Bioinformatics Summer School Report

Name: Faisal Alkulaib
Internship Organization: University of Toronto
Supervisor's Name: Gary Bader
Project title: Enhancing Named Entity Normalization in Biofactoid Using NLP Techniques

1. Introduction

During my internship, I worked on enhancing the gene normalization process in Biofactoid, a web tool that assists authors in curating scientific findings into a standardized, machine-readable format. My primary objective was to improve the accuracy of gene normalization by using natural language processing (NLP) techniques to extract contextual information, which we referred to as "hints." These hints were integrated into the Biofactoid system to reduce gene normalization errors, particularly orthologue mismatches, thereby enhancing the reliability and accuracy of curated biomedical data.

2. Materials and Methodology

The project involved using PubTator 3.0 to extract contextual information from article titles and abstracts. These "hints" were incorporated into Biofactoid's custom search engine called Grounding Search to correct gene normalization issues. The grounding search system was employed to identify biological entities by mapping user-provided names to standardized database identifiers, ensuring accurate data retrieval and integration.

3. Results and Discussion

The implementation of NLP-derived hints significantly improved the accuracy of gene normalization in Biofactoid. The enhanced process successfully corrected prior mismatches, such as associating the gene "Aida" with the correct species, thus improving the reliability of the curated data. This improvement reduces the likelihood of incorrect gene-species associations and enhances the quality of biomedical research data.

4. Skills and Knowledge Gained

Throughout the project, I gained valuable skills in natural language processing and bioinformatics, particularly in the application of NLP techniques for entity normalization. Additionally, I acquired significant experience in using GitHub, learning how to manage a project from development through to deployment. This experience included version control, collaborative development, and deployment in a real-world bioinformatics setting. I also learned JavaScript and significantly improved my debugging skills, which further enhanced my ability to solve problems effectively. Additionally, I developed soft skills such as problem-solving and teamwork through collaboration with my supervisors and colleagues.

5. Challenges and Solutions

One of the significant challenges was dealing with orthologue errors in gene normalization, where the same gene name could refer to different species. This was addressed by implementing the NLP-derived hints, which provided additional context for correct gene identification. Future work will focus on addressing synonym errors to further enhance accuracy.

6. Conclusion

This internship provided me with an opportunity to apply NLP techniques in a practical bioinformatics project, significantly enhancing my understanding and skills in this field. The experience has influenced my career goals, particularly my interest in integrating AI techniques with bioinformatics to solve complex biological problems.

7. Acknowledgments

I would like to acknowledge my supervisors, Jeffrey Wong and Gary Bader, for their guidance and support throughout this project. I also extend my gratitude to the KAUST Academy and the Data Sciences Institute at the University of Toronto for their sponsorship and resources.

8. References (if applicable)

1. Franz et al., (2021). A flexible search system for high-accuracy identification of biological entities and molecules.
2. Wei et al., (2024). PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge.
3. https://github.com/PathwayCommons/factoid
4. https://github.com/PathwayCommons/grounding-search