Dylan Skalman
DSA 688 Internship Report

Introduction

An intein is a kind of protein that facilitates protein splicing, a biochemical reaction involving both the cleavage and formation of peptide bonds. The intein facilitated reaction results in a functional protein, with the intein excluded (1). Inteins are an emerging field of study and have considerable potential for biotechnological applications. The National Center for Biotechnology Information (NCBI) is a resource used by scientists that has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics as per their mission page (2). However, the NCBI has failed to recognize inteins as intervening sequences rather than proteins. This oversight has led to a significant amount of intein records on the US Government owned website that once displayed accurate intein data have been edited or removed, causing a misrepresentation of biological data. This internship project aims to gather and analyze evidence of this data misrepresentation to prompt the correction of inaccuracies in intein data, ensuring the integrity of biological research and biotechnological applications of inteins.

Dr. Novikova's "Intein Clustering Suggests Functional importance in Different Domains of Life" and related scientific publications highlight the importance of accurate data used in the emerging field of inteins. The premise of this paper details a bioinformatic survey of genomic data, revealing a biased intein distribution of inteins among functional categories of proteins in bacteria and archaea. As the growing number of fully sequences and annotated genomes continues (3), more data will become available for researchers, and it is important that the biological data stored by NCBI is accurate so that the integrity of research is not compromised. The methods of this study include "primary intein mining from protein databases available at the National Center for Biotechnology Information 9NCBI). Through mining, 2729 bacterial, 345 archaeal and 6648 eukaryotic genomes were screened for the prescence of inteins. Had this study been repeated today, many of these screened genomes would likely be misrepresented leading to different results and conclusions due to the misrepresentation of this biological data. Intein research such as this publication help to fuel emerging areas of intein research by providing invaluable insights to their distribution and potential applications in biotechnology. The distorted data has the potential to lead researchers to obtain inaccurate results and draw misguided conclusion about the distribution and function of inteins in various organisms, posing a threat to ongoing research.

The work for this internship was conducted at a home workspace and involved NCBI database queries, data manipulation and analysis in python and collaboration through zoom meetings with Dr. Novikova, professor and researcher and Buffalo State University. A research plan of inspecting intein data, collecting intein data and analyzing finding was developed and evaluated weekly.

A significant amount of intein containing protein records hosted on NCBI that were once accurate have been edited or removed, leading to this misrepresentation of biological data. The impact of this inaccurate data can compromise research accuracy and distort scientific conclusions in the emerging field of inteins. As an example, Dr. Novikova's "Intein Clustering Suggest Functional Importance in Different Domains of Life' study indicated a biased intein distribution towards replisome components. This, along with other related scientific publications that make use of available intein related biological data stress the importance for accurate records to maintain the integrity of research in this field. The primary question at hand is the cause behind the alterations of deletions of intein containing protein sequences hosted on NCBI and whether this issue can be resolved to ensure accurate intein data is displayed. For instance, a search for protein sequence ZP_01137525 on NCBI's protein database yields the result "Record removed", indicative of similar instances where sequences have been removed for various reasons. Other instances include these intein containing protein records being displayed with no mention of an intein in their sequence. To address this inaccurate intein data, a comprehensive data search would be conducted with sequences known to be previously accurate and the results would be examined to assess the extent of these inaccuracies. An example of an inaccuracy would be the removal of an intein region from a protein sequence that is still available on NCBI. Alternatively, other intein data hosted on the website would be collected and analyzed to identify missing intein regions. The analysis is focused on determining the patterns, magnitude and trends of these alterations serving as evidence of the once accurate intein containing protein records becoming deleted and misrepresented.

Materials & Methods

To gain a preliminary understanding of removed and inaccurate intein containing protein records, a dataset containing NCBI protein records was provided by Dr. Novikova. This dataset comprised intein containing protein records from three taxonomies that were previously accurately represented on NCBI and isolated from her 2015 paper. These records were manually searched in NCBI's protein database, and the search results were annotated based on whether the records were still actively hosted on the website. For records still displayed, the genetic sequences were inspected to determine if an intein was mentioned in the sequence. Following this initial assessment, additional protein records with a high likelihood of containing an intein were extracted and analyzed for inaccuracies.

The annotated data acquired from the record searches would be analyzed to find the proportion of records that were still active on the website as opposed to those that weren't. For active records their accuracy would be assessed by verifying the presence of an intein annotation in the sequence to find the proportion of accurate to inaccurate intein containing protein sequences.

Data mining for additional intein containing protein sequences involved inspecting taxonomies of pseudomonadota, actinmycetota and eukaryota, the same inspected in Dr. Novikova's 2015 paper. These searches targeted Replicative DNA Helicase proteins, as they are known to be one of the most common intein containing proteins. Protein sequence lengths ranged from 550 – 5000 for pseudomonadota and actinmycetota while the sequence length of actinmycetota was narrowed to 1000-5000 due to the abundance of records which made the download of data difficult. The data of these sequences were also searched for the mention of PRK07773, a profile detailing protein sequences containing inteins.

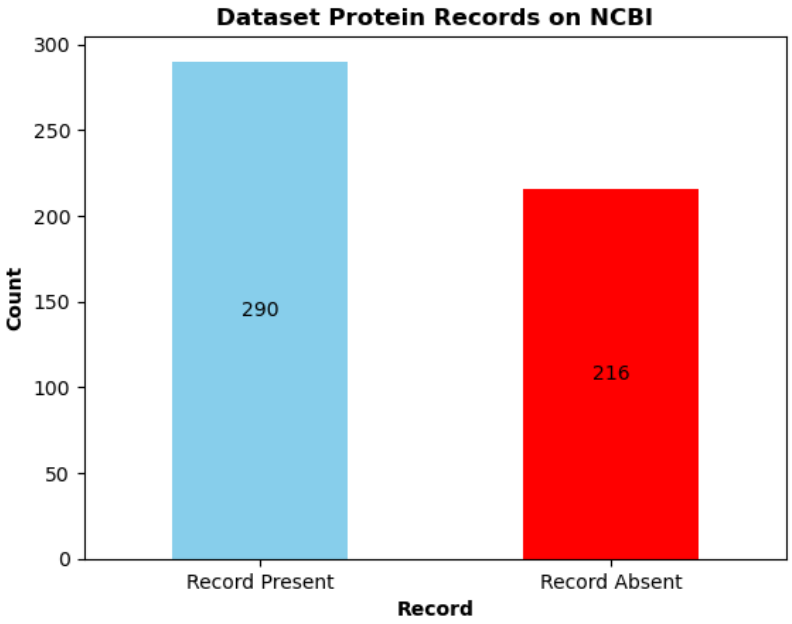
Data mined from NCBI was downloaded as a feature table into a text file for each taxonomy. Python was used to parse and isolate sequences and their data from these text files, creating separate data frames. These data frames were cleaned and joined to create a comprehensive dataset containing each taxonomy. The matplotlib library was utilized to produce charts detailing the proportions of active and inaccurate records.

Results

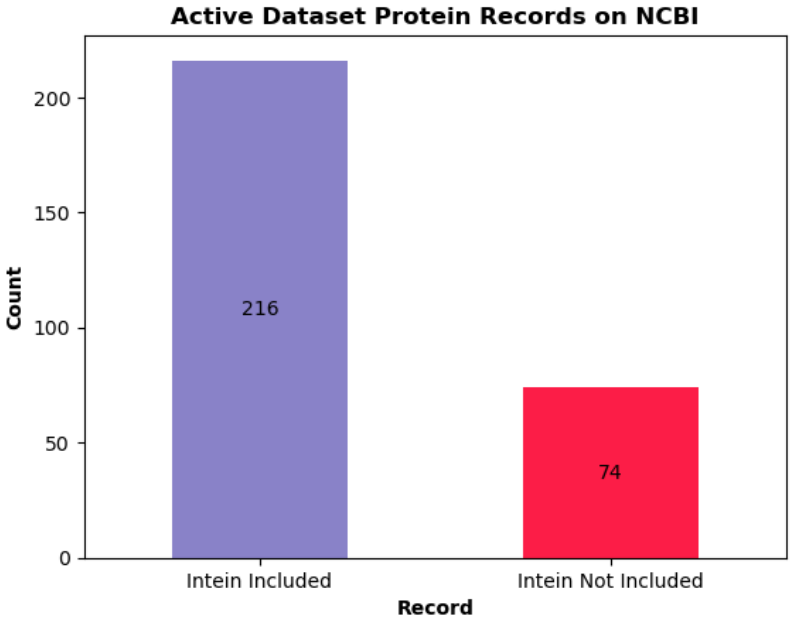
From the dataset provided by Dr. Novikova, a total of 506 unique intein containing protein sequences were manually search on NCBI. Of these 506 unique sequences, 57.3% (290 records) were active while 42.7% (216 records) were inactive, yielding the search result of "Record removed" for a variety of reasons. Of these 290 active records, 74.5% (216 records) were accurate meaning the protein sequences included the mention of an intein while 25.5% (74 records) were inaccurate, lacking the mention of an intein.

From the data mined on NCBI, a total of 1297 sequences contained the PRK07773 profile. Of these 1297 records that contained the PRK07773 profile, 68.3% (886 records) mentioned an intein in their sequence while 31.7% (411 records) lacked the mention of an intein, deeming these records inaccurate. The majority of sources for both accurate and inaccurate sequences were GenBank and RefSeq.

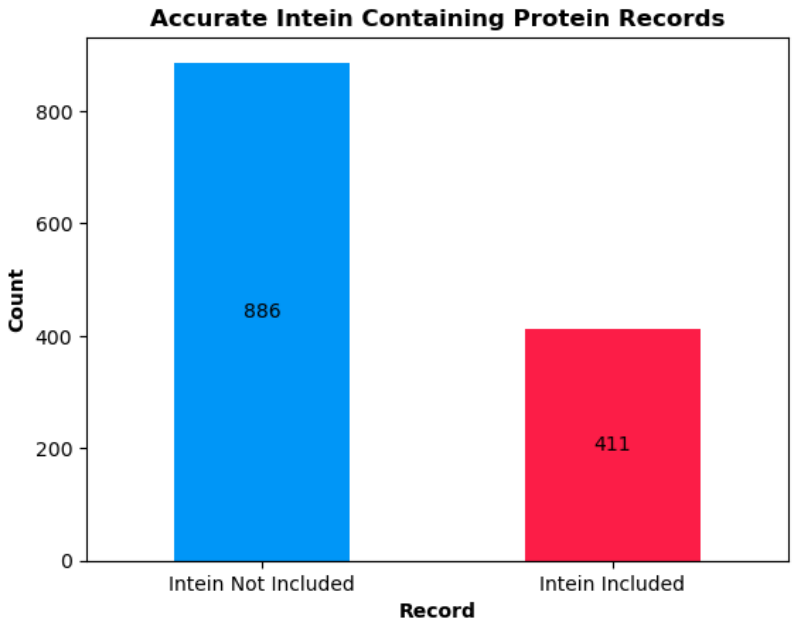
Tables and Graphs



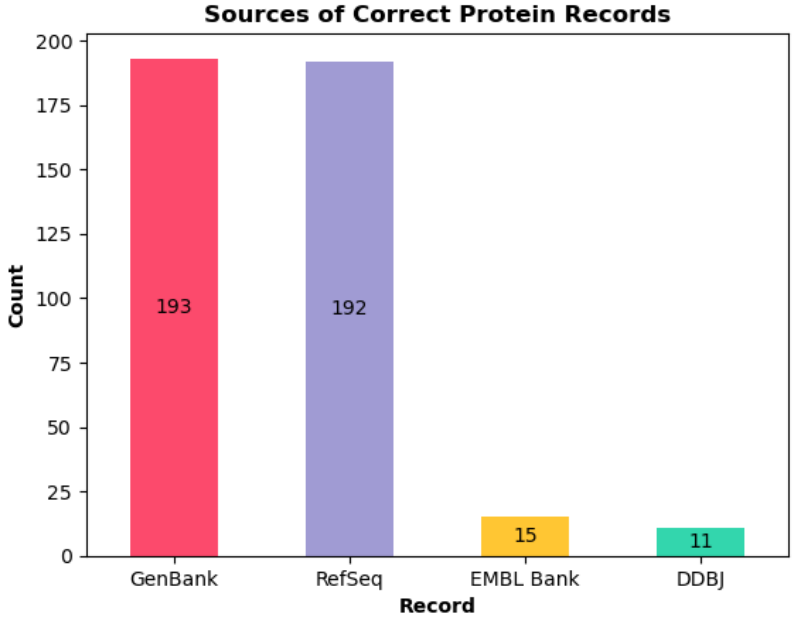
1A. Bar chart showing the count of present and absent intein containing protein records obtained from intein dataset on NCBI.



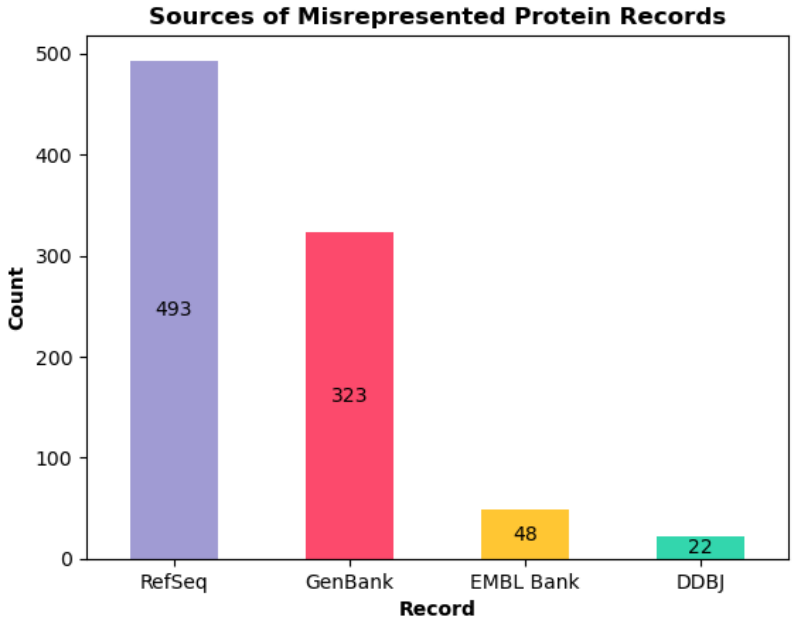
1B. Bar chart showing the count of accurate intein containing protein records obtained from the intein dataset when searched on NCBI. For an intein containing protein records to be accurate, the word “intein” must be included in the protein description or sequence.



2A. Bar chart showing the count of accurate intein containing protein records obtained from actinomycetota, eukararyota and pseudomonadata taxonomies with filters applied on NCBI. For an intein related protein record to be accurate, the word “intein” must be included in the protein sequence.



2B. Bar chart showing the count of sources for correctly represented intein containing protein records obtained from actinomycetota, eukaryota and pseudomonadata taxonomies mined from NCBI



2C. Bar chart showing the count of sources for misrepresented intein containing protein records obtained from actinomycetota, eukaryota and pseudomonadata taxonomies mined from NCBI.

Discussion

Upon following the research plan which involved inspecting and collecting intein data followed up with weekly meetings a comprehensive understanding of intein data was accomplished. Analysis of Dr. Novikova's dataset which entailed a subset of intein containing protein sequences from her 2015 paper, and mining intein records from NCBI revealed that a significant number of intein containing protein sequenced on NCBI had been removed or were inaccurate, lacking the mention of an intein as seen in Figures 1B and 2A. Upon analyzing the intein data, the majority source of records was from 'GenBank' and 'RefSeq' for both accurate and inaccurate records as seen in figures 2B and 2C. These findings did not lead us to draw a conclusion as to why intein containing protein records have been disrupted, but they did serve as evidence of missing and inaccurate intein data. The current state of innacurate intein data on NCBI poses a risk to the integrity of ongoing research where NCBI is used as a data source.

This evidence from this research and data will be submitted to NCBI so that the innacurate and missing biological data can be corrected and validated for scientific use. To explore the scope of misinterpreted biological data on NCBI, more intein containing protein sequences possibly from different taxonomies should be mined and inspected as this study was comprised only of pseudomonadota, actinmycetota and eukaryota. It is important that scientists and researchers involved in the emerging field of the study of inteins are well equipped with accurate data from NCBI to not be misguided by misrepresented data in their findings.

References

(1) Shah NH, Muir TW. Inteins: Nature's Gift to Protein Chemists. Chem Sci. 2014;5(1):446-461. doi: 10.1039/C3SC52951G. PMID: 24634716; PMCID: PMC3949740.

(2) "Our Mission - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/home/about/mission/. Accessed 28 Apr. 2024.

(3) Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI, Belfort M. Intein Clustering Suggests Functional Importance in Different Domains of Life. Mol Biol Evol. 2016 Mar;33(3):783-99. doi: 10.1093/molbev/msv271. Epub 2015 Nov 25. PMID: 26609079; PMCID: PMC4760082.