A beginner's guide to bioinformatics

Krutik Patel (Faculty of Medical Sciences, Newcastle University, Newcastle Upon-Tyne, United Kingdom)

Bioinformatics has revolutionized the modern life sciences and has become a component of many undergraduate training courses and post-graduate research projects. As such, we are seeing more bioinformatics and programming aspects within undergraduate training and so it is important to understand what bioinformatics is, why it is a necessity in modern research and how young academics can begin their journey as bioinformaticians. This article outlines the broad spectrum of what bioinformatics is used for within research labs and provides several resources for beginners to learn how to code and perform bioinformatic tasks.

Introduction to bioinformatics

When imaging a bioinformatician, you may think of a dishevelled individual sitting in a dark room, frantically writing code and speaking in a complex and unintelligible jargon. While this caricature may often be portrayed, bioinformatics has become an integral part of research and is commonly found as a component of undergraduate and post-graduate programs. Bioinformatics or computational biology (which are the same thing) has quickly jumped from a specialist term describing an elite group of biologists that swapped their pipettes for *Python* and *Pearl* to a more generic description which encapsulates any form of computational analysis on biological data. These skills were utilized by higher education institutes during the pandemic as undergraduate and post-graduate students were often asked to undertake alternative capstone projects completely based online. Bioinformatic skills are also heavily required within the research environment as the volume of data being generated by labs is increasing, so the demand is growing for those with the skills to process and analyse large quantities of data.

The demand of bioinformaticians can be reflected by the increasing number of institutions in the UK which enrol prospective students into biology courses with bioinformatic modules or entirely computational post-graduate training courses. Requirements for bioinformatics masters' courses often require a 2:1 in a relevant bachelor's degree. Several drivers have influenced this increase in bioinformatics-related courses such as a greater appreciation of how skills in programming can hasten research outcomes, a data-boom in biology which requires specialist analysis, an ever-increasing demand to increase automation and efficiency in research and industry and the potential benefit machine learning approaches have to offer research. These and other factors are influencing the landscape of modern research and are clearly explained by the references below, which are recommended as additional reading to those interested in the subject.

How does bioinformatics work?

During an undergraduate capstone project, academic staff may instruct their students to perform bioinformatics – which is becoming an increasingly vague term. So, first, I believe it is important to understand what bioinformatics is. At its heart, bioinformatics is the use of computation to understand more about biology and should not be thought of as a replacement to traditional laboratory research – rather a synergistic partner. Another way to phrase it would be to use data science approaches to ask questions in the life sciences.

To help visualize the question 'what is bioinformatics', we can try to frame how bioinformatics works in research. Many questions from life science come from the natural world/*in vivo* – this can include any range of topics such as investigating the molecular causes of diseases, contrasting virulence of different pathogens or researching the origins of human evolution. Researchers could then replicate and test what was seen in the natural world in the laboratory using *in vitro* techniques such as a western blot to measure the effect a gene mutation has on the protein expression level of a known oncogenic gene, or infecting cells with a pathogen and contrasting the viability between infected cells and non-infected cells. A bioinformatician would then use what has been measured *in vitro* to inform computational analysis (*in silico*). The analysis itself can range from measuring samples in parallel to reduce manual effort, running specialist software for quality control or simply producing graphs in a popular coding language like *Python* or *R*. The goal of the analysis could be to generate testable hypothesis which can be validated back in the

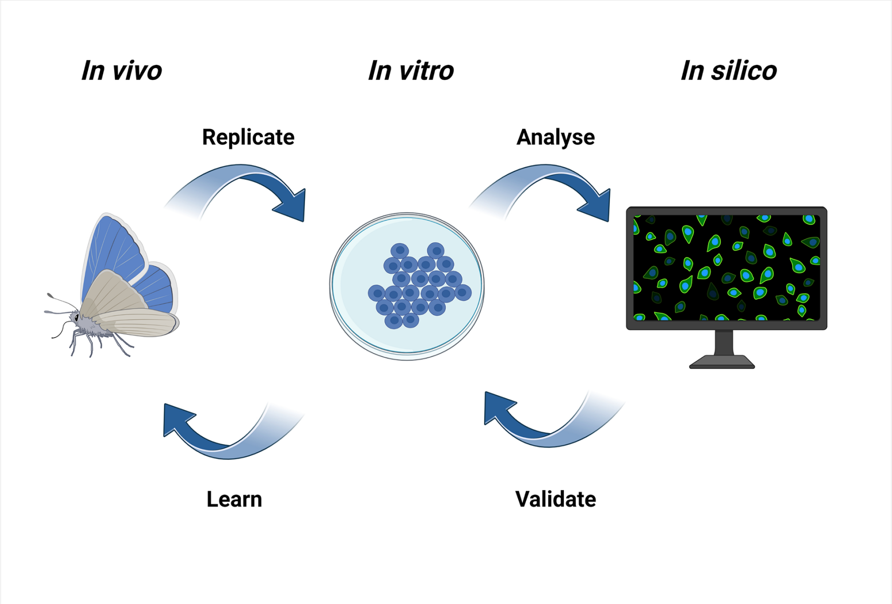

Figure 1. How bioinformatics contributes to biology. Figure made with BioRender. Summary of how bioinformatics works within research. Questions from the natural world can be replicated within a lab and then analysed using computational techniques. Bioinformatic analysis can direct the validatory tests which would need to be conducted within a laboratory setting assess the processed and learn new information.

laboratory, and this in turn would lead to learning something new about the natural world (Figure 1).

Why is bioinformatics necessary in biological sciences?

Bioinformatics supports every aspect in modern biological research. In some ways bioinformatics is not a useful term for the vast array of different tasks that can be performed for uses in biology (Figure 2). Illustrated in Figure 2 is a portfolio which shows the range of tasks from distinct disciplines within biological research that require an expert in bioinformatics. Building machine learning models to make predictions or classifications based on biological data is a novel method of using large complex datasets to find patterns which cannot be detected without the use of powerful algorithms. Such machine learning tasks are becoming a popular avenue with the increasing availability of big data in biology. Big data refers to the exponentially expanding amount of biological data (e.g., sequencing data, gene expression data, population level data), which bioinformaticians must work with. Running a DNA sequencing workflow to detect genetic variations has become the optimal method of diagnosing patients with rare diseases, and this is shown by the success of the 100,000 Genomes project in the UK, which has identified many novel rare diagnoses. Building evolutionary trees based on ancient DNA samples is a necessary step to accurately understand species evolution and this can be powered by intelligent algorithms. Contrasting disease and healthy control samples to identify biological markers is a heavily researched area for complex conditions such as cancers and neurodegenerative disorders. Finally, graphical representation of large data such as clustering graphs (e.g., heatmaps) to identify similarities between biological samples is useful and can lead to efficient methods of identifying patterns. In each of these examples the specific context is less important than the overall theme of using data science approaches to unearth novel insights from biological information.

Skills required for bioinformatics

Now that we have a better grasp of what bioinformatics is, we go through the skills required to become a bioinformatician – either for a project or for career development.

Bioinformatics is centred around programming

Coding is at the core of bioinformatics, as bioinformaticians are expected to produce specialist scripts – lines of code to perform functions for specific tasks. The two most well-used programming languages in modern biological research labs are *R* and *Python*, and both have their merits.

R has a bigger biological community and much of this is driven by the *Bioconductor* project which contains thousands of biology-based tools written in *R*. With many ready-to-use applications, the burden of labour

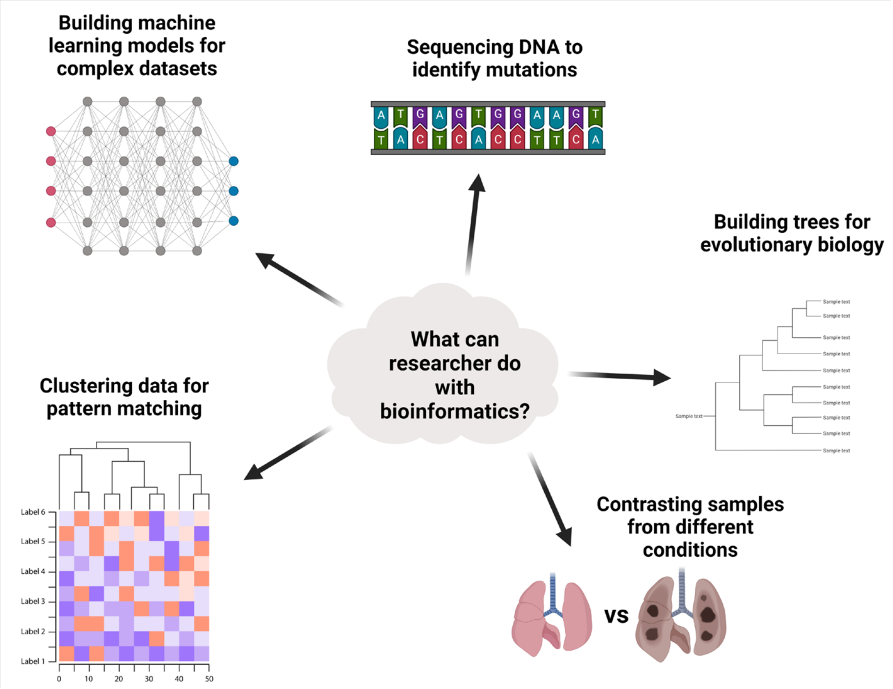
Building machine learning models for complex datasets
Sequencing DNA to identify mutations
A T G A G T G G A A G T
T A C T C A C C T T C A
Building trees for evolutionary biology
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
Sample text
What can researcher do with bioinformatics?
Clustering data for pattern matching
Label 6
Label 5
Label 4
Label 3
Label 2
Label 1
0
10
20
30
40
50
Contrasting samples from different conditions
VS

Figure 2. Examples of tasks which can be classed as bioinformatics. Figure made with BioRender. Several bioinformatic techniques include building machine learning models, sequencing DNA samples, building evolutionary trees, contrasting expression data between disease and control groups and visualizing data.

falls less towards script development, making *R* very useful for shorter projects such as capstone or master research projects. Another advantage of *R* is its popular integrated development environment (IDE) *Rstudio*, which provides a very accessible layout for scripting. The major disadvantages of relying on *R* is that many of the error messages are difficult to decipher and that tools can become out-of-date quite quickly (<2 years).

On the other hand, *Python* also has tools for biological use and, in contrast to *R*-based tools, can often continue functioning for decades; the error messages in *Python* are generally more easily understandable; and *Python* is faster than *R* making it more useful for larger datasets. Another advantage of *Python* is that it has many complete end-to-end and user-friendly machine learning-based tools that contain many algorithms, machine learning processes and structures available for use, such as *scikit-learn* and *keras*, which are becoming more popular for use in biology. However, working in *Python* means that researchers must write more extensive scripts compared to *R*. Additionally, *Python* has many more options for IDEs than *R*, such as *Spyder*, *Jupyter Notebook* and *Pycharm*, and it may take time to find an ideal one. Pycharm is great for experienced coders, and Jupyter is great for learning – making it an excellent choice for your first python project. My personal preference is *Spyder*, as in my opinion, its design is well suited for analysing complex biological datasets.

Overall, both *R* and *Python* have their advantages and most tasks like those shown in Figure 2 can be performed with either. There are distinct advantages when using *R* such as the many ready-to-use tools which can make scripting tasks simpler. But, in biological research right now, *Python* is currently far more useful for machine learning-based tasks.

Resources to begin coding

There are several ways to increase knowledge about coding. Courses are a great way to begin, such as those promoted by Datacamp. Even watching free tutorials on YouTube can be an effective learning strategy to learn programming and more about bioinformatic concepts or tools. Some brilliant channels include StatQuest which goes over statistical and bioinformatic application in short videos and the Bioinformatics-along learning videos which are a collection of hour-long videos going over bioinformatic tool use. I have provided a short playlist of helpful videos for beginners. Another good resource are the people around you – do not be shy about asking for help or saying you do not know how to perform certain tasks. However, by far the best method is to get stuck in and learn by doing. To get started, create a small project which consists of uncomplicated tasks which can be performed in excel, e.g., create a bar graph from some data or calculate the mean of each row in a table and try them out in *R*, *Python* or any other

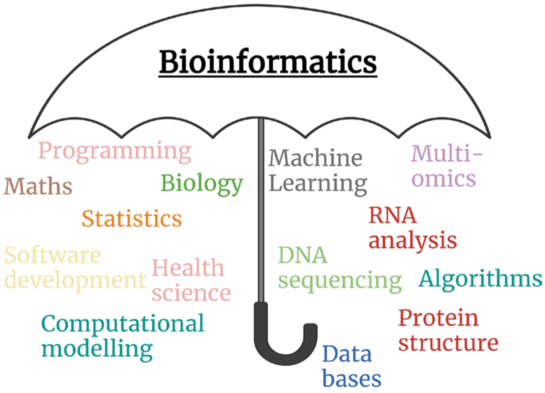

Figure 3. Showing that bioinformatics is an interdisciplinary field. Figure made with BioRender. Bioinformatics is an umbrella term for multiple disciplines within biology.

language you are interested in learning. Many interesting datasets to begin coding can be freely downloaded from kaggle. For more guided project ideas, there are several resources online, such as blogs by Datacamp (*R*) or Upgrad (*Python*). If you get stuck, look online. Most coding-related questions have probably been asked and are probably on a forum on the web. Good forums for bioinformatics-related issues include SEQanswers and Biostars, and most coding issues may have answers written in Stack Overflow.

Resources to begin a machine learning project

Like the points made above, there are good courses to begin learning how to begin a machine learning project on sites such as Datacamp. Again, a brilliant resource is kaggle, which contains downloadable datasets, readymade projects aimed at beginners, and has regular challenges for beginners and more advanced programmers. The biochemical society also runs online courses in *R* and *Python*. Simple machine learning projects could include creating a model to differentiate between cats and dogs or predicting the preferred coffee choices from staff at a school. The skills learnt through such training can be transferred to creating a model to differentiate cancer and non-cancerous tissue images or creating a model to predict genes important in neurodegenerative diseases.

Another part of machine learning is the theory of why certain practices are more efficient than others and a good resource to better understand this is towards data science, which houses blogs from talented data scientists, the blogs containing the code and rationale behind the code and are all clearly written for both experts and non-experts. For those that prefer videos, the Andrew Ng's Stanford lecture course is available on YouTube and is great for those who want to grasp theory and begin developing applicable ideas.

Further skills important in bioinformatics

There are several further qualities which can be very helpful in bioinformatics. First, as expanded on in this article, hands-on experience in coding is very useful as it shows aptitude for creating scripts and performing programming tasks. Creating a portfolio of scripts you have created is a good way to show evidence of your skills. Such scripts could be stored as a public repository in Github, which would also be an impressive resource to have in your CV. Also, there is a big advantage in understanding the fundamentals of the biological questions you are interested in answering using bioinformatics. Being a bioinformatician does not mean one should ignore or be abstinent to the biological questions being researched. This is quite a misconception and, in all honesty, not understanding the data makes the computational tasks more difficult because there is little or no use in performing analysis or creating machine learning models which misinterprets the research question. Another skill is to be flexible and understanding that bioinformatics has become an umbrella term which covers a wide range of interdisciplinary elements including programming, software development, maths, statistics and much more (Figure 3) and so to work in the field one must develop the skills to be flexible. Furthermore, it is also important to know about current trends in the field. Machine learning/AI, big data techniques and multi-omics (combined DNA, RNA and/or protein) analysis are currently the major themes in the frontline of modern bioinformatics and there is a need for skilled individuals to research and work in these areas. ■

Further Reading

Further reading on the current demand on bioinformatics in life science

- Attwood, Teresa K., et al. "A global perspective on evolving bioinformatics and data science training needs." Briefings in Bioinformatics20.2 (2019): 398-404. https://doi.org/10.1093/bib/bbx100
- Gauthier, Jeff, et al. "A brief history of bioinformatics." Briefings in bioinformatics20.6 (2019): 1981-1996. https://doi.org/10.1093/bib/bby063
- Kanehisa, Minoru, and Peer Bork. "Bioinformatics in the post-sequence era." Nature genetics33.3 (2003): 305-310. https://doi.org/10.1038/ng1109

Bioinformatics/coding resources

- Bioconductor - Home. Resource with over 2000 working bioinformatics tools made to work in *R*.
- Learn R, Python & Data Science Online | DataCamp. Full of programming and machine learning courses for a range of skill levels.
- StatQuest with Josh Starmer - YouTube. YouTube channel for bioinformatic and statistical concepts.
- Simon Cockell - YouTube. YouTube channel for longer bioinformatic tutorials.
- Author curated playlist of several YouTube videos which could be very helpful to new bioinformaticians.
- Kaggle: Your Machine Learning and Data Science Community. Resource to begin machine learning practice and datasets to use can be found here. Find Open Datasets and Machine Learning Projects | Kaggle. – Personal preference for free datasets.
- 10 interesting R project ideas and links to data sources by DataCamp.
- 42 interesting Python project ideas and links to data sources by UpGrad.
- Towards Data Science. Machine learning blogs to gather theory and coding examples.
- Bioinformatics Answers (biostars.org). Forum for bioinformatics queries.
- Forums - SEQanswers. Forum for specifically sequencing related queries
- Stack Overflow - Where Developers Learn, Share, & Build Careers. Forum for coding queries
- Stanford CS229: Machine Learning Course, Lecture 1 - Andrew Ng (Autumn 2018) - YouTube. Lecture course on machine learning and practical advice.
- Github is a free and useful method of keeping track of code and projects. Very attractive on a CV.

Krutik Patel recently completed his PhD in bioinformatics at Newcastle University and has since been employed as a research associate/bioinformatician at Newcastle University. His interests are in applying data science techniques for interesting questions in biology and developing software for researchers. Email: Krutik.Patel@newcastle.ac.uk