ECplot: An online tool for making standardized plots from large datasets for bioinformatics publications

Alex Chun-Hong Fok[1,†], Sunny Siu-Nam Mok[1,†], Sau Dan Lee[1] and Kevin Y. Yip[1,*]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

We have implemented ECplot, an online tool for plotting charts from large datasets. The tool supports a variety of chart types commonly-used in bioinformatics publications. In our benchmarking, it was able to create a Box-and-Whisker plot with about 67,000 data points and 8MB total file size within several seconds. The design of the tool makes common formatting operations easy to perform. It also allows more complex operations to be achieved by advanced XML (Extensible Markup Language) and programming options. Data and formatting styles are stored in separate files, such that style templates can be made and applied to new datasets. The text-based file formats based on XML facilitate efficient manipulation of formatting styles for a large number of data series. These file formats also provide a means to reproduce published figures from raw data, which complement parallel efforts in making the data and software involved in published analysis results accessible. We demonstrate this idea by using ECplot to replicate some complex figures from a previous publication.

Availability: ECplot and its source code (under MIT license) are available at https://yiplab.cse.cuhk.edu.hk/ecplot/

1 INTRODUCTION

Bioinformatics research usually involves large-scale datasets. Charts are commonly used to provide a succinct summarization of complex patterns in the data. There are many tools available for chart plotting. Some commercial products include Matlab, Microsoft Excel and Prism. Other tools include Google Chart Tools, Gnuplot and R. These tools vary substantially in terms of the level of programming skills required, cost, and platform specificity. While some of them provide graphical user interfaces, some others require mandatory programming knowledge from users. The costs of these tools range from free-of-charge to hundreds of US dollars per license. Some of the tools can only be run on specific operating systems (OS's), while others have versions for more OS's, or can be run on any platform through a Web interface.

For bioinformatics publications, there are some specific requirements on chart-plotting software:

1. The software should be able to create plots from large amounts of data within a short period of time, ideally without strong requirements on the computing power of users' machines.
2. The software should allow users to easily format a large number of data series with exact style specifications. In particular, having long lists of formatting options distributed across many different sections of the user interface, and asking users to repeat long sequences of mouse clicks at very precise locations for every data series should both be avoided.
3. The software should facilitate seamless re-application of a given formatting style to another dataset, since it is common to have multiple main or supplementary figures of the same type for different sets of data.

We have designed and developed a software tool, ECplot, which fulfills all these requirements. It is a free online tool that can be accessed using any contemporary Web browsers. It targets both users with and without programming experience. The most computationally intensive plotting operations are all performed at the server side, which avoids the need for powerful client machines.

2 BASIC FUNCTIONS

ECplot supports a variety of standard chart types, including bar charts, line charts, pie charts, Box-and-Whisker plots, scatter plots, heat maps and sequence logo plots. The graphical user interface (GUI) provides a spreadsheet for entering data values (Figure 1a). Alternatively, a user may upload a data file in simple delimited formats as input. File formats are clearly specified, while sample data from previous bioinformatics publications are provided as examples. Standard formatting options, such as setting the chart title, position of the legend, font size and series colors, are all provided by the GUI. With a single click of the Apply button, the input data and specified style formats are combined to produce the resulting plot in real time. If further changes to the data and/or formatting styles are made, the figure can be easily regenerated. When a satisfactory version is produced, it can be downloaded in a bitmap (JPG or PNG) or vector (PDF or SVG) format, at a dimension chosen by the user. ECplot provides chart management for registered users. A plot can be saved to its database and later retrieved back from the list of created plots. After any updates to the data and/or formatting styles, the resulting plot can either overwrite the original plot, or be saved as a new plot.

[†] These authors have made equal contribution to the work.

[*] To whom correspondence should be addressed.

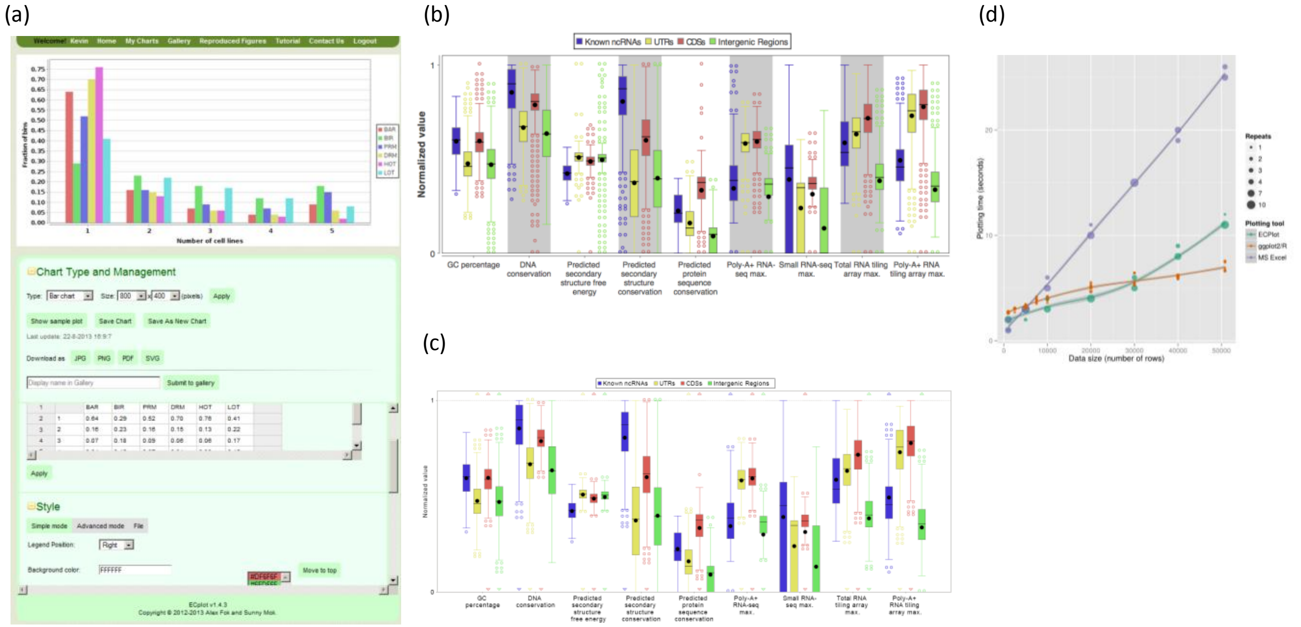

Chart Type and Management

Type: Bar chart Size: 800 x 400 (pixels) Apply

Show sample plot Save Chart Save As New Chart

Last update: 22-8-2013 16:9:7

Download as JPG PNG PDF SVG

Display name in Gallery Submit to gallery

1		BAR	BIR	PRM	DRM	HOT	LOT
2	1	0.64	0.29	0.52	0.70	0.76	0.41
3	2	0.16	0.23	0.16	0.15	0.13	0.22
4	3	0.07	0.18	0.09	0.06	0.06	0.17

Apply

Style

Simple mode Advanced mode File

Legend Position: Right

Background color: FFFFFF #DF6F6F Move to top

ECplot v1.4.3
Copyright © 2012-2013 Alex Fok and Sunny Mok

Fig. 1. (a) The ECplot interface. (b) Figure 1a of Lu and Yip et al., 2011. (c) The replicated figure produced by ECplot. (d) Comparing the time of plotting scatterplots of various sizes by three different tools. ECplot and R were run on a machine with an Intel Xeon CPU E5645 at 2.4GHz (6 cores). Microsoft Excel was run on a machine with an Intel Core i5-3550 CPU (4 cores) at 3.3GHz.

3 ADVANCED FUNCTIONS

To provide more functions for advanced users without making the GUI complex, ECplot offers two ways to perform some operations not supported by the GUI.

The first way is to edit the data/style files directly. All data and formatting styles are stored in XML files with formally defined schemas. An advanced user can edit the files directly. The relevant controls in the GUI will be updated accordingly. This feature is particularly useful when handling a large number of data series. A user can copy the XML section corresponding to the formatting for one data series, and paste it a number of times, with the required formatting changes, for other data series. One may also generate the XML files offline manually or programmatically and upload them to ECplot for plotting.

The second way is to use the programming option of ECplot. ECplot was implemented using the JFreeChart library[1]. An advanced user can issue commands in XML format to call corresponding methods of JFreeChart Java classes with the specified parameter types and values. This programming option also makes ECplot customizable and extensible, in that it can incorporate new functionality provided by new versions of JFreeChart without changing its GUI or source code.

By default ECplot rejects incorrectly formatted XML files. To help users with using the advanced options, syntactic and semantic errors in the XML and command codes can be ignored without affecting the normal operations of ECplot if instructed by the user.

4 FIGURE REPLICATION

Reproducibility of results is a fundamental requirement of scientific research. Many journals now require authors to make the data and detailed methods involved in producing published results available. For example, in a recent publication of the ENCODE consortium (ENCODE Project Consortium, 2012), the exact data files and computer programs used to produce the figures were made publicly available on a virtual machine. ECplot represents another way to facilitate reproduction of published results. The data and style XML files (and the command files, if any) of the figures of a publication can be put at the supplementary web site of it for interested parties to replicate the figures. As a proof-of-concept example, we used ECplot to replicate two complex figures in a previous publication (Lu and Yip *et al*., 2011). Except for parts that were manually added in the original figures, the reproduced figures are identical to the original ones (Figures 1b and 1c; another example available on the ECplot web site), which were originally created using a Java program specifically written for producing the figures. One of the figures involves a data file with about 67,000 data points and a total file size of about 8MB. ECplot was able to create the replicated figure within several seconds, demonstrating its ability in handling large bioinformatics datasets. Figure 1d shows the running time of three different tools when plotting different sub-samples of this data set. It can be seen that at the various data sizes, ECplot had a plotting time comparable to ggplot2.

REFERENCES

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Lu Z.J. and Yip K.Y. et al. (2011) Prediction and characterization of non-coding RNAs in *C. elegans* by integrating conservation, secondary structure and high throughput sequencing and array data. *Genome Res*. 21, 276-285.

[1] http://www.jfree.org/jfreechart/