Text Mining over biological data to predict protein stability and solubility

Jean-Marc Mahoro

University of New Brunswick Saint John, Saint John NB E2K 5E2, Canada

jean_marcc@hotmail.com

Abstract. Many biomedical and biotechnological applications find use in proteins. We need to know the stability [4] of these proteins and whether these proteins are soluble because their production and engineering critically depends on it. To know the stability and solubility of these proteins can be done by reading a biological corpus including such data. But, for research scientists to physically collect all the data, and also have to read through each paper to find out the stability and solubility of the protein would be time consuming. We'd need more employees specialized to do this job which means more time and funds being used. Knowing that there is a more efficient way of doing this would make this a waste of time and resources. The more efficient method this project proposed is to use text mining to accurately predict stability and solubility on a corpus [6]. There are factors that may affect the process of stability and solubility, by using these factors inside our code we hope to be able to determine a prediction of protein stability [4] and protein solubility. This will be important to save both laboratory experimental effort and financial resources. To attempt to predict protein stability and solubility we will be text mining [3] over biological corpora which would leave us to find out which factors give a more accurate prediction of stability and solubility [1]. This method of text mining could be used for much more than just stability and solubility prediction, this project will show research scientists that need different data mined how to do it.

Keywords: Solubility · Soluble · Stability · Protein · GATE · Text Mining · Prediction · Mutation · Jape Rules · JAPE · Corpora · Corpus.

Introduction

Biomedical and biotechnological applications find use in stable and soluble proteins. Only certain stable proteins are suitable for development, but as of now finding stable protein adds to the cost and time of finding new drugs. This is because research scientists and drug companies find determining protein stability a headache [7]. The methods are time consuming and costly [8].

There is a software called GATE [2] which allows for text mining over data. This project suggests cutting the cost and time spent by simply text mining using GATE

over a corpus [6] of papers from laboratory results, and combining this software to have the Java program predict the stability and solubility for the entire corpus with the click of a button.

Motivation

Dr. Baker who works with research scientists from all over the world approached me with a project which could benefit these research scientists. Since proteins find use in numerous biomedical and biotechnological applications, it would be important to know its solubility and if the protein worked with is stable. Factors such as high temperature, pH, protein concentration, etc. can affect the process of solubility and stability [9].

Knowing whether protein is stable or soluble during an experiment would allow for many resources to be saved, this is where this project comes in. It would be critical to be able to predict protein stability and solubility to save time and money for the research scientists. This project is for the research scientists and the motivation is to save them on both time and money.

Problem Statement

An example of why this is a project of interest are Monoclonal antibodies. They are a type of protein derived from natural antibodies. Monoclonal antibodies steadily help transforming the way we treat and prevent diseases from cancer and conditions affecting the immune system to viral infections. There is even more interest in mAbs (Monoclonal antibodies) since the coronavirus pandemic since they're showing potential to treat Covid-19, and are currently being trialled in humans. The catch is that mAbs need to be stable for it to be appropriate for development [7]. Since the current methods [8] to determine stability causes research scientists and drug companies a headache [7], this project focuses on bringing a solution to the current headache by text mining over data to predict solubility and stability.

Some of the known challenges are integrating GATE inside Java. When doing this you can simply add jape rules [2] in the Java program.

The unresolved issues are the fact that this program cannot be downloaded and work right away, the README file needs to be completed step by step in order for this to work on another person's device.

Researchers and scientists interested in proteins will benefit from the outcome [9], drug companies will benefit from the outcome [7], and most important of all humanity may benefit from the outcome if the prediction of protein stability correctly classifies a protein as stable which then turns out to be the protein to treat Covid-19 [7].

Goals

My personal goal was to add jape rules designed by Shawn Kroetsch into the Java program which would then be able to predict stability and solubility. These rules would text mine over data for mentions related to mutation stability, solubility, the direction, and more. The Java program need to be combined with the jape rules from the GATE [2] pipeline. After adding the jape rules into the java program and running the GATE pipeline through the java program, this program should be able to predict stability and solubility.

Dogma

Some of the background information required to understand my work theory wise is that only stable proteins are appropriate for development [7]. Stable protein has potential to treat Covid-19 [7].

Common knowledge about protein stability and solubility. Protein is soluble if it prefers to interact with water, rather than with other protein molecules. Protein is also stable if it is properly folded.

Core Methodology

The technical step this project is using to achieve its goals are adding stability and solubility rules to the current Java program. This is done by adding JAPE rules that Shawn Kroetsch worked on. I then plan to combine this Java and GATE [2] to text mine [3] over the collected data and activate these rules. The next step is to achieve a desirable stability and solubility prediction using the precision/recall measures from GATE [2]. The following picture is a diagram of the simple steps followed:

Fig.1. A visual of the architecture model

Resources

My personal hardware used is a 2016 MacBook Pro. This is not one of the best devices to use today but it delivers the solutions without problems.

A third-party software used is called GATE [2]. GATE is a software that allows for you to text mine over a corpus [6] with specific jape rules [2] attached which determines the language to look for. Another third-party software would be Java. Java is a popular programming language which I will be using alongside GATE to text mine and get stability and solubility results.

Benchmarking data sets

For knowledge resources I searched articles online and links sent by Dr. Baker.

Describe your Results

As I am currently able to run the program without the added jape rules [2] and only the old jape rules, I am still able to go over the results as the only difference will be a bit more information on the result data.

The code I added was a call to the new jape rules [2] to add them when running the program.

Some new knowledge I generated is that you can integrate a third party software that can text mine and use it inside Java. This is efficient and allows for quick text mining and getting result by just running the program. The results are more efficient than existing studies.

Updated GANTT Chart

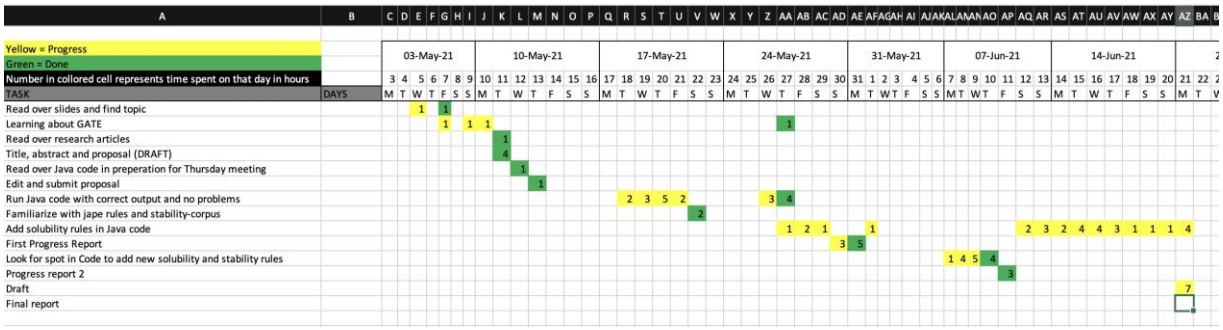

Fig.2. GANTT chart May 31, 2021.

Evaluation

I evaluate my work by the amount of time spent, the effort put in and the results that then benefit other people. I spent a fairly amount of time on the project. Since the project is not one hundred percent yet, I cannot say that I evaluate my work to be great.

Future

Most important is get full functionality.

To improve the outcome of this project, we need to add all the new jape rules [2] and reach our goal to have the code run the program and give the stability and solubility predictions. Some design improvements could be adding some more comments in the text, so the next person has a less difficult time looking for stuff.

In future work I'd have to use different corpora to find stability and solubility in new protein that haven't been tested yet.

Research scientists who are interested in protein should know about my results.

Conclusion

As you can see, this project is important and bring breakthrough results in the medical field. This project will deliver easy access to stability and solubility predictions which will benefit research scientists from all over the world and drug companies. I already gave an example of the possibility of protein stability being able to treat Covid-19.

The contributions made by me to the state-of-the-art is that there's a new tab in the result file for stability. Instead of doing all these costly methods, you can now check the stability with the click of a finger.

My prototype is still an infant and needs to be tweaked for full functionality.

Critical Commentary

I'd say that the most crucial lesson I learned is that a bug should be fixed as soon as possible, not later.

The tools I worked with are Java and GATE.

I followed the methods already in the code but changed them according to the new jape rules.

Basic knowledge of GATE and Java (probably a bit more understanding of Java) is required to be able to advance this project, and from there you'll gain more deeper knowledge in the subject.

Updated References

1. Musil, M., Konegger, H., Hon, J., Bednar, D., and Damborsky, J., 2019: Computational Design of Stable and Soluble Biocatalysts. ACS Catalysis 9: 1033-1054.
2. Cunningham, et al. Developing Language Processing Components with GATE Version 8. University of Sheffield Department of Computer Science. 17 November 2014.
3. Witten, I.H., Don, Katherine J., Dewship, M. and Tablan, V. (2004). Text miningin a digital library. International Journal on Digital Libraries, 4(1), 56-59.
4. Winnenburg, R., Plake, C. Schroeder, M. Improved mutation tagging with gene identifiers applied to membrane protein stability prediction. BMC Bioinformatics 10, S3 (2009). https://doi.org/10.1186/1471-2105-10-S8-S3
5. Klein A, Riazanov A, Hindle MM, Baker CJ. Benchmarking infrastructure formutation text mining. J Biomed Semantics. 2014;5(1):11. Published 2014 Feb 25. doi:10.1186/2041-1480-5-11
6. Merriam-Webster. (n.d.). Corpora. In Merriam-Webster.com dictionary. RetrievedJune 21, 2021, from https://www.merriam-webster.com/dictionary/corpora
7. University of Bath. (2021, May 27). Technology predicts protein stability. ScienceDaily. Retrieved June 21, 2021 from www.sciencedaily.com/releases/2021/05/210527150202.htm
8. Man, T. P. (2018, September 18). Methods of Determining Protein Stability.https://info.gbiosciences.com/blog/methods-of-determining-protein-stability.
9. SoluProtMutDB. (n.d.). https://loschmidt.chemi.muni.cz/soluprotmutdb/