BINF sights.

ISSUE 28
2025-T2-WK7

MACHINE LEARNING IN BIOINFORMATICS

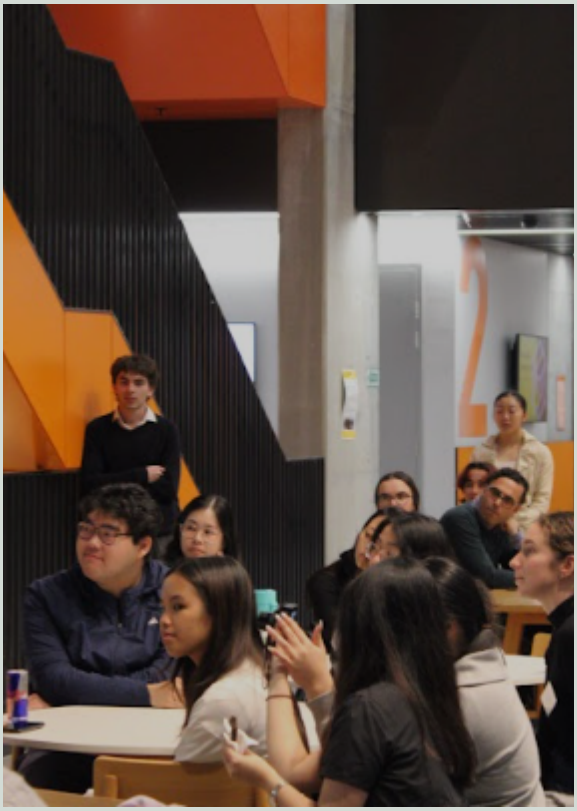
Where Biology Meets Intelligence:
Harnessing Machine Learning to Decode Biology

IMAGE SOURCE:
Freepik

NETWORKING NIGHT

On Thursday, June 19th, BINFSOC hosted our Annual Networking Night. There were guests and students from many different backgrounds, who were all given the opportunity to connect and engage with each other. The evening was separated into two valuable parts: a set of 8 round-table discussions, where students could meet all of our distinguished guests, and some open-floor networking, for continued discussions and personalised questions.

We are proud to have hosted such a successful evening, but it could not have been done without the organising team, attendees, and guests. We thank the organising team for putting together such a wonderful event, the attendees for being interested and engaged, and, of course, our guests, named in the list below, - for making the evening special. You helped the future of bioinformatics navigate the future, and provided valuable insights into the industry.



NETWORKING NIGHT

Finally, we would like to thank our sponsors for helping us to elevate this event and make it as incredible as it was. We will continue to provide opportunities for students to interact with bioinformatics - so stay connected via our socials!

Australian Proteome Analysis Faculty	Dr Ignatius Pang
CSIRO	Carol Lee Letitia Sng
CCI	Piyush Mundra Mark Cowley Gabi Ryan
Garvan	Hani Kim John Reeves
Commonwealth Bank	Aravind Venkateswaran Gina Brichacek
UNSW	Nona Farbehi Raymond Louie Hamid Alinejad-Rokny Sara Ballouz
Fold AI	Alex Gavryushkin
UTS	Hao Chen

INDUSTRY MENTORING PROGRAM

BINFSOC has officially kicked off our Industry Mentoring program!

The program aims to connect students with industry experts, academic mentors and with their fellow peers. The first event in the program was a two-part resume workshop. Students were able to learn the components of a good resume before diving into a resume review and mock interview workshop. Students were able to use their new resume skills and develop their interview skills with real industry professionals.

With more to come, BINFSOC is excited to be able to play a role in connecting students with each other and the industry!

MATCHA CRAWL

On Sunday, June 22nd, the members of BINFSOC embarked on the society's first-ever matcha crawl! Students explored all over Sydney's CBD to try different renditions of the trendy, earthy drink, and even dressed in green to match the occasion. It was a great experience for students to socialise and wind down from studies during a stressful term. Thank you to the following cafes for hosting our matcha crawl - be sure to check these places out if you haven't already (they're definitely BINFSoc-approved)!

MATCHAYA

Envoy Cafe

Chubby Cubby Cafe

Children's Cancer Institute

INDUSTRY GUEST FEATURE

GABRIELLE RYAN

RESEARCH ASSISTANT AT CHILDRENS CANCER INSTITUTE

GENERAL ADVICE AND INDUSTRY INSIGHTS

Gabrielle Ryan
Research Assistant at CCI

Can you talk about your journey as an undergraduate student to where you are now?

I started my journey in the area of computer science and mathematics during my Bachelor of Science. Although I enjoyed the use of computational tools for problem solving, I also found the application to be too "dry". It wasn't until my third year when I took a bioinformatics course and discovered the potential of the field.

Thus, I decided to continue my university career with a Master of Bioinformatics to gain further understanding and experience. During my degree, I completed two research projects where I became hands-on in some real-world data from some algae that grows on coral, as well as some wet-lab experience with yeast!

With these projects under my belt, I applied to some of the few bioinformatics positions available in Australia at the time, and managed to snag a job at the Australian National University in Canberra as a Bioinformatician for a small consultancy group. In this role, I worked on some front-end development of a user-interface for some mice and human data, helped to develop a lab-interface system using Python, and did some RNA-seq analysis for malaria parasites. From this role, I had gained experience in a variety of very different areas within bioinformatics, and was ready to delve deeper into a specific area with real world impact, which led me to my current position as a Research Assistant at the Children's Cancer Insititute.

Can you tell us about your current role and what a typical day looks like for you?:

Currently, I'm working as a research assistant within the Germline Childhood Cancer Risk (GCCR) team at CCI. My role involves working on some long-term research projects, while also assisting with analysis from other team members in my team, or collaborators. A typical day for me will involve working on the tasks that are due soon and that other people rely on me for - generally for presentations or grants. This includes tasks such as using R to run gene set enrichment analysis on methylation data, or do some statistics analyses and produce figures. In between this I will attend meetings or seminars at CCI, and finally get some work in for my long-term projects, which currently involves setting up a workflow on Databricks for accessing variants from our patients.

What kind of projects are you working on, and what problems are you trying to solve through your current work or research?

Within my GCCR team, our overall goal is to discover new genes or variants that are potential drivers of cancer using the data collected from a program called ZERO. Childhood cancer is unique from adult cancer in that children have less exposure to environmental factors that might cause cancer (e.g. smoking, sun exposure). Thus we expect a major cause of cancer would be genetic risk inherited from their parents. However, so far we can only account for ~16% of these potential genetic causes. Our team is aimed at trying to find uncover potential genetic reasons for the rest of these children. My current project is using statistics and the latest platforms to uncover genes with a significant amount of mutational burden in our ZERO cohort, in comparison to the cohort in gnomAD (considered to be generally healthy).

What technologies, tools, or methods are central to your work

Currently, I am involved in the use of a platform called Databricks to analyse all of the SNVs and InDels from ZERO. This also involves the use of PySpark as a language. I also consistently use R to pre-process and analyse data, along with bash scripting. Some tools I use for variant calling include samtools, bcftools, bedtools for processing data, Docker for accessing bioinformatics tools, Git and Github for version control and sharing, and VEP, ClinVar, gnomAD, and CADD for annotation and interpretation of variants.

Are there any underrated tools, platforms, or resources you think more people in bioinformatics should know about?

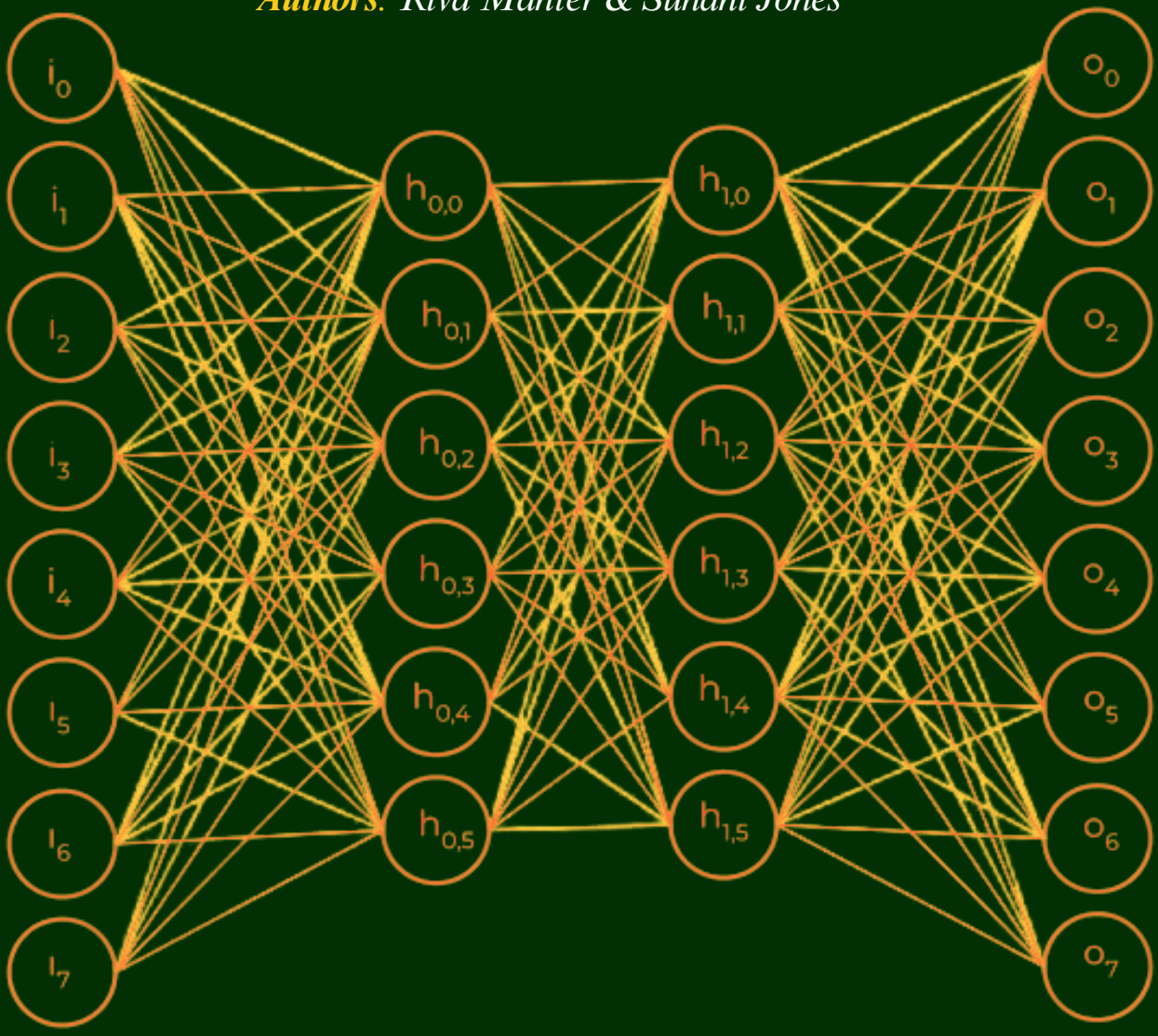
As a resource, I love the Youtube channel StatsQuest! Josh explains statistics by breaking everything down and going through the principles, which has been essential to my learning.

What advice would you give to students just starting out in bioinformatics? Are there skills they should focus on early?

I believe an understanding of molecular biology and statistics is essential and something I wish I had a deeper understanding of. While using computational approaches is very useful as a tool, if we don't understand the underlying biology, then it's possible to overlook important biological patterns. Additionally, try to get involved in as many projects as you can, you're still learning what area you enjoy, and it's great to get hands on experience with real-world data.

Machine Learning

Authors: *Riva Manter & Suhani Jones*



An Introduction to Machine Learning

Where it Fits in to AI

We are currently in a huge technological transition, quoted as the "Fourth Industrial Revolution", with the booming wealth of data. Part of this transition has been the emergence of different systems, such as machine learning, to manage and analyse these pools of data. You may have heard of machine learning being used interchangeably with Artificial Intelligence (AI), however they do differ. AI broadly refers to the "ability of computers to emulate human thought and perform tasks in real-world environments", however machine learning is defined as the "technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience and data". To put it simply, AI is a broad field that requires machine learning which processes data to make algorithmic decisions, through pattern recognition.

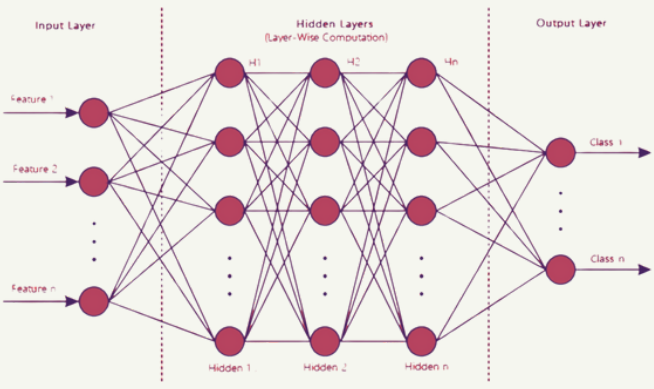
Some Types of Machine Learning

Supervised learning: a task-driven approach, taking sample input-output pairs to apply to new data i.e. "learning" a function based on labeled training data training examples. An example would be predicting the classification or genre of a text based on other texts with identified genres.

Unsupervised learning: a data-driven process, analysing unlabelled datasets without the need for human interference. This is generally used for understanding trends in data such as, extracting generative features, identifying meaningful structures, groupings in results, and exploratory purposes.

Semi-supervised learning: a hybrid approach of supervised and unsupervised learning models. It operates on both labeled and unlabelled data. This approach is useful to improve upon unsupervised learning models, providing a better prediction comparatively.

Reinforcement learning: an environment-driven approach, using evaluative methods to design the optimal behaviour. This idea is based on rewards/penalties to minimise risks or increase rewards of tasks. Areas where this approach is most useful are automation tasks such as robotics, autonomous driving tasks, and manufacturing.



A structure of an artificial neural network modelling with multiple processing layers
Image: (Sarker, 2021)

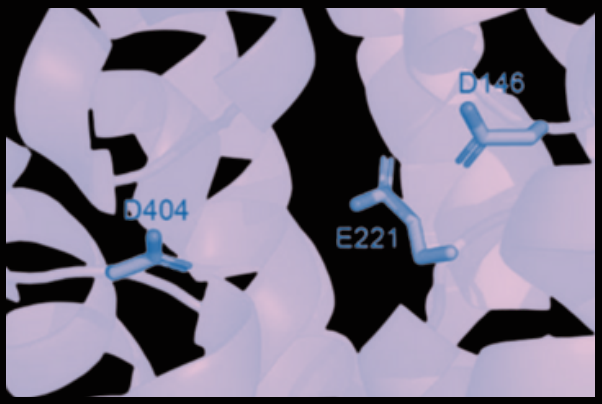
Each learning model accounts utilises different algorithms themselves in order to carry out tasks, such as classification analysis, regression analysis, data clustering and other deep-learning models, which themselves have their own different theoretical approaches useful for a range of required tasks.

Machine Learning and Bioinformatics

The reach that machine learning has in real life applications is wide, as well in bioinformatics, where it is used ubiquitously to identify patterns from given data and has revolutionised ways of research.

Some supervised learning applications include algorithms like support vector machines (SVM), K-nearest neighbours (KNN), regressions and random forests, which in turn can lead to representations of gene expression and single nucleotide polymorphisms (SNPs). These algorithms may be used to classify, analyse and draw trends between cancer vs healthy patterns, and many other domains of study involved in biomedicine.

Types of supervised learning algorithms used for biomedical research
Image: (Auslander et al., 2021)



Predicting charged structures within proteins
Image: ResearchGate)

Importantly, structure and hence function can be determined through softwares such as AlphaFold2, a deep neural network that shows the structure of every single protein, allowing researchers to link the importance of protein structure to function. This piece of machine learning software was revolutionary to the science field, where the function of many proteins once unknown were able to be determined.

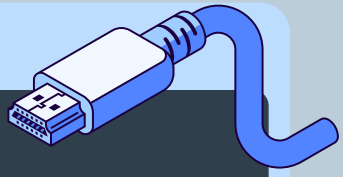
Conclusion

It is clear that the applications of machine learning in the bioinformatics field has no limit. From phylogenetic trees to pinpointing which mutations may cause certain diseases, machine learning is vital for making inferences on data.

References and further reading:

- *https:///link.springer.com/article/10.1007/s42979-021-00592-x*
- *https:///link.springer.com/article/10.1007/s42979-021-00592-x*
- *https://ianjwehba.medium.com/unfolding-alphafold-and-what-it-means-for-biology-46a763d30a10*

Code Your Way



```
$ chmod +x print_info.py
$ ./print_info.py

Welcome back to...
                    << CODE YOUR WAY! >>
Today, we have an introduction to R coding! R is an open
source programming language, with many libraries and
packages used for specific statistical and data
visualisation purposes. Many statisticians choose to use
RStudio, a graphical user interface for R (think of it
like coding in any language on VS Code), so Code Your
Way is exploring the basics here.

R is the most widely used bioinformatics tool, working
on different platforms and over large datasets.
```

ATTGCCTTATR
ACTGATAGCCSTUDIO_

Is there a coding concept you'd love to see featured in **Code Your Way**? Let us know by emailing **binfsights@binfsoc.com** — we'd love to hear from you!

basics of R studio

Let's start off with some basic R commands!

Mathematic Operations:

Think of R like a powerful calculator. Here are some mathematic operations used:

- Basic Arithmetic:
  `+, -, *, /`
- Special Operators:
  `%% (modulus), %*% (matrix multiplication)`
- Special variables:
  `pi, TRUE, FALSE, NULL, NA, NaN, Inf`
- Fancier operators:
  `log(), exp(), sqrt(), round()`
- Logical operators:
  `|, &, <, >`

Data Types/Structures

Some basic data types & structures, and what they look like:

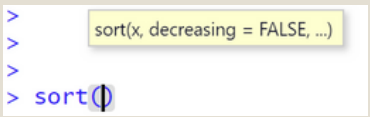
- Numbers:
  `1 [integer], 3.1415 [double]`
- Characters:
  `'A', 'B', 'C'`
- Strings:
  `"BINFSoc", "Rules"`
- Array or vector (multiple elements of the same data type):
  `c("Hello", "Goodbye"), seq(0,200,10), rnorm(100), rep(1,10)`

Functions

Like many other programming languages, function in R take in defined inputs to return outputs. Functions can be user defined or from packages/libraries. To find out necessary inputs type:

`?function`

- RStudio also suggests inputs above a function once typed in:



Visualisation

An important part of statistics is data visualisation. Here's visualisation of the famous Iris dataset in R:

Want to know a bit about the data?

Type these in:

```
summary(iris)
```

What does this code do?
Statistical summaries for each set of data

```
class(iris)
```

What does this code do?
Determine the data type/structure

```
colnames(iris)
```

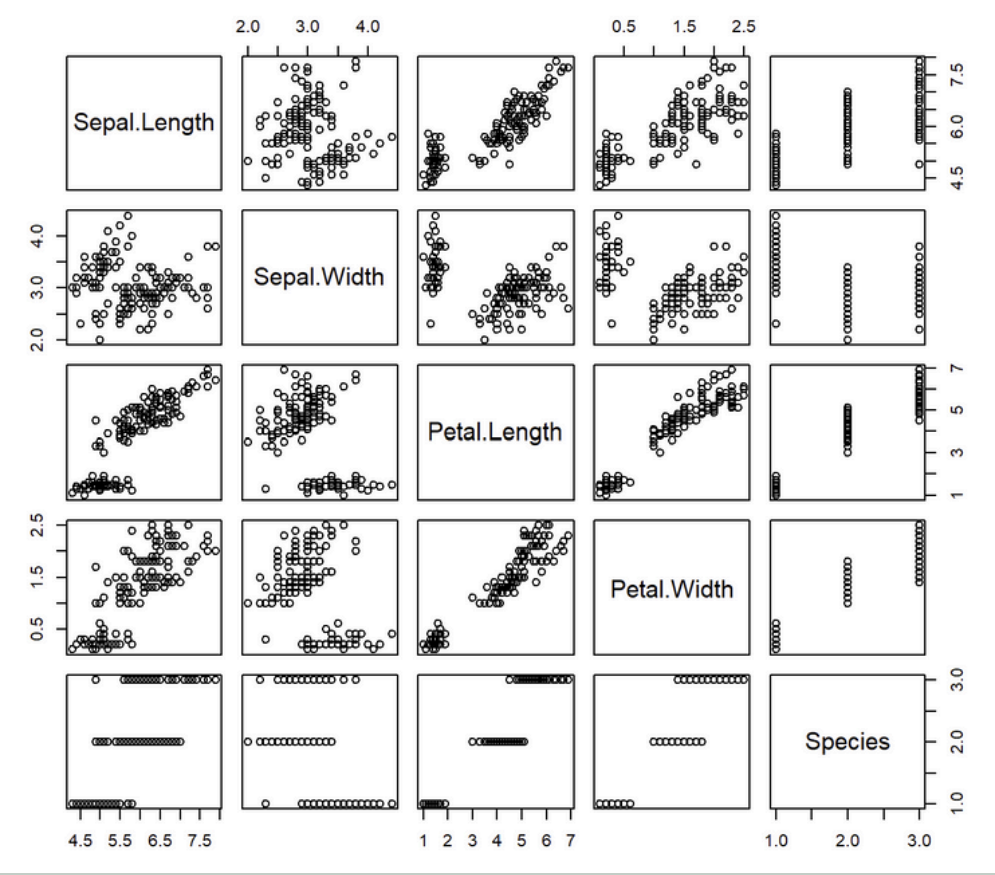
What does this code do?
Get the names for each column

Plotting

We can choose to plot the data in multiply ways based on what is required:

1) Let's try plotting everything:

```
plot(iris)
```



This may not be the best plot for our data, as the it is hard to extract meaning from the visualisation.

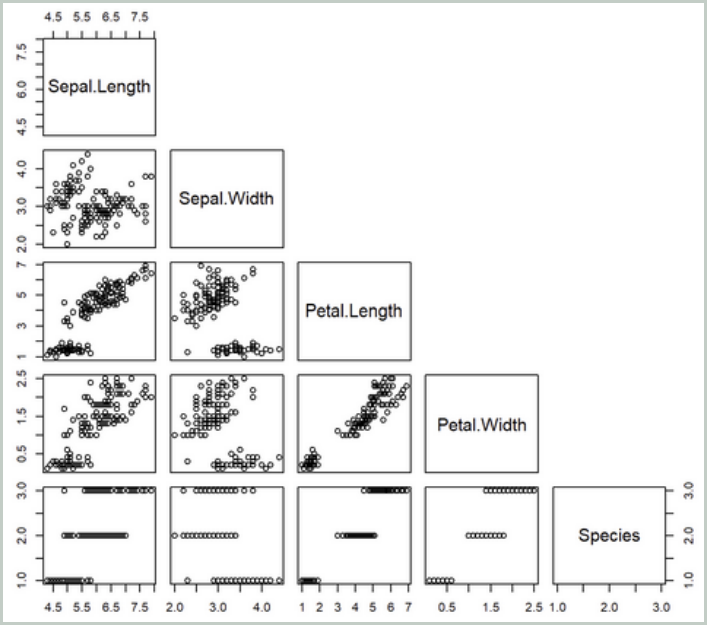
Instead, let's try to see the data plotted once: i.e. the lower half of this full plot.

Plotting Cont.

2) Plotting only the lower half of the scatter plot:

```
pairs(iris, upper.panel = NULL)
```

What does this code do?
Plots a matrix of scatterplots, where we have specified no upper panel



Alternatively, to see the upper half of the scatter plot:

```
pairs(iris, lower.panel = NULL)
```

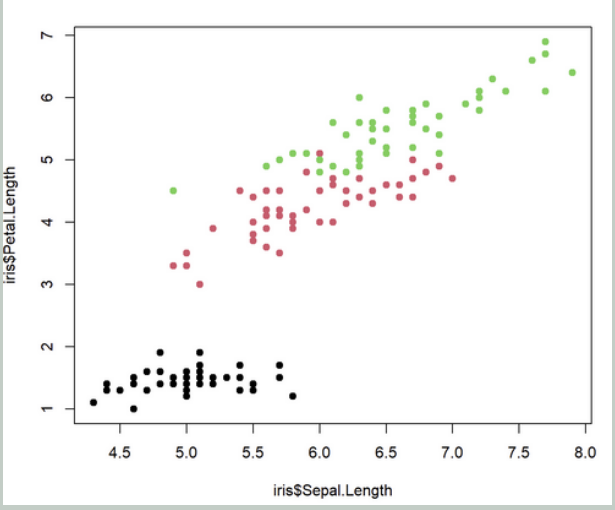
Now that we can understand the data better, let's try to visualise individual parts separately

3) Plotting Sepal Length vs Petal Length:

```
plot(iris$Sepal.Length, iris$Petal.Length, pch=19, col=as.numeric(iris$Species)
```

What does this code do?
plot function with:

- x-axis: Sepal Length
- y-axis: Petal Length
- pch=19: setting the plots to be solid circles
- col=as.numeric(iris$Species): setting each iris species as a different colour

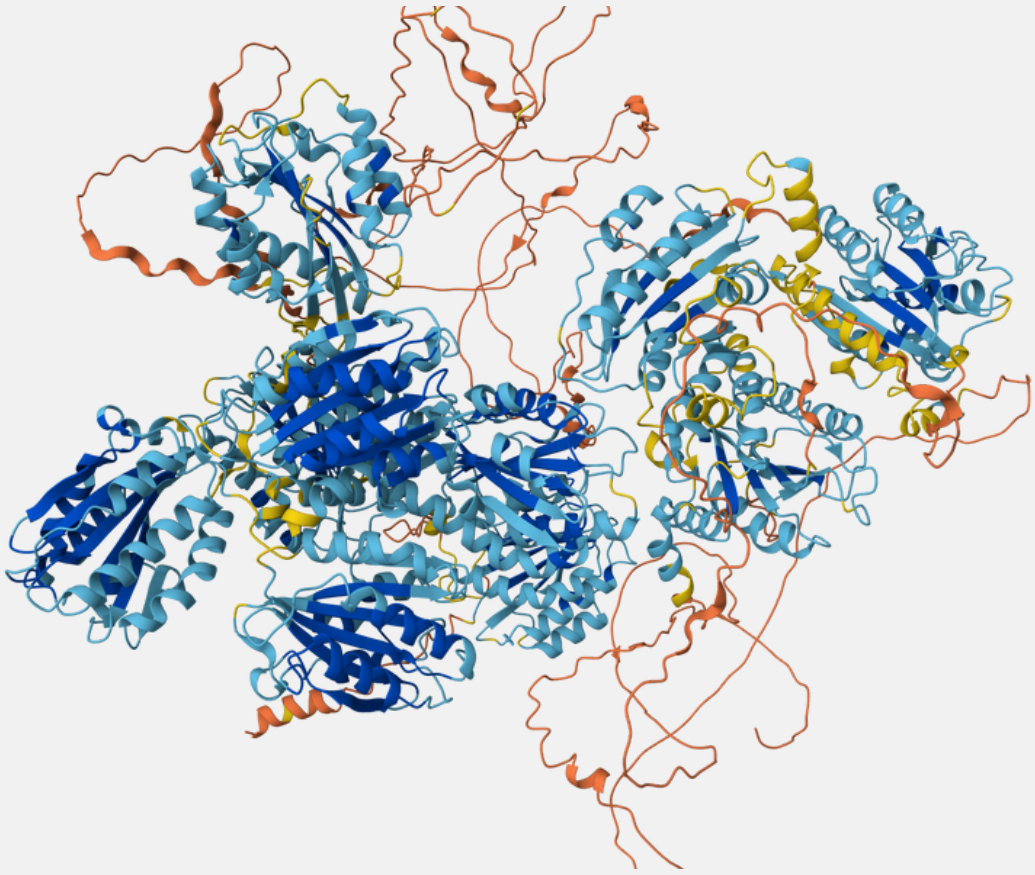


You now know the basics of data visualisation in R!
Try playing around with the plot function for different outputs:

```
plot(iris$Sepal.Length,
iris$Petal.Length, pch=12, cex=3,
lwd=4, lty=4, type="b", col=colors()
[sample(600,5)]
[as.numeric(iris$Species)] )
```

PROTEIN OF THE ISSUE

Collagen alpha-5(VI)



The protein responsible for **strengthening and supporting** tissues in the body, including skin, cartilage, tendon and bone.

- Its **triple helix structure** comprises of two alpha1 chains and one alpha2 chain.
  - The alpha chains are each composed of **three polypeptide chains.**
  - Each alpha chain contains approximately **1000 amino acid residues!**
- Without it, you'd be floppy and fragile!

Our Sponsors

BINFSOC

Thank you to all our partners and sponsors for their continuous support.

Red Bull is the number-one energy drink manufacturer, partnering with leading athletes and organisations around the world

Sponsor Us

If you would like to become a sponsor of BINFSOC, please reach out to sponsorships@unswbinfsoc.com

Contact us

IF YOU HAVE ANY COMMENTS or feedback regarding BINFsights, please write to us at binfsights@unswbinfsoc.com

We also encourage anyone to share with us anything you'd like us to take a look at, be it a bioinformatics tool that you have made or find useful; or news in the bioinformatics world that you'd like to see written about in future issues.

TO VIEW PAST AND PRESENT issues of BINFsights, check out our website at unswbinfsoc.com/binfsights
Stay tuned on our Facebook page for updates regarding events and society news.

-- The BINFSOC Team

BINF sights.

Acknowledgements

HAFSA FAHAD

PUBLICATIONS/IT EXECUTIVE & TEAM LEAD

RIVA MANTER

PUBLICATIONS/IT DIRECTOR

PHOEBE TANDJIRIA

PUBLICATIONS/IT DIRECTOR

Angelina Tran

PUBLICATIONS/IT SUBCOMMITTEE MEMBER

Suhani Jones

PUBLICATIONS/IT SUBCOMMITTEE MEMBER

Ethan Morritt

PUBLICATIONS/IT SUBCOMMITTEE MEMBER

Zoe Brookes

PUBLICATIONS/IT SUBCOMMITTEE MEMBER