Eurographics Conference on Visualization (EuroVis) 2017
B. Kozlíková, T. Schreck, and T. Wischgoll
(Guest Editors)

Supplementary Material for Molecular Visualization of Computational Biology Data: A Survey of Surveys

N. Alharbi[1], M. Alharbi[1], X. Martinez[2], M. Krone[3], A. Rose[4], M. Baaden[2], R.S. Laramee[1], M. Chavent[5 6]

[1] Department of Computer Science, Swansea University, UK
[2] Laboratoire de Biochimie Théorique, UPR 9080 CNRS, France
[3] Visualization Research Center, University of Stuttgart, Germany
[4] University of California, San Diego, USA
[5] Department of Biochemistry, University of Oxford, UK
[6] IPBS, Toulouse, France

Methods (extended version) To construct Figures 1 and 2 of the survey of surveys, we extract the references from the Scopus database [Sco] and analyze them using in-house Python scripts. For Figure 1, the references are curated by the experts to define in which category a reference belongs. Briefly, if the reference is published in an ACM, IEEE or related conference and journal it is categorized as a computer visualization paper, otherwise it was tagged as a *"computational biology"* paper. This category is kept very simple due to the paper format. The analysis of word concordance contains two primary phases. In the first phase, the Natural Language Toolkit (NLTK) [Bir06] along with Python is used to process the textual data. The process includes tokenizing the text, lemmatizing the words, removing stopwords and non-alphabetical characters, and finally producing a concordance of important words for each survey. The individual concordances of each individual survey are compared to a collective concordance of all the surveys - the left-most axis in Figure 3. The individual concordances are normalized by survey paper length to make each comparable. In the second phase, the list is passed to a D3.js library that renders a parallel coordinates plot. Each axis represents a survey in the list. The axes are normalized and scaled up to the maximum term frequency value over the axes. This transforms the concordances from raw term frequencies to relative percentages. Each word is mapped to a color that represents the general field the word relates to. The word cloud in Figure 3 is generated using the script by Müller [Mue].

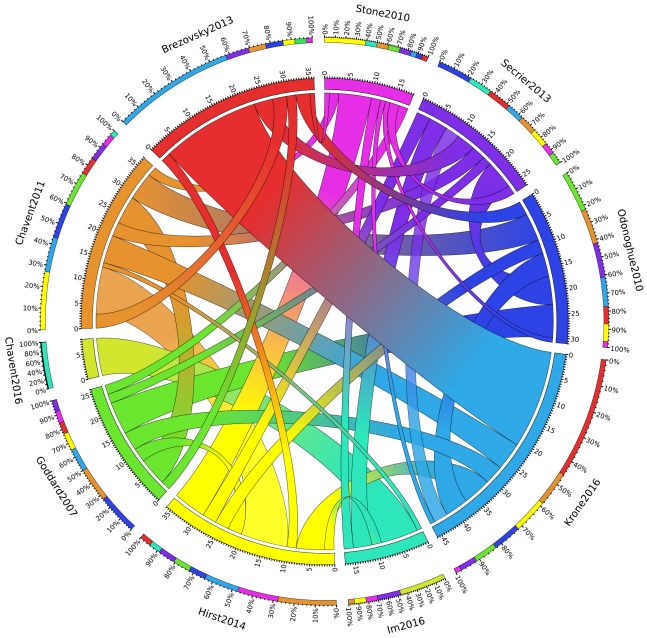

Figure 4: *Radial representation of the connection between CV literature surveys and CB literature surveys based on the number of citations. Each color represents a survey. Thickness of Ribbons encodes the number of common references between two survey papers. The data is visualizaed utilizing Circos [KSB*09].*

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

References

[Bir06] BIRD S.: Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (2006), Association for Computational Linguistics, pp. 69–72. 1

[KSB*09] KRZYWINSKI M., SCHEIN J., BIROL I., CONNORS J., GASCOYNE R., HORSMAN D., JONES S. J., MARRA M. A.: Circos: an information aesthetic for comparative genomics. *Genome research 19*, 9 (2009), 1639–1645. 1

[Mue] MUELLER A.: word_cloud: A little word cloud generator in python. https://github.com/amueller/word_cloud (last accessed: 31.01.17). 1

[Sco] Scopus. (https://www.scopus.com (last accessed: 09.02.17). 1

© 2017 The Author(s) Eurographics Proceedings © 2017 The Eurographics Association.