An Extendible clustering package for bioinformatics.

Project plan

Karan Alreja – 2024/25

Final Year Project

Supervised by: Sunas Sourjah

Department of Computer Science Royal Holloway, University of London

1 Abstract

Biologists began sequencing proteins due to their critical role in cellular activities such as metabolism, cell structure, and communication. The first protein's amino acids which were successfully sequenced was insulin by Sanger, F., and E. O. P. Thompson, this proved that proteins could be systematically and successfully analyzed which further sparked more study. (SANGER, 1953) (F., 1955)

Protein analysis also helped biologists understand how mutations in protein sequences could lead to diseases. For example, the analysis of hemoglobin revealed that sickle cell anemia is caused by a single amino acid substitution . (VM, 1956)

Following Sanger's success with insulin, other biologists also started their attempts on protein sequencing. Margaret Dayhoff, known as the "mother and father of bioinformatics," played a pioneering role in this bioinformatics field (G, 2004) She, along with Robert S. Ledley, created COMPROTEIN, the first software designed to determine protein structure, marking one of the first uses of computers in sequencing. Dayhoff also founded the first biological sequence database, the Atlas of Protein Sequence and Structure. (Margaret O. Dayhoff ... [etal.], 1969)

Following the success of protein sequencing, biologists recognized that proteins, while crucial for cellular functions, were ultimately products of genetic information encoded in DNA. This shift came from the understanding that - specifications for any living being (more precisely, its 'proteins') are encoded in the specific nucleotide arrangements of the DNA molecule. This view was formalized in Francis Crick's sequence hypothesis (CRICK, 1970), in which he postulated that RNA sequences, transcribed from DNA, determine the amino acid sequence of the proteins they encode. In turn, the amino acid sequence determines the three-dimensional structure of the protein, therefore if one could understand how to translate DNA into protein sequences, we could get the all possible proteins producible by an organism by looking at its DNA. (Jeff Gauthier, 2019)

As seen by COMPROTEIN, computers became invaluable in bioinformatics because protein and DNA sequencing requires vast amounts of comparison, pattern matching, and calculations, tasks that computers excel at (R, 1979). The rapid development of sequencing techniques for proteins and DNA led to an explosion of biological data. This "big data" challenge was compounded by the growth of the World Wide Web in the mid-1990s, allowing for unprecedented data sharing between biologists.

This increase in data coincided with Moore's Law, which predicted the exponential growth of transistors in CPUs. However, by 2008, the sheer volume of biological data had far outpaced the growth of computational power. This biological "big data" has since been organized into large online databases such as GenBank (1995), PubMed (1997), and the Human Genome Project (1999) (Jeff Gauthier, 2019)

The rise of bioinformatics as a discipline has led to the emergence of bioinformaticians— as per the international Society for Computational Biology bioinformaticians are – "[using] current techniques, skills, and tools necessary for computational biology practice', '[applying] statistical research methods in the contexts of molecular biology, genomics, medical, and population genetics research' and 'knowledge of general biology, in-depth knowledge of at least one area of biology, and understanding of biological data generation technologies."

With the development of bioinformatics, scientists are able to handle and examine the enormous volumes of data produced by sequencing projects, revolutionizing biological research. Better computational tools and publicly available biological databases have not only sped up the rate of discovery but also made accessible data, enabling researchers from all over the world to work together and add to the pool of knowledge. From mapping genetic variants to predicting protein structures, bioinformatics will continue to be crucial for understanding complicated biological systems as sequencing technology develop. In the future, this multidisciplinary subject will likely influence biological discoveries by pushing advances in environmental science, agriculture, and medicine.

As biological data grew larger and more complex, machine learning techniques, like clustering algorithms, became essential tools for interpreting this data. Clustering methods such as k-means, hierarchical clustering, and self-organizing maps are being used to classify gene expression profiles, protein structures, and other biological data, helping scientists to discover patterns and relationships that were previously hidden in vast datasets. For example, clustering algorithms have been used to identify subtypes of diseases, such as cancer, by grouping patients based on gene expression profiles, which lead to better and more personalized treatments (Eisen MB, 1998). Machine learning also help biologists to predict protein functions and interactions by recognizing similar sequences across organisms (Marcotte EM, 1999).

Processing such data effectively requires advanced software tools that can manage the computational level. Software has become vital for automating tasks like pattern recognition, statistical analysis, and visualization, enabling researchers to process enormous volumes of biological data with speed and accuracy (as talked about earlier). Programs like BLAST, used for comparing biological sequences, and tools for protein folding prediction show how software has improved bioinformatics. By using clustering algorithms and visualization techniques in the software, biologists can transform raw data into important insights （Altschul SF, 1990）. In this project, my plan is to create a software that doesn't only implements these clustering methods but also provide an interface for visualizing and these interpreting biological datasets, making it a useful tool for researchers in genomics, proteomics, and systems biology.

2 Timeline

My plan is to start reading up on SE principles (plug in architecture specifically) that can be applied with biological data(microarray data and sequence data) with having skeletal software which can run some clustering algorithms(k means clustering, hierarchical clustering.) on some biological datasets, this will ensure that my basics are covered and I can work on polishing and revision in term two.

My main goal for term two is to complete the work on user-interface (JavaFx), visualization techniques such as Heatmaps, Scatter plots, Dendrograms and ensure complete implementation of the software for extendibility with plug-in software, also work on validity and bug fixing.

Reading up on SE principles, clustering algorithms, and software systems (spring boot, JavaFx) at the start of term one will ensure that the requirements are well understood, and the project has a good base to build work on. Working on polishing and finishing project in term two will ensure that I understand the requirements of the project from term one, and I can build a better finished product while also giving myself time for bug fixing and any future addition that I might want to make.

My tech stack that I plan to use:

Programming language: Java with Spring boot for full-stack management.

For parsing and storage of biological data I plan on using an PostgreSQL database.

For user-interface I plan on using JavaFX since I have used it for prior projects.

2.1 Term 1

Week 1-2: Do a background study on bioinformatics, understand its role within computer science and study project specifications – while also working on project plan.

Week 3-4: Study SE principles (java plug-in technology to make the software extendible), clustering algorithms (k means clustering), parsers for biological databases/data, software's to be used.

Week 5-6: Start working on code, designing software, and converting clustering algorithms into code.

Week 7-8: Work on getting biological data set up in database, make clustering algorithms work on these databases and start working on user interface.

Week 9: Bug-fixes and polishing up software.

Week 10-11: Work on interim report and presentation.

2.2 Term 2

Week 1-2: Revise project requirements, and check project against specifications. Go over additional clustering techniques and biological datasets that can be feasible to be implemented into the project.

Week 3-4: Work on implementing clustering algorithms and biological datasets.

Week 5-6: Work on user interface, work on visualization techniques.

Week 7-8: bug-fixing, testing, and implementing any extensions that seem appropriate.

Week 9: Final reviews and check against specifications.

Week 10-11: Work on final report and presentation.

3 Risks

With any software project there exists associated risks, though this section I will talk about few general risks and few specific ones.

3.1 Data loss

It is possible during this project that some hardware might fail which can cause my project to fail, therefore, I will make sure to keep committing my code, report and diary on suitable cloud platforms that will ensure that my data is backed-up and safe, e.g., GitHub, Overleaf.

3.2 Report-Code balance

Code and Report are both vital part of any software program, but it is possible that I might end up fixating and working on one and leaving the other behind, to ensure that this doesn't happen - I will make sure to work on important parts of the code through the term and leave a few weeks at the end to wrap up the report.

3.3 Biology overhead

During the project it is possible for me to work out computational problems, but there are a few biological aspects to this project that might cause hiccups, if I find myself lost, I will consult relevant sources and get in touch with my supervisor so that she can guide me.

3.4 Machine learning

Over the course of this project, it is possible for me to go down a rabbit hole of over fixating on clustering algorithms rather than making a software that allows them to be used with flexibility in the program. To ensure that I don't do this I will ensure to stick to my timeline, work according to specifications and update my advisor through consistent meetings through the term.

3.5 User-Experience

It Is possible that I end up making the code, user interface too complicated to use and without proper documentation which might make it very hard to add plug ins or use as a stand-alone software. To ensure that the software and interface remains as simple and user-friendly as possible - I will make sure to have plenty of testing with people and consistently work on documentation.

3.6 Code issues

Since this code is to be used with plug-ins and heavily relies on extendibility, I might end up making the codebase to rigid or out of spec, which won't allow it to work as intended. I will make sure to learn and stick to proper Software engineering principles and periodically go over the specifications of the project.

Bibliography

Altschul SF, G. W. M. W. M. E. L. D., 1990. Basic local alignment search tool. *J Mol Biol,* 215(3), pp. 403-10.

CRICK, F., 1970. Central Dogma of Molecular Biology. *Nature ,* Volume 227, p. 561–563 .

Eisen MB, S. P. B. P. B. D., 1998 . Cluster analysis and display of genome-wide expression patterns.. *Proc Natl Acad Sci U S A,* 95(25), pp. 14863-8.

F., S., 1955. The structure of insulin. *Bulletin de la Societe de Chimie Biologique,* 37(1), pp. 23-35.

G, M., 2004. *Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and business.*. London: Wiley.

Jeff Gauthier, A. T. V. S. J. C. N. D., 2019. A brief history of bioinformatics. *Briefings in Bioinformatics,* 20(6), p. 1981–1996.

Marcotte EM, P. M. T. M. Y. T. E. D., 1999. A combined algorithm for genome-wide prediction of protein function.. *Nature.,* 402(6757), pp. 83-6.

Margaret O. Dayhoff ... [etal.], 1969. *Atlas of protein sequence and structure. [Vol. 1].* Volume 1 (M. O. Dayhoff, Ed.). ed. Silver Spring: National Biomedical Research Foundation.

R, S., 1979. A strategy of DNA sequencing employing computer programs.. *Nucleic Acids Res,* 6(7), pp. 2601-2610.

SANGER, F. &. T. E. O., 1953. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal,* p. 353–366.

VM, I., 1956 . A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin.. *Nature,* 178(4537), pp. 792-794.