Marie Vendettuoli[*] Heike Hofmann[†]
Department of Statistics

Human Computer Interaction Program
Bioinformatics and Computational Biology Program
Iowa State University

Abstract

Circle graphs are gaining popularity for the visualisation of gene relationships. This presentation is problemmatic because it requires readers to make comparisons using polar coordinates and does not provide information regarding conditional probabilities without introduction of additional complexity. Additionally the graphic is of low datato- ink ratio, a violation of Tufte's principles. In this paper we assess the information a participant is able to elucidate from traditional circle graphs compared to a new implementation of the hammock plot.

Index Terms: H.5.2 [User Interfaces]: Screen Design—

1 Introduction

Circle graphs have become an increasingly popular visualization tool to depict genomic data and other concept relationships derived from large datasets both in academic and mainstream publications [1], [10]. In its most basic form, each category is arranged on the circles perimeter. Lines drawn through the center connect categories that share a relationship. A circle graph may be extended to include additional information by weighing connectors via color or line thickness to map the strength of a relationship. Other plots summarizing information regarding a specific category may be displayed on a concentric axis just outside the circles perimeter.

A major challenge when visualizing large datasets is the need to balance disparate scales. The existence of relationship trends between categories can only be seen when displaying entire data set while the nature of individual relationships and any supplementary information regarding a single category is better suited for a much smaller scale. These two perspectives provide a strong argument for distinct visualizations when presented as static graphics, while separate images places the burden of cognitive load of comparison and relatedness on the reader.

In the case of many relationships across the circle, overplotting occurs and the order in which the lines are drawn affects the message to reader. Key relationships may be completely obfuscated due to an abundance of non-informative connections. While it is possible to select a color scheme to highlight relationships of interest, this technique may only be applied with a priori knowledge of a connections importance, information which is not available during exploratory analysis.

In contrast hammock plots, designed specifically for instances of categorical data and those of mixed categorical and continuous data [11] allow for linear comparisons. Categories are displayed as a series of univariate labels, grouped by variable. Related categories that are measures of the same variable. Connections are drawn as a bicariage graph of rectangles where the width is proportional to interaction. In the case where rectangles are of width zero (lines), the hammock plot simplifies to the special case of parallel coordinate plots. Of particular interest when applying hamock plots to large datasets is the ability to minimize line crossings - each bivariate graphing region only displays information regarding two variables, which is a maximum of (n_var1)(n_var2) - 1 crossings. This is many orders of magnitudes less than $\Pi_{i=1}^{nvar} n_i - 1$ possible crossings that may occur if all connections are displayed in the same region, as for circle graphs.

2 Methods

Data Preprocessing We first identified three datasets: set A is the fate of passengers on the Titanic, summarized by travelling class, gender, age and survival [3]. Set B is the 8 x 8 table found on the Circos website [9]. Set C is the gene pathway information compiled via KEGG [7] [6] and UCSC genome browser [8] [5]. Genes involved in human metabolism were identified using the KEGG database and the Bioconductor package KEGG.db [2] was used to map

Generating figures

User Study To compare the effectiveness of hammock plots versus circle graphs, we performed a user study of XX participants, X female and Y male ranging in age from a to b. Participants first shown the same short tutorial explaining the features of each plot type (circle graph, hammock plot). The plots were then shown using data set ordering A, B, C, and random assignment of plot type (circular graph, horizontal hammock plot and vertical hammock plot). Each participant viewed a total of three graphs viewing each data set and each plot type exactly once . For data set A, participants were asked: ????? ... No time limit restriction was placed on the tester, but time spent on each question was collected.

3 Results

Data set	Question	Plot type	No. Right	No. Testers
A	?	Circos		
	?	Hammocks, vertical		
	?	Hammocks, horizontal		
B	?	Circos		
	?	Hammocks, vertical		
	?	Hammocks, horizontal		
C	?	Circos		
	?	Hammocks, vertical		
	?	Hammocks, horizontal		

4 Conclusion

Acknowledgements

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

[*] e-mail: mariev@iastate.edu
[†] e-mail: hofmann@iastate.edu

References

[1] Y. Aumann, R. Feldman, Y. B. Yehuda, D. Landau, O. Liphstat, and Y. Schler. Circle graphs: New visualization tools for text-mining. *Principles of Data Mining and Knowledge Discovery*, 1704:277:282, 1999.

[2] M. Carlson, S. Falcon, H. Pages, and N. Li. *KEGG.db: A set of annotation maps for KEGG*, r package version 2.6.1 edition.

[3] R. J. M. Dawson. The 'unusual episode' data revisited. *Journal of Statistics Education*, 3(3), 1995.

[4] S. Ekdahl and E. Sonnhammer. Chromowheel: a new spin on eukaroytic chromosome visualization. *Bioinformatics*, 20(4):576:577, 2004.

[5] P. Fujita, B. Rhead, A. Zweig, A. Hinrichs, D. Karolchik, M. Cline, M. Goldman, G. Barber, H. Clawson, A. Coelho, M. Diekhans, T. Dreszer, B. Giardine, R. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. Kuhn, K. Learned, C. Li, L. Meyer, A. Pohl, B. Raney, K. Rosenboloom, K. Smith, D. Haussler, and W. Kent. The ucsc genome browser database: update 2011. *Nucleic Acids Res.*, October 2010.

[6] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[7] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.*, 40:D109–D114, 2012.

[8] W. Kent, C. S. T. Furey, K. Roskin, T. Pringle, A. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[9] M. Krzywinski. Visualizing tables and tabular data with circos, 2011.

[10] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. Jones, and M. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 2009.

[11] M. Schonlau. Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics*. RAND Corporation, American Statistical Association, 2003.