankey(fontsizeProportion=FALSE, font="times new roman", pathHover=0.6)
template2 <- makeSankey(List, averagePath = TRUE, noughtPath=-0.3,
noughtPathColor="#cccccc", pathMin="#333333", nodeMin="orange",
nodeMax="purple")
Sankey2 <- drawSankey(template2,
caption="Top SPIA pathways with pGFdr < 0.1.<br><font size=-1>Nodes are purple if upregulated in good diagnostic samples.</font>",
width=1000, height=400,
legendfont="times new roman",
legendsize=18, nodeValue="Fold change")

# Write latest Sankey (Figure 2B) to working directory as 'poorprognosis.html'
writeSankey(name="poorprognosis")

# Figure 2C
library(networkD3)
simpleNetwork(List, Source="source", Target="target", fontSize=15, opacity=0.8,
charge=-200)

Box 2. R script to reproduce Fig. 2A, B and C.

molecular function in complex diseases will also lead to improved prediction algorithms and better treatment options. Comparison of overlapping miRNA-gene connections between our case studies and the regulatory network of Yang et al. [5] show two benefits to using Director’s customized Sankey diagrams over standard network visualization approaches in representing gene regulation. First, the organization of nodes and paths in a directional flow makes clear the biological levels represented in the data. Strong regulatory and co-regulatory effect of key miRNAs are emphasized, highlighting potentially interesting regulatory cascades and roles in mediating cell-matrix interactions. Second, the ability to custom-define path and node colors with a value-based gradient scale adds the visual information necessary to depict the estimated contribution of a gene’s altered expression on biological pathways. This feature is not readily available with other R packages utilizing the D3 library to render Sankey diagrams [27,28].
Visual representation of the miRNA-gene network differentiates good and poor prognosis HGSOC patients with Director is given in Fig. 2A and B. For comparison, a basic network graph was rendered with the R package networkD3 [28] in Fig. 2C. The graph also uses the dynamic abilities of the D3 library and could be further customized with directional arrows for paths, node colors, and path widths reflecting node interaction strengths. However, the structural layout of the network graph would have remained the same. Regulatory paths and the estimated contribution of each molecule are clearer in the directed structure of Sankey diagrams. Additionally, node colors would have been assigned by group classification in the network graph and not a value-based gradient scale like in Director's Sankey diagram.

The resulting diagrams are dynamic but not interactive. Though multiple diagrams can be produced quickly and efficiently, data written to HTML documents cannot be further manipulated via R to preserve the information content. Packages like Shiny [29] and Rserve [30] utilize web servers to power on-demand data manipulation to change the information content of a diagram. Code must be re-executed with Director to update the information content of a diagram. In the future, we will explore the incorporation of such web applications for more interactive browsing and manipulation of custom diagrams.

The generation of high-volume data has lead to the popularity of web applications to interact with and visually explore data [17, 26, 29, 31]. Escher [32], for example, is a web-based tool that allows exploration of biochemical reactions in the context of pathways. Reactive user-interfaces like these, however, rely on user interaction to manually build and publish visualizations. This can make the replication of visuals and the creation of multiple visuals in a workflow tedious. Director provides an efficient solution to both data visualization challenges in R. The few lines of required script are easy to implement in analysis workflows (see Box 1 and 2).

We plan to expand Director’s abilities beyond custom Sankey diagrams and develop other dynamic visualization techniques for multidimensional biological data. While packages like rCharts [27], networkD3 [28], and Plotly [31] facilitate JavaScript-based visualizations through R, support of the D3 library is limited to basic plots and graphs. Assembling custom visualizations like Director’s Sankey diagram is possible with these packages, but requires additional knowledge of JavaScript, CSS and R to accomplish. Director thus provides a convenient means of making custom dynamic visualizations for biological data in a readily available R package.

6. Conclusion

Director is an exploratory and visualization tool that demonstrates how the visual interpretation of multidimensional data can lead to an intuitive, ‘big picture’ understanding of complex diseases not achieved with statistical methods alone. Despite advances in high-throughput technologies, the exploration of data after analysis for biological insights remains essential manual work. Dynamic visualization can accelerate the discovery process in a way that is both engaging and aesthetically pleasing. For projects containing multiple levels of molecular data, visualization is necessary to meaningfully link and interpret analysis results within a biological context.

The heart of Director, custom Sankey diagrams, utilizes dynamic visualization to discover influential, potentially interesting molecules and biological pathways in regulatory cascades. Implementation in R/Bioconductor enables generation of custom Sankey diagrams to be easily incorporated into analysis workflows. Dynamic elements of the diagram, powered by the D3 library, engage users to explore the data for actionable insights and form hypotheses as to the underlying biological mechanisms represented. Such tools are necessary in modern medical research to reveal useful knowledge from large-scale, heterogeneous molecular and clinical data.

Acknowledgements

This work was supported financially by the Academy of Finland (projects 305087 and 284598), the Sigrid Juselius Foundation and Finnish Cancer Association. The results published here are in whole based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. Computing resources from CSC – IT Center for Science Ltd. are gratefully acknowledged.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2017.10.013.

References


