Efficient Hybrid String Matching Algorithm for Protein Sequence Analysis

Protein sequence analysis is a fundamental task in bioinformatics, essential for identifying conserved motifs, predicting protein functions, and understanding evolutionary relationships.

To identify relationships many protein string matching algorithms are introduced such as Berry Ravindran, Index Based Search and many more. In all algorithms Berry Ravindran, Barry Moore and Index Based Search algorithms are considered to best. All this algorithms lack support of optimization so in propose work we are developing hybrid MAC algorithm by combining features Berry Ravindran and Index Based Search algorithm.

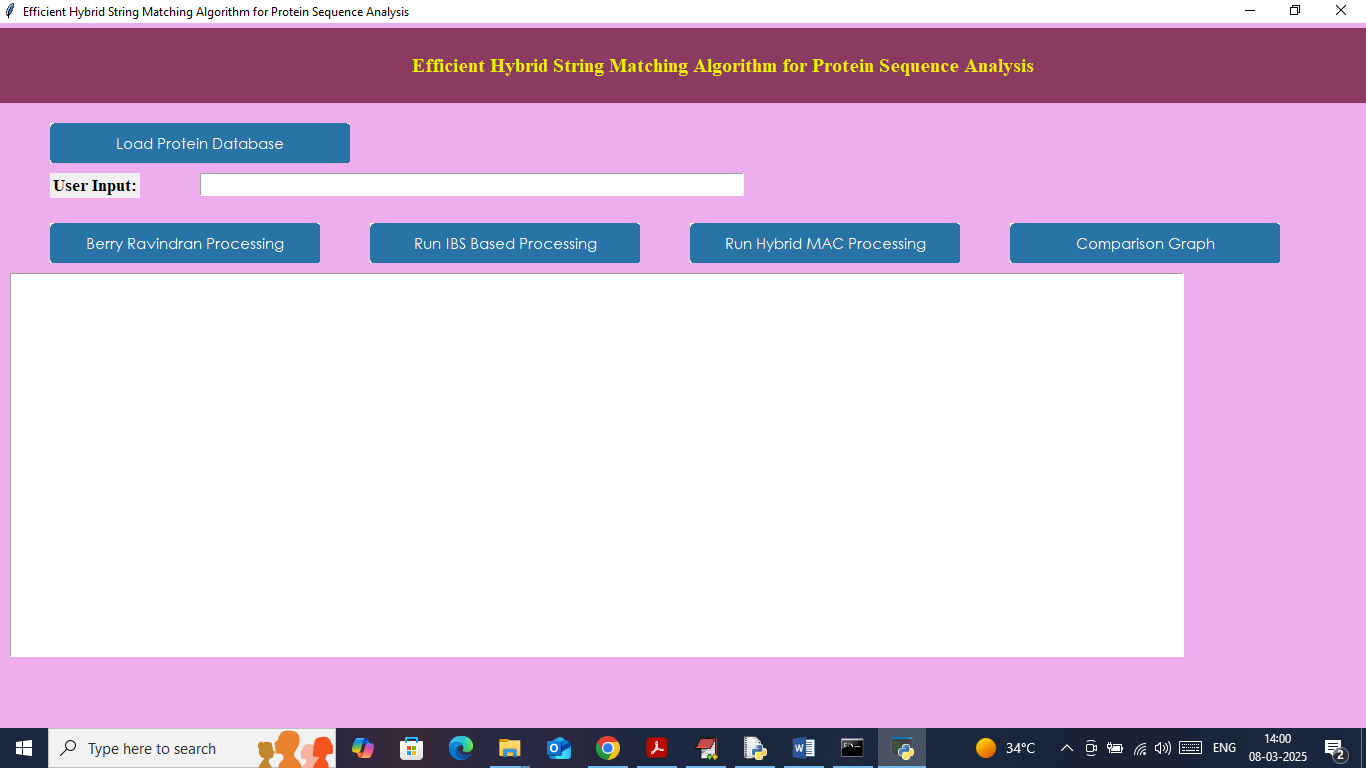
Berry Ravindran and Index Based Search does not support any heuristic technique to break loop to avoid unnecessary comparison by employing brute force technique. So we have enhance propose MAC (Minimum number of Attempts and Character Comparisons) algorithm by combining both algorithms and then apply Brute force technique to avoid unnecessary comparison when no pattern matching exists. By employing propose technique we can loop only minimum number of attempts for character comparison.

To implement this project we have designed following modules

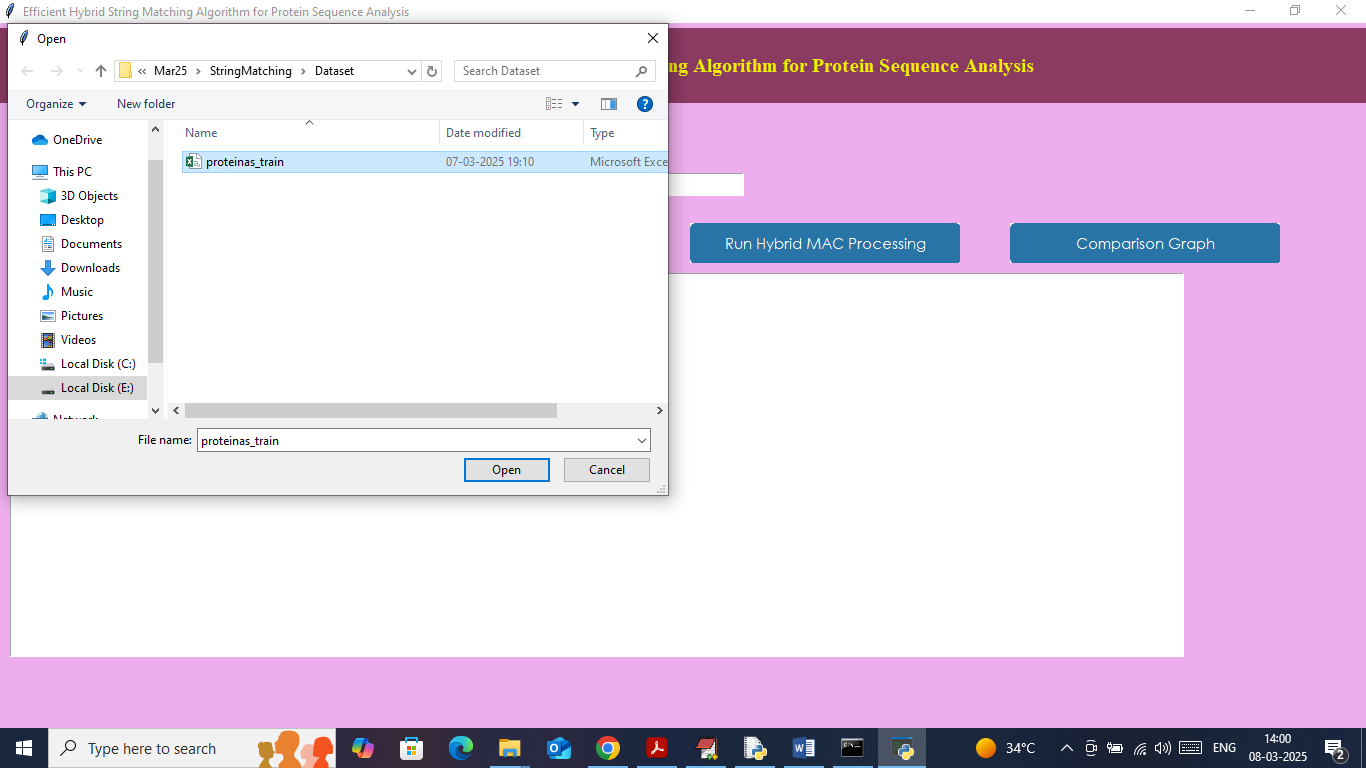
1. Load Protein Database: using this module user can load dataset to application
2. User Input: using this user can enter some pattern text which is used for string matching from protein sequences text
3. Berry Ravindran Processing: user input will be applied on loaded dataset to match string using Berry algorithm and then calculate computation cost
4. Run IBS Based Processing: user input will be applied on loaded dataset to match string using IBS algorithm and then calculate computation cost
5. Run Hybrid MAC Processing: user input will be applied on loaded dataset to match string using Hybrid MAC algorithm and then calculate computation cost
6. Output: each algorithm will be displayed output along with protein class name and index of string found in sequences
7. Comparison Graph: will plot comparison graph between all algorithms execution time.

SCREN SHOTS

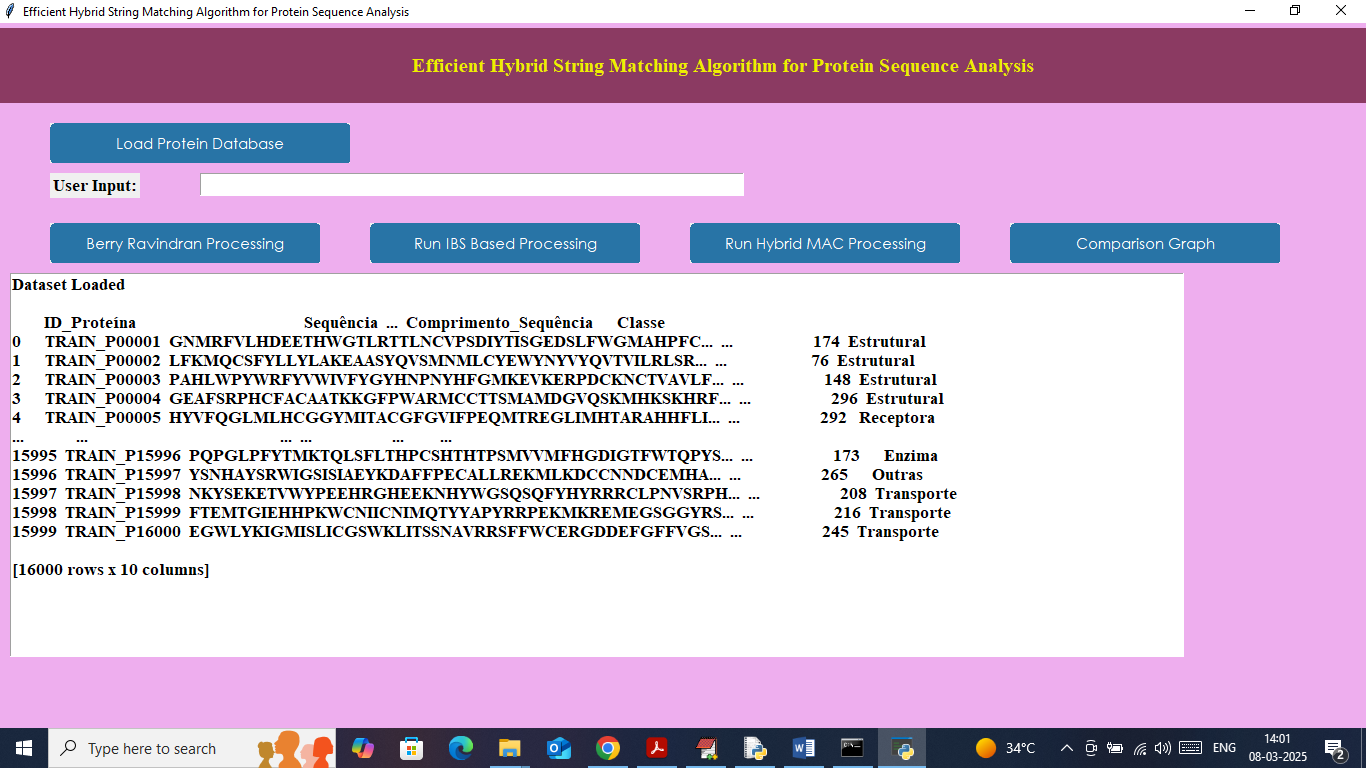
To run project double click on ‘run.bat’ file to get below page



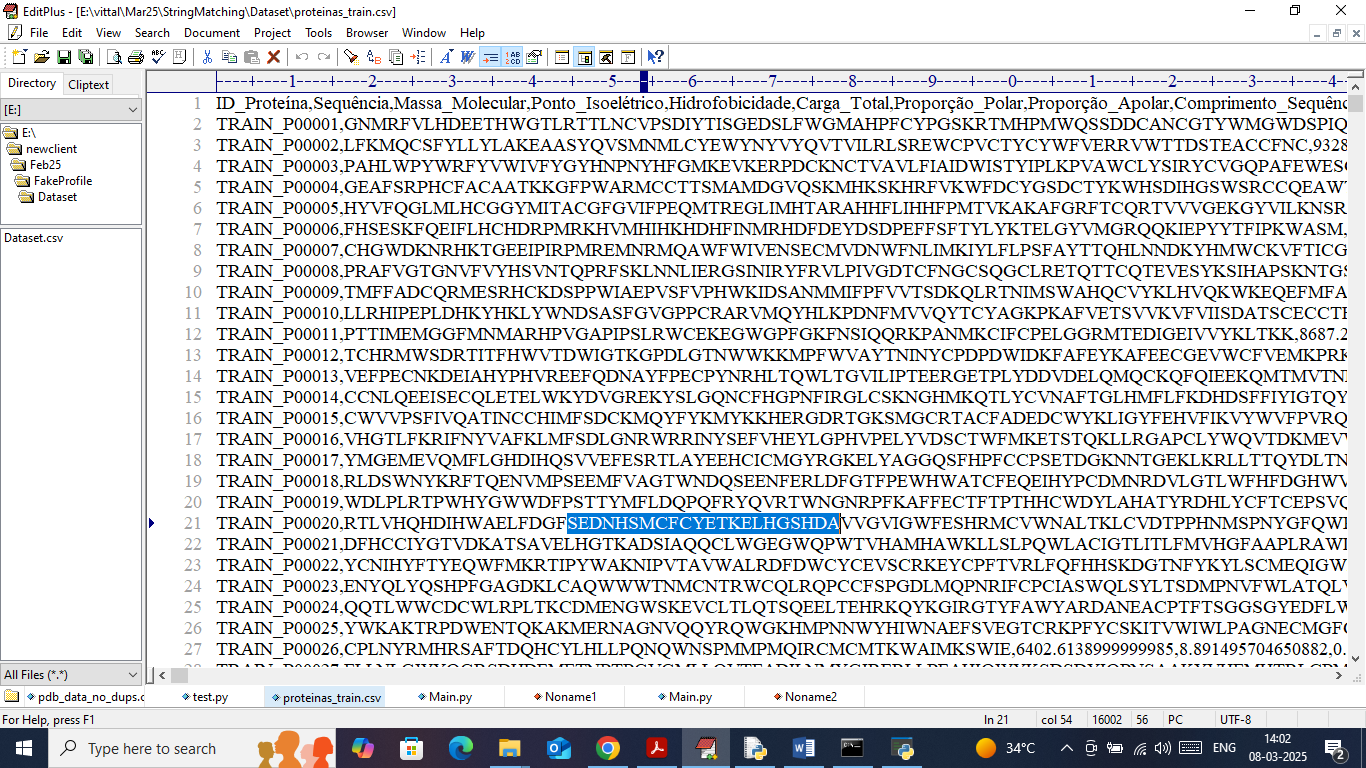
In above screen click on ‘Load Protein Database’ link to load dataset like below page



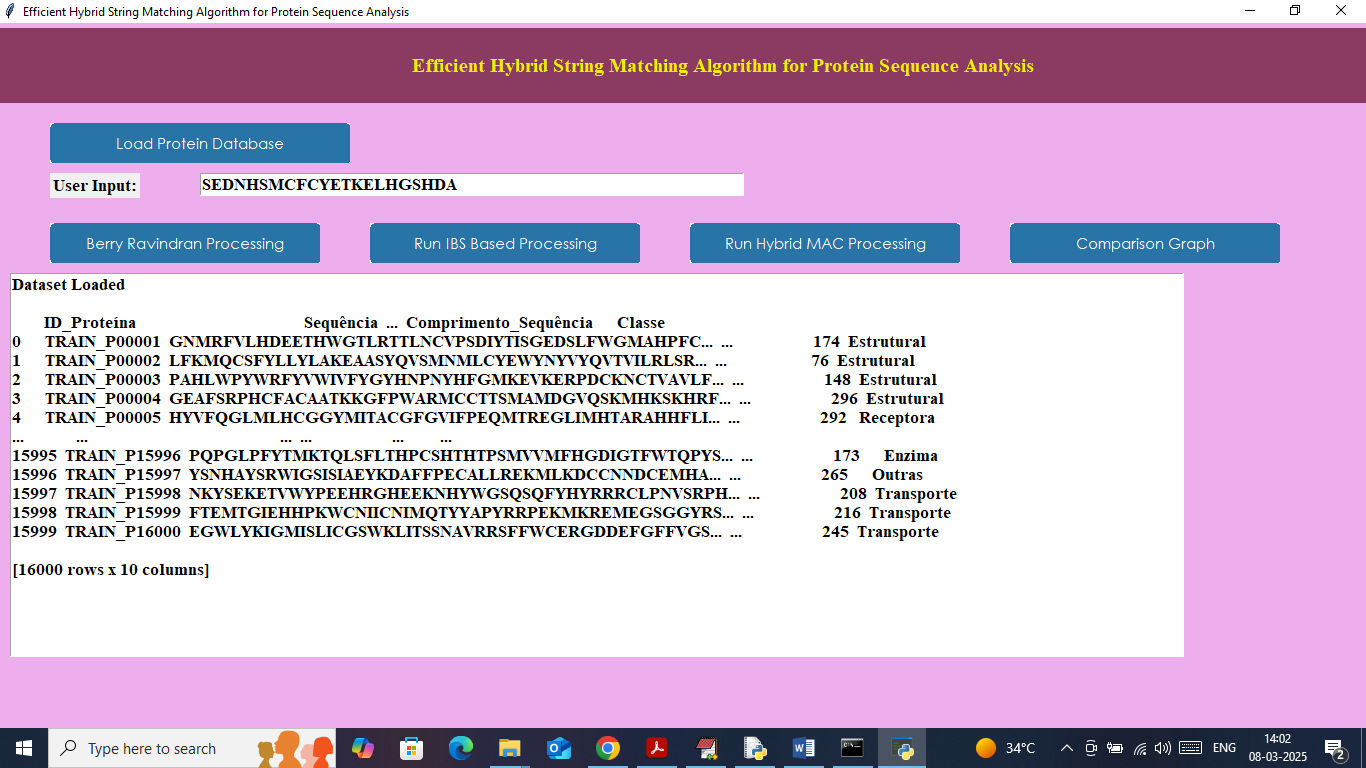
In above screen selecting and uploading ‘Protein Dataset’ and then click on ‘Open’ button to load dataset and then will get below page



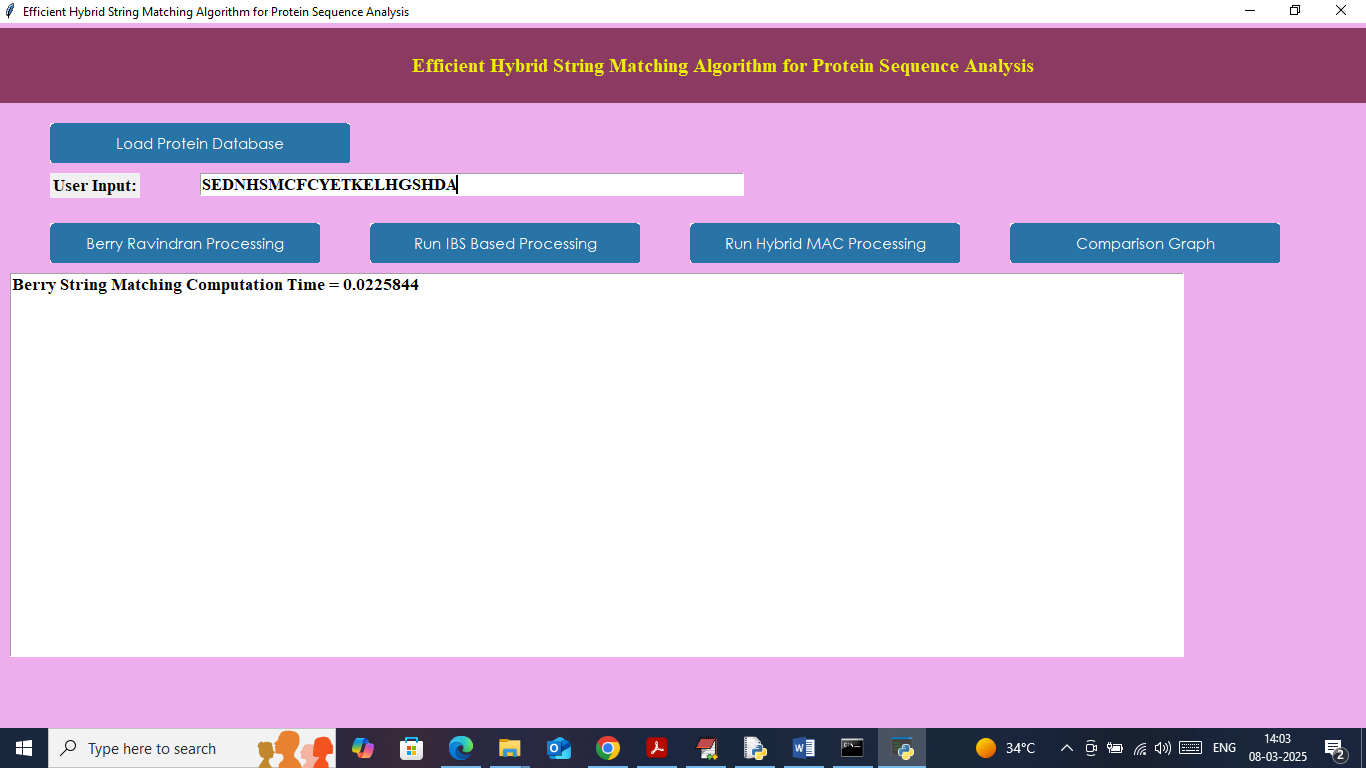
In above screen dataset loaded and now enter some pattern from protein sequences to search or copy and paste in text field



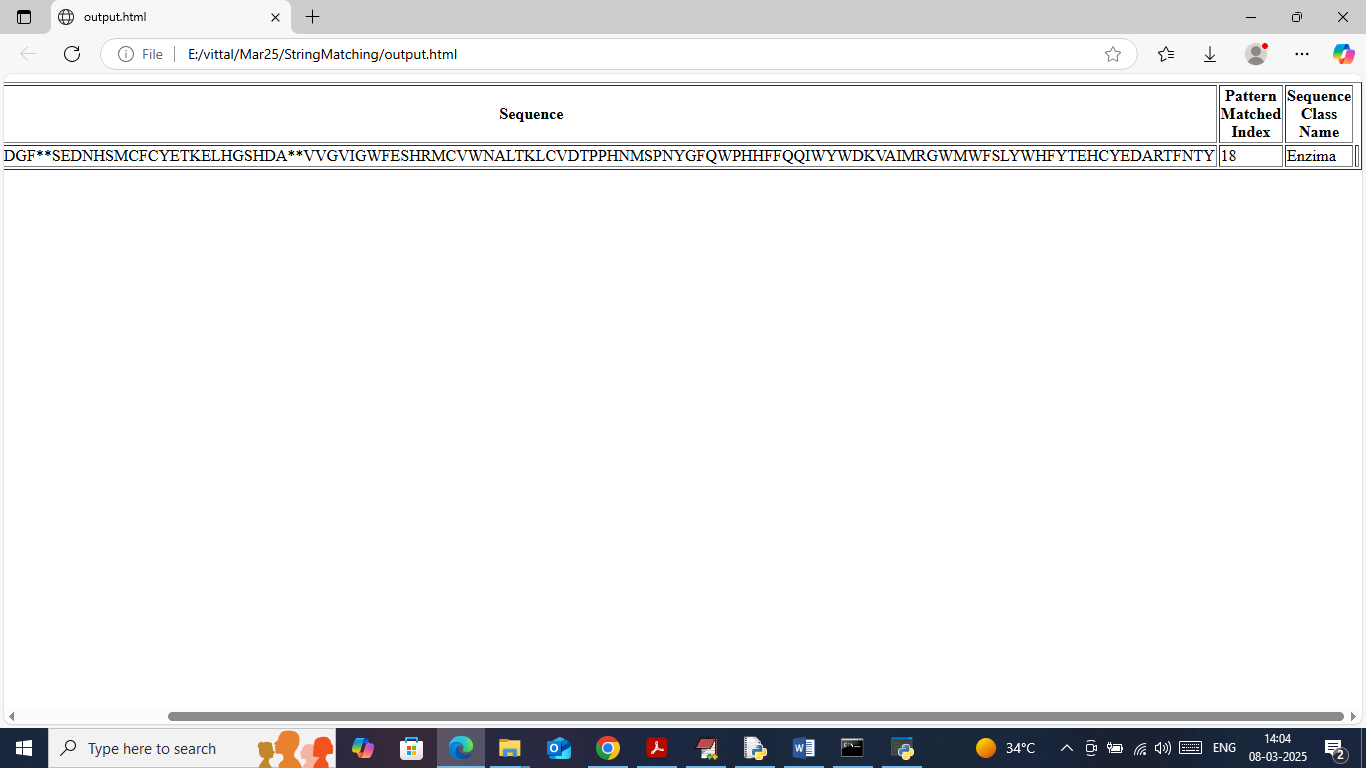
From above screen I am copying some text from sequences and then paste in below page



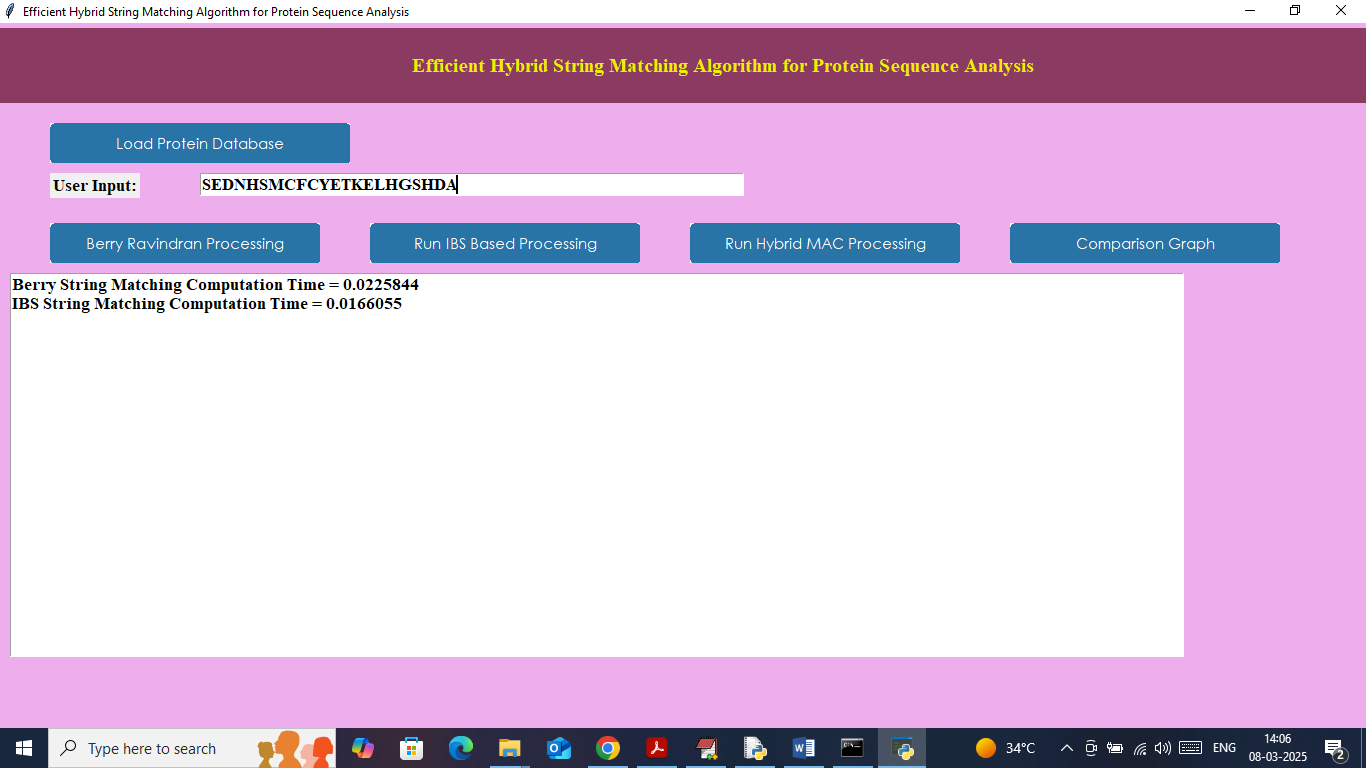
In above screen after entering some pattern click on ‘Berry Ravindran Processing’ button to match using Berry algorithm and then will get below page



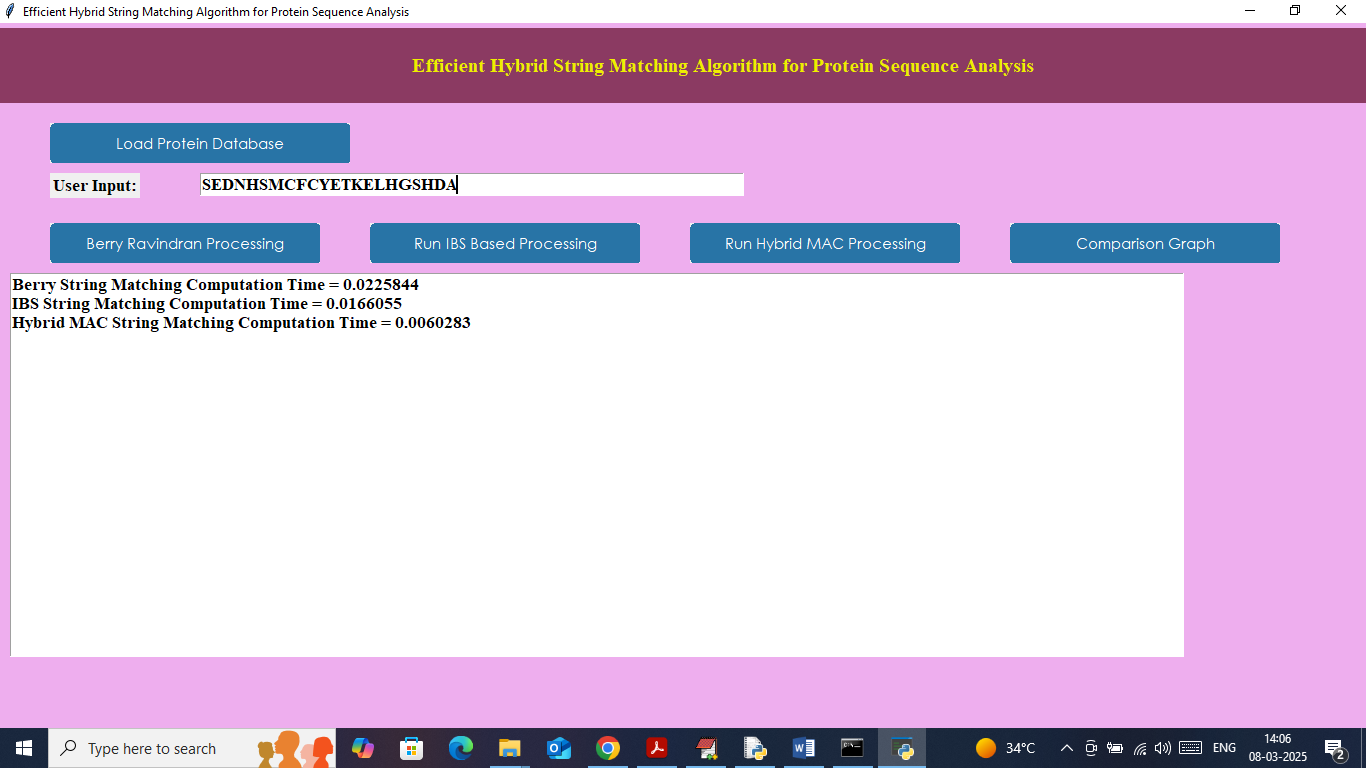
In above screen Berry took 0.022 seconds to search pattern and then will get below page



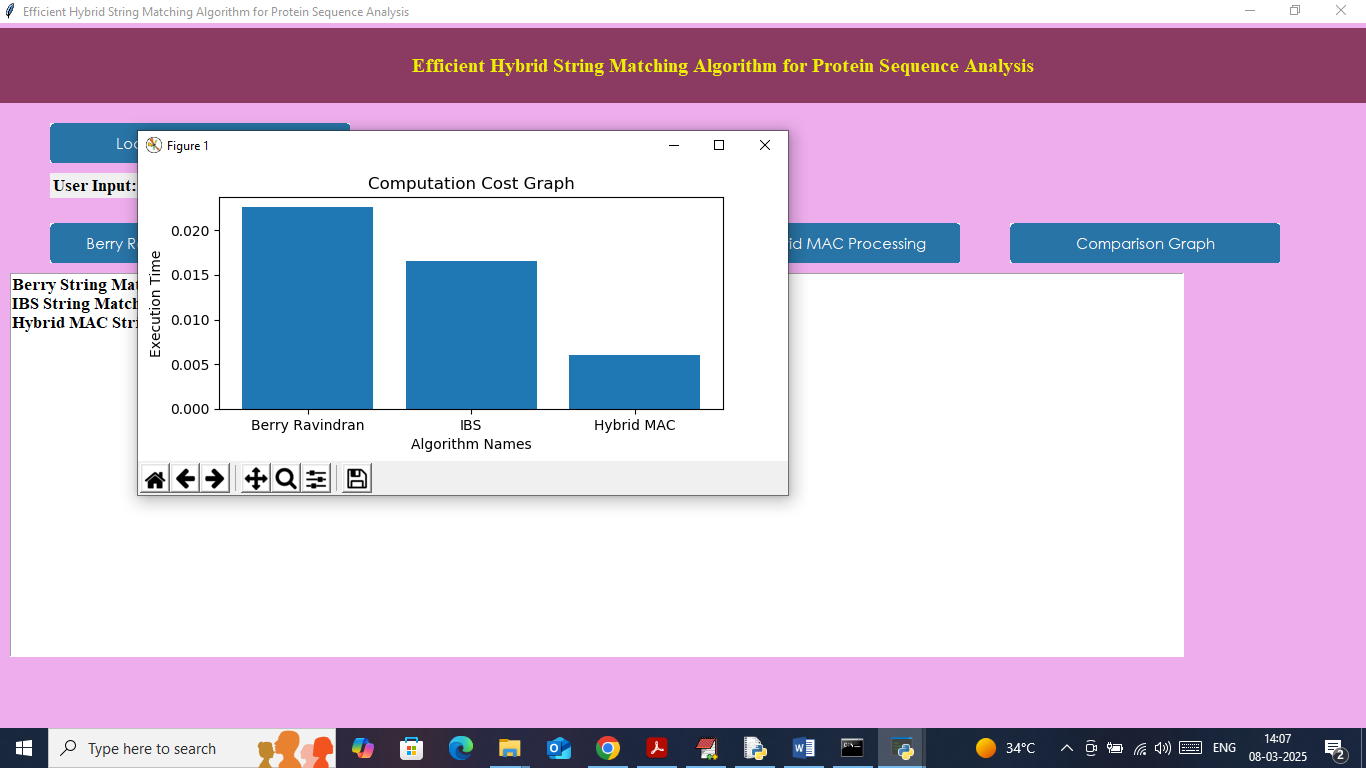
In above screen in first column we can see the lengthy Sequence and then in between that sequence we can see ‘\*\*’ symbol which indicates where actually search pattern string found and in second column can INDEX no from which pattern occur and in 3rd column can class name of that protein. Same above output will get for all algorithm. In below screen will show computation time of next two algorithms



In above screen IBS took 0.016 seconds



In above screen propose Hybrid MAC took 0.006 seconds and now click on ‘Comparison Graph’ button to get below page



In above graph x-axis represents algorithm names and y-axis represents computation time or cost and then in all algorithms propose MAC took less execution time cost.

Similarly by following above screens you can search for any pattern text from dataset sequences.