Protein Secondary Structural Analysis (PSSA)

Introduction

Protein secondary structure is the three-dimensional form of local segments of proteins. There are three common secondary structures in proteins, namely alpha helices, beta sheets, and turns. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its 3-D tertiary structure. The Protein Secondary Structural Analysis (PSSA) database provides the simulated 3-D structure of the proteins as well as the experimentally determined secondary structure information pulled out from the UniProt database. The users can observe the 3-D protein structure and proportion of the three main secondary structures within a protein sequence for a given accession code.

This project is a part of creating the database that focuses on extracting the secondary structure information from the UniProt database and measures the proportion of each secondary structure for a given accession code. This information includes the helix percentage, known partial helix percentage, beta strand percentage, known partial beta strand percentage, turn/random coil percentage, and known partial turn percentage in each protein. In UniProt database, 'Helix' is used to indicate the positions/ length of experimentally determined helical regions within the protein sequence. Likewise, 'Beta strand' is used to indicate the positions of experimentally determined beta strands and 'Turn' is the positions of experimentally determined hydrogen-bonded turns within the protein sequence.

Database Structure

The overall project of building the database consists of three main steps, 1) scraping the UniProt database to calculate the secondary structure information for each accession code, 2) creating 3-D protein structure model, and 3) integrating steps 1 and 2 so that they can be shown independently in the PSSA database. Python is used as the programming platform for all three steps. In this report, step 1 is described in detail. The program outputs a table consists of seven columns for each accession code and each row represents an accession code. The first column indicates all accession codes, and the rest of the six columns represent their corresponding structural information, which are in float format.

Methodology

For scraping data from UniProt, 'BeatifulSoup' package is used. Then 'Requests' package is used to send HTTP/1.1 requests to https://www.uniprot.org/uniprot/ website. The secondary structure table was extracted from the database, which contains the length of helix, beta strand, and turn. From this table, six parameters were calculated, which are: helix percentage (hp), known partial

helix percentage (kphp), beta strand percentage (bp), known partial beta strand percentage (kpbp), turn/ random coil percentage (rp), and known partial turn/ random coil percentage (kprp).

Each of the structure's proportion was calculated by dividing their total length by the sum of all parameter's length. Known partial of each parameter was calculated by dividing their total length by the length of PDB Structure known for that area. For instance, the total length of helix structure of P08670 was 194 and the sum of all structures' (helix, beta strand and turn) length was 203. Therefore, the helix percentage of the accession code P08670 was calculated (194/203)*100 or 95.5665%. The length of PDB Structure known for that area was 466. Hence, the known partial helix percentage (kphp) of P08670 was calculated (194/ 466)*100 or 41.6309%.

Finally, the output will be a CSV file containing all accession ID (e.g. P35579) and their corresponding parameters (hp, kphp, bp, kpbp, rp, kprp). Therefore, the CSV file will have seven columns for each accession code (Figure 1).

	A	B	C	D	E	F	G	H
		ID	bp	kpbp	hp	kphp	rp	kprp
0		P35579	23.33333	0.357143	76.66667	1.173469	0	0
1		P08670	1.970443	0.858369	95.5665	41.6309	2.463054	1.072961
2		P21333	75.77566	23.98942	21.59905	6.83793	2.625298	0.83113
3		Q9Y490	39.13043	1.416765	57.6087	2.085793	3.26087	0.118064
4		P60709	30.56769	18.66667	64.19214	39.2	5.240175	3.2
5		P14618	36.67546	26.17702	53.82586	38.41808	9.498681	6.779661
6		Q09666	NA	NA	NA	NA	NA	NA
7		P06733	25.45455	16.12903	69.09091	43.7788	5.454545	3.456221

Figure 1: Output CSV file.

Additionally, there will be a text file logging all errors while running this program. The 'logging' package was used for this purpose. The log file contains the error type, error name, and the date and time of when the error occurred (Figure 2).

```
2021-10-05 17:14:01 ERROR-list index out of range
2021-10-05 17:14:01 ERROR-[Errno 13] Permission denied: 'Python_parameters_in_dataframe.csv'
2021-10-05 17:14:49 ERROR-list index out of range
2021-10-05 17:14:55 ERROR-list index out of range
2021-10-05 17:18:39 ERROR-list index out of range
2021-10-05 17:18:46 ERROR-list index out of range
2021-10-05 17:19:31 ERROR-[Errno 2] No such file or directory: './HC1_BB1jj.csv'
2021-10-05 17:28:04 ERROR-[Errno 2] No such file or directory: './HC1_BB1jj.csv'
2021-10-05 17:28:21 ERROR-list index out of range
2021-10-05 17:28:26 ERROR-list index out of range
```

Figure 2: Text file that logs error.

Conclusion

The PSSA database allows users to visualize the 3-D structure model of a protein and their secondary structure information independently. This project is a part of creating the PSSA database, where the secondary structure information of the proteins is collected from the UniProt database and proportion of three main structures were calculated of the protein sequence.