Data visualization with the programming language R

Paul Brennan (Cardiff University, UK)

Data visualization is an extremely valuable skill in science, finance and journalism. Learning to program will help reproducible data analysis and will increase the different types of visualization that can be generated. The statistical programming language R is a very useful programming language. The R community is friendly, supportive and very diverse including students, academics, health scientists, journalists and professional data scientists. An experience of R or another programming language such as Python or JavaScript will improve your science and your employment opportunities in and outside of research. Programming is a useful skill in education, finance, journalism and other areas too.

Introduction

Data visualization is a vital part of sharing our research. Common data visualizations start with bar charts and line graphs: images for your supervisor and presentations for your department. Then perhaps you need to summarize your data for your thesis and hopefully some papers. As such, making figures is an important part of every scientist's career and good visualizations will improve your data and your presentations will have more impact. Data visualization is a valuable skill outside of research and the knowledge and technical ability to make effective visualizations is a very useful transferable skill. Many colleagues have won their next job, inside and outside academia, because of effective data analysis and visualization skills.

Our task is to create data visualizations that really communicate our science with the minimum of distractions. My aim is always to make figures that can be understood by themselves requiring as little reading of text as possible. There are some great good practice guides for improving your data visualization skills. One of the pioneers in visualization is Edward Tufte who published a beautiful book called *The Visual Display of Quantitative Information* in 1983. He was one of the first to identify key principles for data visualization. These include avoiding distortion, presenting many numbers in a small space and encouraging the eye to compare different pieces of data. He also wrote about data-ink maximization and chart junk. For a more modern analysis, the book *Visualization Analysis and Design* by Tamara Munzner provides a detailed and systematic analysis of data visualization. The book provides lots of good examples. Focussing on science, Nature Blogs have gathered a collection of pieces that were published in *Nature Methods* entitled Points of View. These are described as practical advice on visualizing scientific data. They are bite sized and I found them very readable.

Learning to program to create visualizations

I recommend learning how to program to generate good-quality data visualizations. My favourite tool is the statistical programming language R. Python is also good and I have used JavaScript in the past. These programming languages will allow you to make your work more reproducible. This is great for your future self as you move towards writing and publishing projects that are months and years long. Reproducible data analysis and visualization is good practice for the whole of biochemistry and molecular biology.

By using a programming tool such as R, you will probably have access to a wider variety of visualization methods. I first learned R, in 2014, to allow me to make heat maps and cluster analysis and it immediately helped me analyse a proteomic data set. The book *Visualize This* by Nathan Yau provided me with code that worked immediately and I was hooked. Writing this article, I went back to the original data from 2014 and I was able to reproduce that cluster diagram from my code (Figure 1a). This would be much more difficult using non-programmatic workflows, for example, those based on Excel. Did I mention that R is open source and free? This makes me happy too. Figure 1b and c show other examples of visualizations that R can produce: a violin plot and a phylogenetic tree.

Violin plots are a good way to represent data but are not very commonly used in biochemistry, where the bar plot is more common. There have been calls to move away from bar plots or dynamite plots as they are sometimes called, as they are not great ways to summarize data or reflect variation. Programming

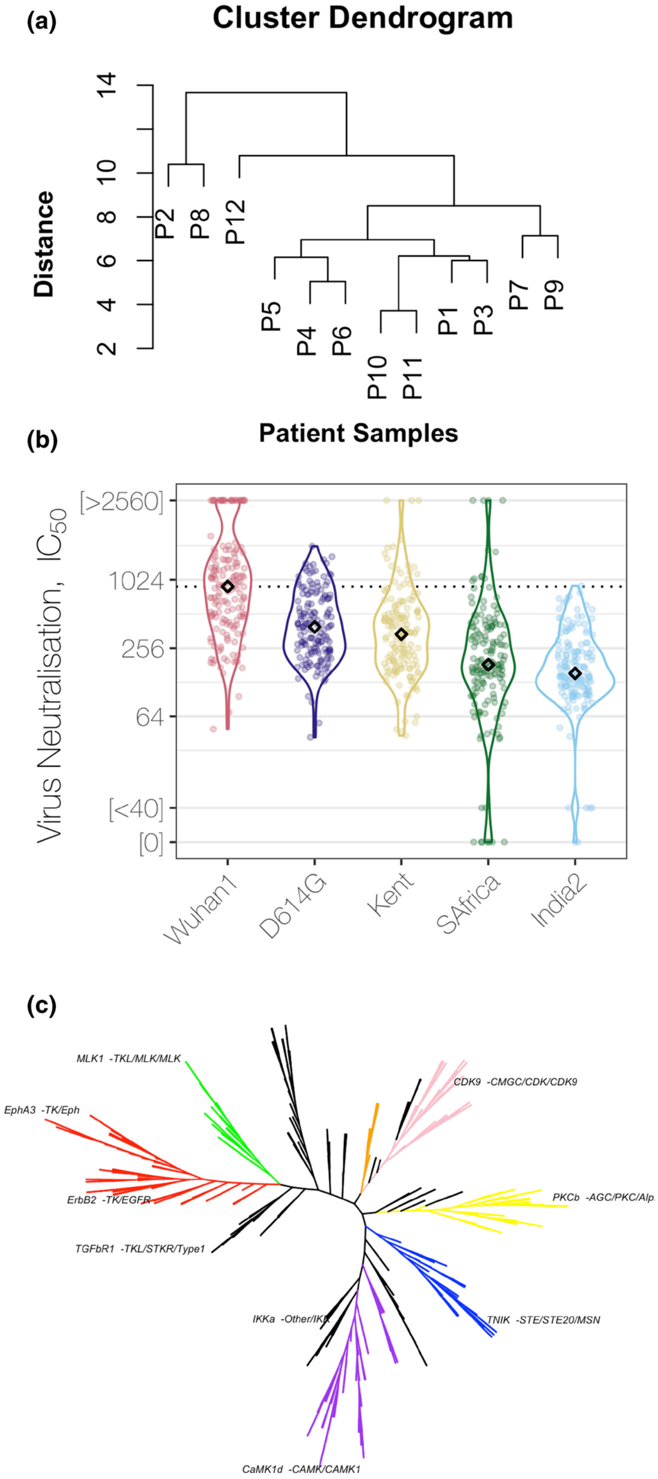

Figure 1. R allows reproducible data visualizations. (**a**) My first published R data visualization – a cluster dendrogram from a proteomic data set. First published in 2014 and reproduced in 2021 (Alsagaby et al, 2014, https://doi.org/10.1021/pr5002803). (**b**) A violin plot showing neutralizing antibody titres against five SARS-CoV-2 strains reproduced from shared data and code from Wall et al (2021, https://doi.org/10.1016/S0140-6736(21)01290-3). (**c**) Phylogenetic tree of human kinase domains inspired by visualization by Manning et al (2002, https://doi.org/10.1126/science.1075762). All made with R: code and data required to make these visualizations are available from https://rforbiochemists.blogspot.com/ or https://github.com/brennanpincardiff/

allows you to easily create lots of different types of plot as you explore your data, again in a reproducible way. The violin plot in Figure 1Bb is from a paper published in 2021 in the *Lancet* by Emma Wall and others from the David Bauer laboratory working in the Francis Crick Institute in London. They shared their data and R script which generated their graphs and statistical analysis through Github and so I was able to reproduce their graph in just 20 minutes. This is the benefit of learning and using programming to analyse your data – easy reproducibility and sharing – good practice for your science.

First steps in learning R

I recommend the R-Studio integrated development environment (IDE) for R. It made R much easier to use and really helped my learning. You can write scripts which record your workflow. These can be saved, shared and opened later. You can keep a record of everything you do. For bioinformatics, there is a large collection of R packages – packages are collections of R functions that facilitate data analysis. One important collection is held at *Bioconductor*, which collects and develops R tools for the analysis and comprehension of high-throughput genomic data and biologic data.

Learning R does take some time. Regular practice makes the difference – every day at the beginning if possible. There is a free online book entitled *R for Data Science* (https://r4ds.had.co.nz/) that can help. This book promotes an opinionated set of packages (groups of functions) that works together in R called the TidyVerse. They represent a good framework for data analysis and have a lot of free resources available online. A typical data analysis workflow involves:

1. Importing the data
2. Tidying our data (also known as wrangling or munging)
3. Transforming or summarizing
4. Visualizing

If you want to try to make your first data visualization in R, see Box 1 for how to start. There are plenty of sample scripts on the R for Biochemists blog site for you to explore. You could consider signing up for an online course with the Biochemical Society.

Perhaps you have a data set of your own that you want to play with and a visualization that you want to try to make. This will help to inspire you to keep learning. Keep trying and perhaps share your results. I'm not saying that it is easy. Sometimes learning programming is quite frustrating. However, solving the challenges that arise can be very rewarding. My recommendation is regular practice for a couple of months and you will see a difference in your work.

Box 1. Steps to making your first graph in R

- Download R
- Download R-Studio
- Open R-Studio
- File > New File > R Script
- Go to R for Biochemists blog
- Select a Starting point for yourself
- On that page, cut and paste the script into your new file
- Run the script line by line and see your first R data viz appear
- Change the code and repeat
- Learn and try to have fun

Teaching R to biochemists and molecular biologists

As part of the Training Theme panel for the Biochemical Society, we created an online course focussed around teaching R to biochemists and molecular biologists. It is called R for Biochemists 101. We created this 5 years ago and over that time, over 500 learners have engaged with the course. For a first person experience of this see Box 2 written by Ellie Davis. In R for Biochemists, I have focussed on real data that has been generated by researchers in the biomedical or life sciences. I usually find this more interesting than data about airlines or Star Wars. I think it makes the analysis and visualizations more relevant to biochemists. During the course learners draw a protein standard curve, an enzyme kinetics plot and a volcano plot. We calculate some kinetics constants. We also extract data from images and draw some maps.

In parallel, I also created a blog site called R for Biochemists which has been going since 2015. There are many practice example R scripts on the site and you are welcome to explore this at your leisure. Perhaps the most interesting place to start is the visual index. Since it began, the blog has had over 150,000 visits. The most popular page is a script that shows how to make a volcano plot (Figure 2a). It has had >10,000 views from 16 different countries. It has been interesting to share teaching materials in this way. Figure 2 shows three of the most popular visualizations that have been viewed – the volcano plot, an LD50 plot and a graph from some flow cytometry data.

Last year, we developed another R course entitled R for Biochemists 201. The first run of this course will be in November 2021 and future runs are planned. The material builds on R for Biochemists 101. R for Biochemists 201 will teach participants the key concepts of tidy data, about good coding practice

Box 2. Ellie Davis' R journey

- My journey with R began with the R for Biochemists 101 online course and has taken several guises; I have completed the course as a learner, managed the course as an administrator and, most recently, contributed to a content update. Not so very long before my own adventures with R, I believed any sort of programming to be a completely unattainable skillset for me, even if its value has always been very apparent. I don't think I'm alone in this; words like "script" and "code" are daunting for many, and their associated skills reserved for an exclusive group. I have since discovered that this isn't the case. In fact, the R community is welcoming and supportive with a culture founded on open source and collaboration. While working with Paul on the maintenance of R101 and the development of R201, I have taken great pleasure in seeing this culture continued in the comment sections; learners are encouraged to use the wealth of resources available to support their training, and it is wonderful to see them share what they have found with their fellow learners. These same learners are often complete beginners to programming, and yet they are quickly able to harness its power to visualize and analyse their data. I really enjoyed reading learner comments, especially their plans to apply newly learned code to a complicated data set. Sometimes, they have been sitting on the data for a while due to feeling unsure of how to handle it and now they feel empowered. That's why it's important to break down the walls to computational skills and continue to equip our community with the tools they need to maximize the potential of their research.
- *Ellie Davis worked in the Biochemical Society Conference Office from 2018–2021, managing the Society's Training from 2020 and has now moved to Historic England as a Training Adviser.*

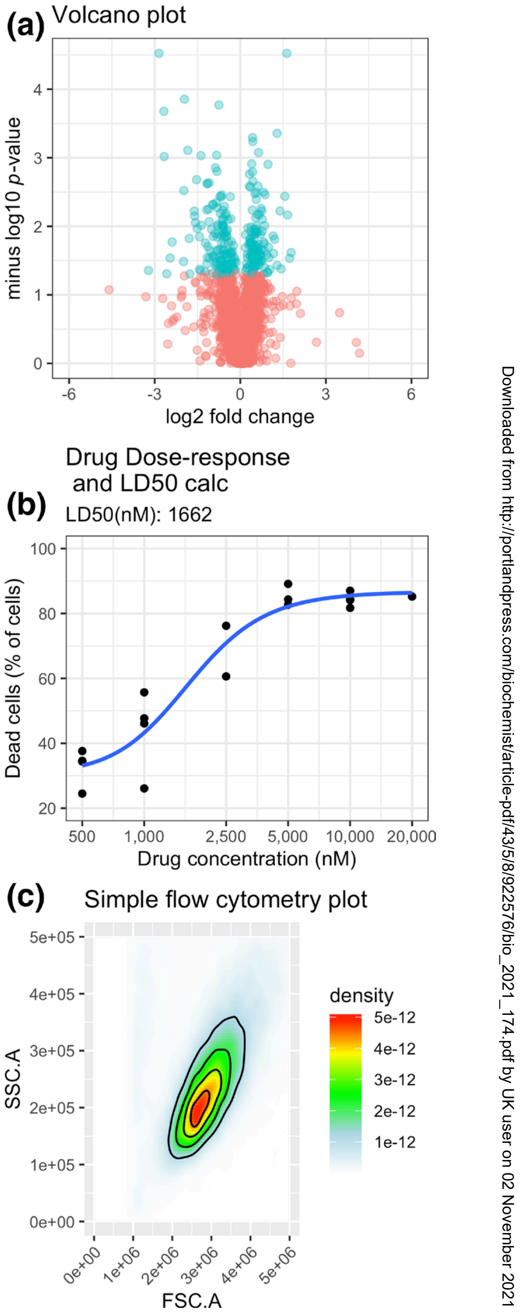

Figure 2. Three of the most viewed data visualizations from R for Biochemists blog**.** (**a**) A volcano plot. (b) A drug dose–response curve and LD50 calculation. (**c**) A simple flow cytometry plot. All made with R: code and data required to make these visualizations are available from https://rforbiochemists.blogspot.com/ or https://github.com/brennanpincardiff/

such as developing reproducible workflows and how to create more complex data visualizations in R. Please get in contact with the Biochemical Society if you would like to join the next run of R101 or R201. The Biochemical Society also runs online training in Python as well and provides training in a wide variety of biochemical areas. If you would like to get involved in designing and delivering training please make contact.

Lots of journals now require authors to share their data. One of the exercises I recommend is to try to reproduce a figure the authors have made from their data. This is a great way to play with data in the public domain. Having a target figure to try to reproduce is a good test of your understanding of the data and the message of the paper. This approach is often quite a challenge because very few authors will supply the code they used to make the data. Sometimes authors have

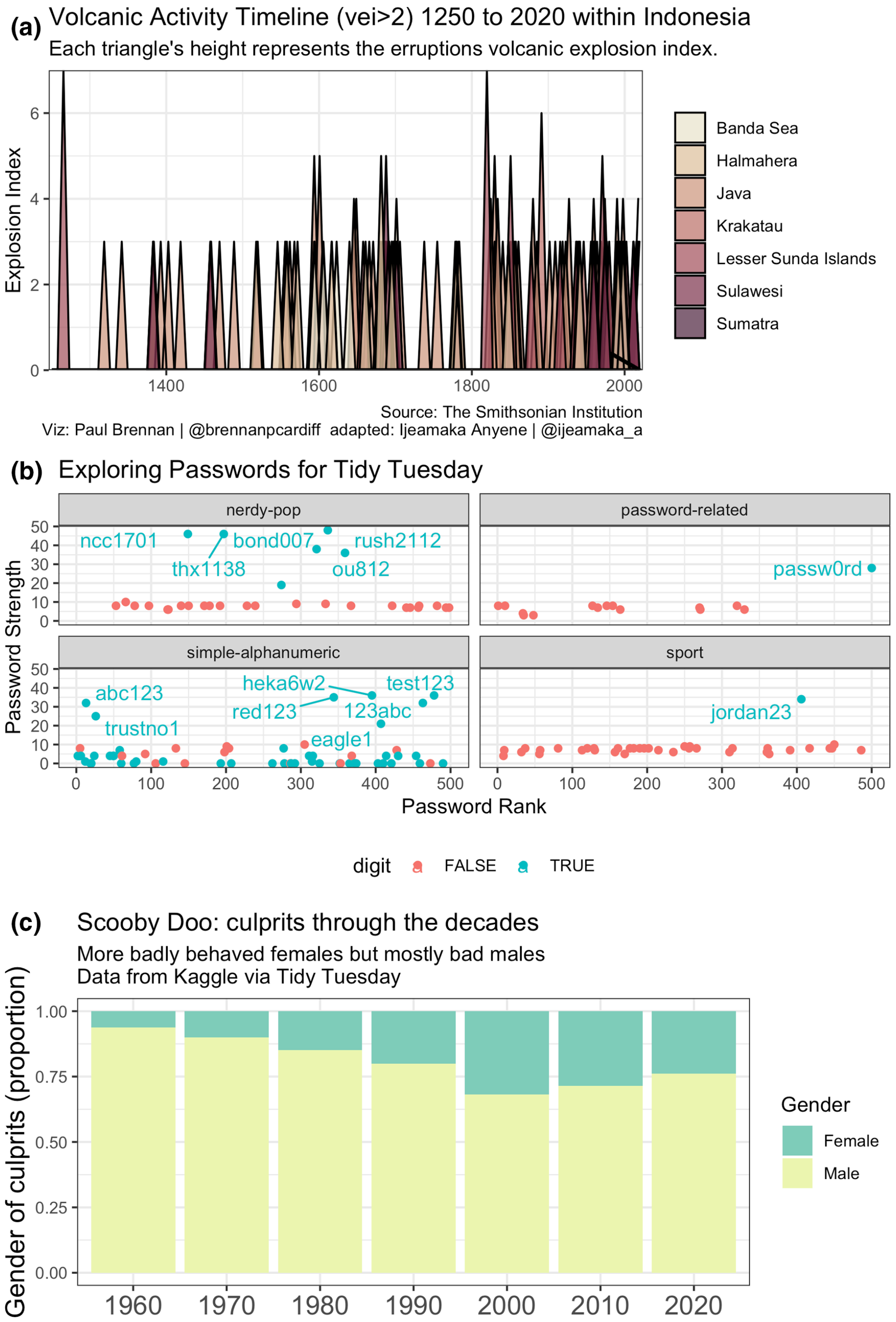

Figure 3. Three data visualizations made while participating in Tidy Tuesday – an R data visualization showcase. (**a**) A volcanic activity time line inspired by @ijeamaka_a. (**b**) Illustrating the importance of numbers in password strength. Across a range of password types, inclusion of numbers increases password strength. (**c**) Showing the proportion of female culprits in Scooby Doo shows from 1960s to 2020s . Tidy Tuesday site: https://githubcom/rfordatascience/tidytuesday. All made with R: code and data required to make these visualizations are available from https://rforbiochemists.blogspot.com/ or https://github.com/brennanpincardiff/

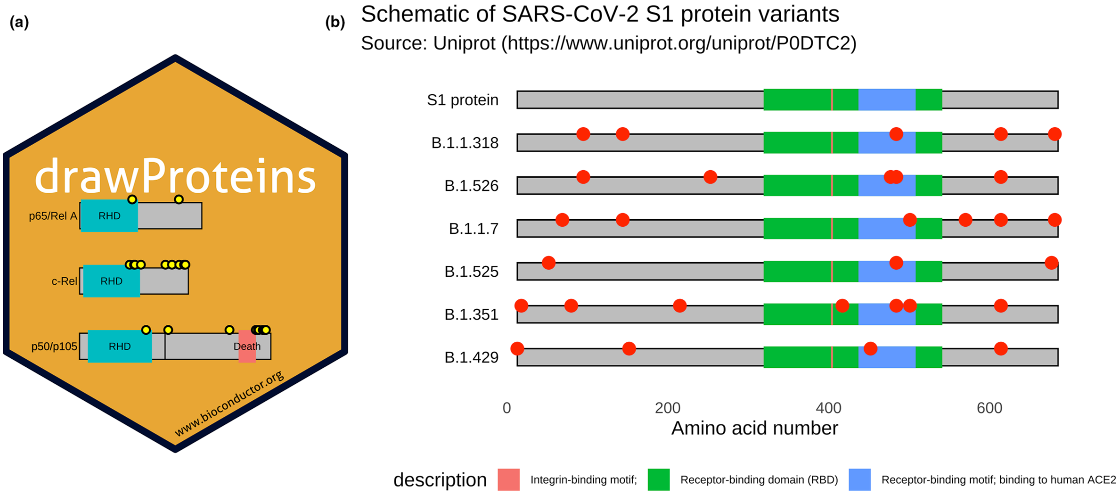

Figure 4. Showcasing drawProteins. (**a**) The Hexsticker for the Bioconductor package drawProteins. (**b**) Schematic of SARS-CoV S1 protein variants. Protein and variant information from Uniprot (https://www.uniprot.org/uniprot/P0DTC2). All made with R: code and data required to make these visualizations are available from https://rforbiochemists.blogspot.com/ or https://github.com/brennanpincardiff/

made some changes to the data that you don't appreciate the first time you look at things. I've written a few blog posts on R for Biochemists about how I have done this.

Engaging with the wider R community

The R community extends well beyond science and I am inspired by it regularly. By attending a local R User group, I met colleagues from local companies, the NHS and other universities. I learned some really important computer programming lessons including the value of version control, how to create my own package and more about machine learning. We learned together, gave talks, ran SatRday One Day Conferences and online workshops. I have seen colleagues in the User group use their knowledge of R to move from academia to industry, specifically into data science consultancy and finance.

I have found the R community friendly and inclusive. As a result of COVID-19, there are many online workshops. It is possible to engage with learners and conferences from all over the world. There are over 200 active R-Ladies groups (@RLadiesGlobal) in many different countries that you can join for free. The R-Forwards group (@R_Forwards) aims to create teaching resources that reflect the diverse community of R users. The broader data community can be excellent too. As examples, you can find @BlackInData, @R_LGBTQ and @QueerinAI on Twitter. I follow all of these. I recommend reaching out to those who interest and inspire you to find your own supportive group. This may be in person or online as suits you.

Tidy Tuesday is a data visualization showcase that runs every Tuesday. They share data on Github. If you can search Twitter or other social media platforms for #TidyTuesday, you will see inspiring data visualizations. I have been inspired by @ijeamaka_a, @dokatox, @juliasilge among others. Many of the people posting visualization include their code so you can try to reproduce their work. Sometimes it requires an investment of time. When I participate, I always learn something. Here are three data visualizations I made while participating in Tidy Tuesday (Figure 3). The topics vary: volcanic activity, password strength and the gender of Scooby Doo culprits!

Making protein schematics with R – drawProteins

Over my research and teaching career, I've made many schematics of proteins. These are not always regarded as data visualization but they are. My tool of choice was PowerPoint. In 2018, I decided to write R code to allow me to reproducibly and programmatically generate protein schematics. Moving from data analysis to creating my own functions and publishing these as a package was my next step in becoming a computer programmer. I had to learn about code development, unit testing and continuous integration. I engaged with the Bioconductor community and published my package called drawProteins. Figure 4 shows a little of what is possible with the package. I've used drawProteins to visualize Spike protein S1 proteins modifications from various SARS-CoV-2 corona virus strains using data downloaded from UniProt, a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. drawProteins can also be used to visualize multiple proteins and post-translational modifications such as protein phosphorylation sites. These protein visualizations can be generated programmatically and in a reproducible way.

Final words

I hope that I have persuaded you that creating data visualizations is vital for scientists and is a transferable skill. If you can do it with a programming language it will help you in the long term: the months, years and decades ahead. We generate lots of data of increasingly sophisticated type both in biochemistry, molecular biology and business. If you can learn to summarize data into inspiring visualizations, this skill will be useful for your future career no matter where it leads. Please be aware that all the data visualizations shown in the figures were made with R. You can reproduce these figures using the code and data which are available at the R for Biochemist blog and Github.

Anyone interested in learning more about the Society's programming courses and training work should visit www.biochemistry.org/events-and-training/training or contact the Training Team at conferences@biochemistry.org. ■

Further Reading

Resources for starting to learn R

- This website and the published book is a great starting point for learning R: https://r4ds.had.co.nz/ or Wickham, H. and Grolemund, G. (2017) *R for Data Science*. O'Reilly Media, Sebastopol
- There are lots of biochemistry-inspired R scripts to try on the R for Biochemists blog site: https://rforbiochemists.blogspot.com/
- More general R information is available through R-Studio: R-Studio website: https://www.rstudio.com/ and https://www.rstudio.com/resources/webinars
- Biochemical Society Training Courses (https://biochemistry.org/events-and-training/training/): R for Biochemists 101 and, coming in 2021, R for Biochemists 201.

Learning more about data visualization

- This collection is very interesting and well worth a read: Evanko, S. (2013) Data visualization: A view of every Points of View column. *Methagora* http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html
- This book inspired me to learn R: Yau, N. (2011) *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, Wiley
- Inspirational talks and visualizations: VizBi website https://vizbi.org/ Lots of videos, talks and poster about visualizing biological data.
- A detailed theoretical knowledge: Munzner, T. (2015) *Visualization Analysis & Design*, CRC Press, Boca Raton
- Edward Tufte information and books https://www.edwardtufte.com/tufte/books_vdqi

R packages and inspiration

- TidyVerse website https://www.tidyverse.org/
- Bioconductor home page https://www.bioconductor.org/
- Tidy Tuesday site https://github.com/rfordatascience/tidytuesday
- Tidy Tuesday Twitter links: https://twitter.com/search?q=%23tidytuesday&lang=en
- drawProteins on Bioconductor https://www.bioconductor.org/packages/release/bioc/html/drawProteins.html
- R Open Sci – open tools for open science https://ropensci.org/

Dr Paul Brennan is an educator and biochemist with over 25 years experience of teaching and research. He works in the Centre for Medical Education at Cardiff University. As well as programming, he also facilitates biomedical teaching and has a leadership role in equality, diversity and inclusion. Email: BrennanP@cardiff.ac.uk. University home page: https://www.cardiff.ac.uk/people/view/122818-brennan-paul. Twitter: brennanpcardiff. Github: bennanpincardiff