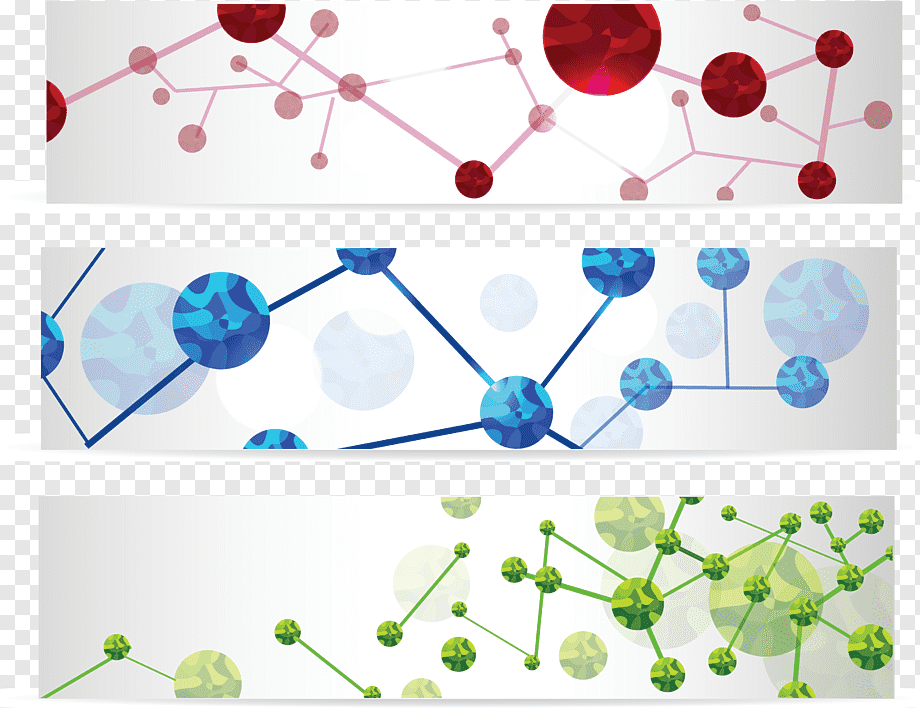
****

Software Engineering Department

Ort Braude College

**Capstone Project – Phase** B**, 61998**

Hidden homology detection via the protein connectivity network analysis \_ 23-2-R-13\_



Student name: Fadi Harb.  
 ID: 316239151.

Supervisor: Dr. Zakharia Frenkel   
GIT link : https://github.com/fadih1996/Hidden-homology-detection-via-the-protein-connectivity-network-analysis\_FinalProject

***Abstract:***

The field of homology detection plays a crucial role in understanding the evolutionary relationships between proteins and peptides.

MDPs (mitochondria-derived peptides) are small bioactive peptides encoded by mitochondrial DNA and are involved in protecting cells from stress.

While their functions have been extensively studied in humans, there is limited knowledge about MDPs in other vertebrate species, and their prevalence across different organisms is unknown. The short length and hidden location of MDPs make them difficult to study using existing tools and databases.

The proposed approach utilizes innovative techniques, along with the reconstruction of a protein connectivity network, to overcome these limitations and improve the identification and analysis of hidden proteins.

**Keywords:** protein sequence, protein structure prediction, electrical network, protein annotation, mitochondria-derived peptides (MDP).

***Introduction:***

“One of the tasks of computational biology is in protein structure prediction based on known 3D-structures of related proteins (homology modeling) (Kopp and Schweder, 2004).”[4]

For both basic research and drug development, data gathered from recently developed methods for structural annotation of proteins, such as efforts to predict protein subcellular or peptides localization, membrane protein types and regions, enzyme functional classes, and even signal peptides, is extremely helpful.

A huge number of protein sequences presents an important challenge for protein homology modeling.

In some cases, proteins that are physically similar have almost no sequence similarity, so it puts us in front of very large challenge for distant homologs with divergent sequences.

In this study, we investigate whether nature maintains a record of the protein's evolutionary history in the context of sequence space.

If thus, how far does this path extend? How continuous is the protein sequence space, in other words? And what is the structure of the sequence space over the long term?

The proposed method uses a network-based approach using a graph to examine the relationships between proteins and identify patterns that indicate hidden homology, meaning proteins that share a common ancestry but have diverged in sequence.

We will use the graph (a bunch of connected edges) to show how the sequences are related.

By counting how many different paths there are between the edges in the graph, then we could recognize the similar the sequences were.

This gives us a way to measure how similar sequences are and how sure we can be about our prediction.

So, this new method could be useful for comparing sequences in the future by analyzing the proteins or peptides sequences to their functionality, we can uncover these hidden homologies, which can contribute to our understanding of protein’s functions and their evolution.  
So, in this project I will focus on a tool called protein connectivity network (PCN) to detect hidden similarities between proteins.

The main objective is to develop a computational framework that leverages the structural and functional relationships between proteins and peptides, rather than relying solely on sequence similarity.

By constructing and analyzing connectivity networks, which capture the intricate interactions and dependencies between amino acids, it is possible to uncover hidden homologies that have eluded traditional sequence-based methods.

The constructed tool will be applied to the detection of the MDPs (mitochondria-derived peptides).

This information based on the references that mentioned above in [X], they are: [1],[2][3],[4],[5],[7]

*Background and related work*

**2.1 Protein – Peptides**, **illustration images**

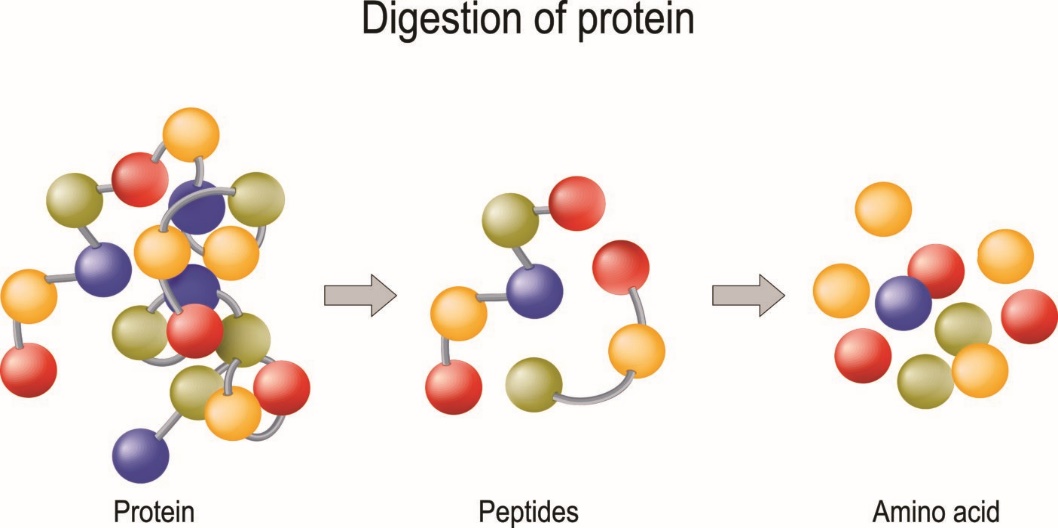
****

Fig 1

Proteins and peptides are necessary macromolecules that include amino acids which have a variety of functions in the body, including signaling within cells, structure, and enzymatic activity.

Peptides are short chains of amino acids which are held together by peptide bonds.

they typically consist of include less than 50 amino acids, despite this definition could be change, they are smaller chains of amino acids, while proteins are bigger and more complicated molecules, the peptides are serving as building blocks or fragments of proteins.

The body uses proteins for  many important reasons.

Proteins play an important position in the development and maintenance of all body parts, including bone, muscle, blood cells, skin, and hair.

They are also the main building block of enzymes, proteins that aid in a variety of bodily chemical processes, including digestion.

The production of hormones including insulin, thyroid hormones, estrogen, and testosterone also depends on proteins.

They also can be found in many foods, and consuming too little or too much of them can be harmful to the body.

****

Fig 2

**2.2 Exploring Mitochondrial-Derived Peptides (MDPs):**

**Unveiling Hidden Regulators of Health and Disease**

Mitochondrial-Derived Peptides (MDPs) are a recently discovered class of peptides encoded by small open reading frames (sORFs) within known mitochondrial genes. The first identified MDP, Humanin, demonstrated neuroprotective properties and holds great therapeutic potential for neurodegenerative diseases.

Alongside Humanin, other well-known MDPs include MOTS-c and SHLP, which act as cyto- and neuroprotective agents while regulating energy metabolism.

MDPs are implicated in various senescence and aging-associated diseases, chronic inflammation diseases, cancer, neurodegenerative diseases, and cardiovascular diseases. However, the study of MDPs has been challenging due to their length and the limitations of available research tools, hindering comprehensive analysis.

Understanding the evolution and prevalence of MDPs, particularly the pseudogenisation process of Humanin, not only sheds light on their evolutionary significance but also holds substantial implications for biomedical and pharmaceutical research.

Moreover, investigating MDPs may lead to the identification of novel functions and evolutionary insights across different species, expanding our understanding beyond known Vertebrata species.

**2.3 Implications of MDPs in Disease:**

MDPs (Mitochondrial-Derived Peptides) have implications in various diseases, including neurodegenerative diseases, cardiovascular diseases, metabolic disorders, age-related diseases, chronic inflammation diseases, and cancer.

They exhibit neuroprotective effects in neurodegenerative diseases and cardio-protective effects in cardiovascular diseases.

MDPs are also involved in regulating energy metabolism and have been linked to metabolic disorders like obesity and diabetes.

They play a role in age-related diseases and possess anti-aging properties.

Additionally, MDPs have anti-inflammatory effects and may impact cancer cell growth and apoptosis.

Understanding the roles of MDPs in these diseases can lead to potential therapeutic interventions, diagnostic markers, and a deeper understanding of disease mechanisms.

**2.4 Protein Connectivity Network (PCN)**

Protein Connectivity network is exhibited by a graph  to show the framework. Protein or peptide sequence segments are represented by the graph's nodes.

Each node in the network can have one or more edges connecting it to other nodes.

The network approach of the current invention can be set up to determine a protein's function, relationship to other proteins in terms of structure, or annotation of a specific sequence of amino acids or protein.

The structures and activities of the protein are mostly unaffected by evolutionary changes since they mostly occur at the sequence level.

As a result, sequences with comparable structures or functions might differ greatly.

However, we can determine their relationship by identifying the intermediary sequences that connect them by using PCN.

The most significant characteristic of this network is that the associated segments of protein retain their structural and functional characteristics when the protein sequences gradually shift from one to a completely other one.

In our study, a network is used to represent the similarity of short sequences made up of 20 amino acid segments as nodes connected to edges, with the weight of similarity determined by the sequence kinship.

Only if the similarity between correspondent fragments is greater than or equal to 60% does the network include an edge.

The size of the PCN might vary depending on the number of proteins used to build it; for instance, a network made up of 320,000 proteins will result in 92,000,000 nodes of 20 amino acid fragments. [9]

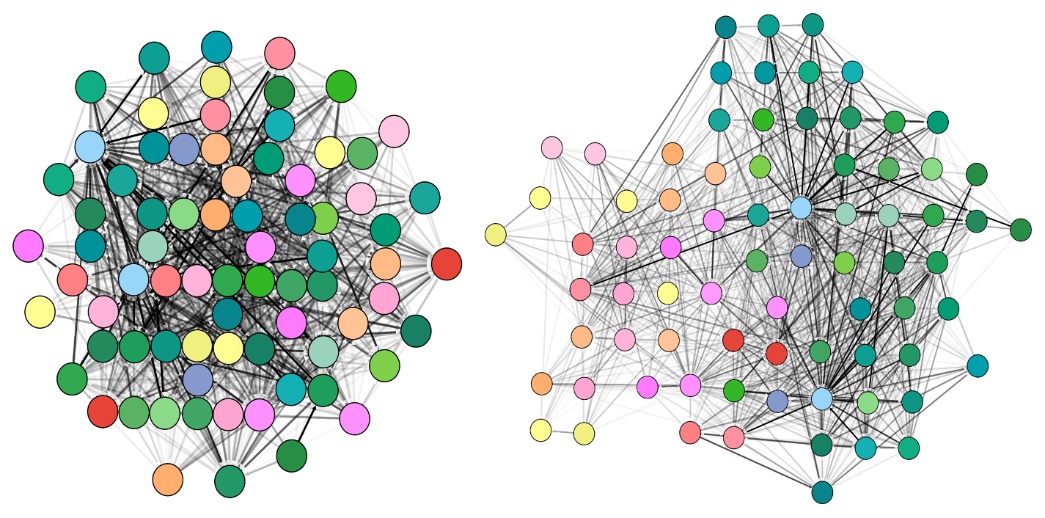


Fig 3

**2.6 UniProt DataBase**

According to this study [10], there were two separate protein databases called:

Swiss-Prot + TrEMBL and the other PIR-PSD.

They had different coverage and priorities for annotating protein sequences.

In 2002, these two groups joined together with the aim of creating a unified and

high-quality protein database called UniProt.

UniProt is maintained by the Swiss Institute of Bioinformatics (SIB), European Bioinformatics Institute (EBI), Georgetown University Medical Center, and National Biomedical Research Foundation.

The main goal of UniProt is to support biological research by providing a reliable and comprehensive database of protein sequences.

This database is well-classified, accurately annotated, and contains a wealth of information.

It includes cross-references to other relevant resources and offers user-friendly interfaces for easy access by the scientific community.

UniProt is built upon the expertise and efforts of the consortium members accumulated over many years.

UniProt consists of three layers of databases:

UniParc (UniProt Archive): This database stores a complete collection of publicly available protein sequences. It aims to provide a stable and non-redundant set of protein sequences.

UniProt: The UniProt Knowledgebase serves as the central database of protein sequences.

It ensures accurate, consistent, and rich annotation of protein sequences, including information about their functions.

UniRef (UniProt NREF databases): The UniRef databases are derived from the UniProt knowledgebase and provide non-redundant data collections. They offer different resolutions to cover the entire sequence space effectively.

**2.7 DISCOVERING Hidden Homology with (PCN)**

Protein structure prediction, function annotation, and evolutionary analysis all depend on the ability to detect homology, a key problem in computational biology.

Traditional techniques of homology detection rely on approaches based on sequence similarity, which can miss distant homologs with low sequence similarity.

This limitation encourages the creation of novel techniques that can find hidden homology based on various kinds of data.

In recent years, Protein connection network analysis has been a promising method for finding hidden homology in recent years.

According to this method, proteins can be seen as nodes in a network with edges representing interactions between them, such as physical contact or co-expression patterns.

Even if proteins have little in common terms of sequence similarity, it is still possible to identify proteins that are functionally and structurally connected by looking at the network's topology.

**2.8 Algorithms and computational system**

Bioinformatics protein homology detection relies on a range of algorithms and computational systems to compare protein sequences, identify similarities, and predict evolutionary relationships.

These methods utilize various techniques to achieve accurate and efficient homology detection.

In this study, I want to focus on these algorithms within a specific calculation system.

**Key feature TATA-BOX**

The "TATA box" is a DNA sequence that is crucial in the process of gene transcription in many eukaryotic organisms. While it is not a feature of proteins directly, it plays a significant role in the synthesis of proteins by influencing the transcription phase of gene expression.

**Key feature NAPD**

The acronym "NAPD" could be associated with a typo or confusion with the more commonly referenced molecule "NADP+" in the context of biochemistry and molecular biology. "NADP+" stands for Nicotinamide Adenine Dinucleotide Phosphate, which is crucial in cellular metabolism and not a protein itself but often tightly interacts with proteins.

**Electrical network model**

The presence of the sequence fragments in the same network may serve as a new criterion of the sequence relatedness. To describe this relatedness a "Structural Flow" will be introduced. The links between the nodes mean sequence similarity of corresponding sequence fragments and may "transmit" with some probability the biological properties, such as 3D structure organization and/or functions, from one sequence fragment to another. The number of independent paths between two nodes through the network (or, in other words, maximal flow) reflects structural similarity between corresponding protein fragments.

To consider also the length of the paths, more complicated models will be applied. The network in the sequence space can be considered as an electrical network. The parameter reflecting relatedness of fragments will be "conductivity" (inverse resistance) between corresponding nodes. According to the laws of resistance:

for series connection of and , and for parallel connection.

Thus, this parameter considers both the number of independent pathways and their length, allowing the simple introduction of specific properties of connections.

The calculation of the resistance between two points of the network will be carried out as follows, with example in fig 3.

The first and the second points (A and D) are considered possessing potentials '0' and '1' respectively. For the network of N (4) nodes and K (5) edges (connections) a linear system of K+N-2 variables will be composed: K – currents through the K edges, and N-2 – potentials at remaining N-2 nodes.

N-2 equations are composed following Kirchhoff's law, according to which the sum of all currents at every point (excluding two mentioned above) is equal to zero.

iAB = iBC + iBD

iAC = iCD - iBC

Other K equations are obtained from the Ohm's law Uj–Ui = Uij\*Rij, where Uj and Ui are potentials at nodes i and j, and Iij and Rij are current and resistance between these nodes, respectively.

UA = 0

UD = 1

iAB = R1\*(UB-UA)

iAC = R2\*(UC-UA)

iBC = R3\*(UC-UB)

iBD = R4\*(UD-UB)

iCD = R5\*(UD-UC)

From this system one can find a total current i, and total resistance 1/i (since total difference of potentials is equal to '1'). The values of parameters Rij in a simple case can be taken to be equal to '1', but for a more realistic model their values should depend on similarity between the nodes 'i' and 'j', as well as on other parameters. (Calculation of these Rij is the main purpose of our project) Using this approach, a new score for sequence relatedness will be developed.

The score will be obtained from analysis of calculated resistances between the nodes with known properties and will reflect the probability of these properties to be similar for both nodes.

A picture containing line, diagram

Description automatically generated

Fig 8

Fig 3: Example of calculation for the resistance between two points A and D. The potentials (U) of node A and D are given and

This model referred to reference[9]

**Related work**

**6.1 NetworkBLAST**

Protein connection network analysis has been used in several research to investigate homology detection.

For instance, a technique called NetworkBLAST, which uses a network matching algorithm to identify homologous proteins based on their connectivity patterns, this method was proposed by Le et al.

(2017). It was demonstrated that NetworkBLAST works well in identifying distant homologs than standard sequence similarity-based techniques.

It is a platform for locating protein complexes in interaction between proteins networks offered by the NetworkBLAST web service.

It could examine one network or two networks from several species. In the latter situation, it generates a list of potential complexes that are supposedly conserved throughout evolution in the two networks. [17]

WORK PLAN

My Work Plan for this project is about using K\_mismatch Algorithm, Graph Analysis Algorithm - GraphAligner, and Protein Database:-

1. **Exploring and recognizing:**

\* Understand the concept of hidden homology and its significance in protein analysis.

\* Study the PCN and its application in identifying hidden homologies.

\* Learn about the K\_mismatch algorithm and its use in sequence comparison and alignment.

\* Exploring the GraphAligner algorithm and its capabilities in aligning sequences to a graph representation.

\* Gaining knowledge about protein databases and their role in providing reference sequences for analysis.

1. **Definition of the Research Objective:**

\* Clearly define the goal of your hidden homology detection project.

\* Determining the specific criteria for identifying hidden homologies in protein or peptides sequences.

\* Set performance metrics for evaluating the effectiveness of the algorithms used.

1. **Preparing the Protein Database:**

\* Select a suitable protein or peptides database that contains diverse and representative sequences.

\* Download and preprocess the database to ensure consistency and compatibility with the analysis algorithms.

\* Organize the database in a format compatible with the K\_mismatch and GraphAligner algorithms.

1. **Implementation of the K\_mismatch Algorithm:**

\* Developing an implementation of the K\_mismatch algorithm.

\* Configuring the algorithm to handle the protein or peptides sequences from the database.

\* Running the algorithm to compare the query sequences against the protein database, considering various K\_mismatch thresholds.

1. **Applying Graph Analysis - GraphAligner:**

\* Utilizing the GraphAligner algorithm for aligning protein sequences to a graph representation.

\* Converting the PCN into a suitable graph structure.

\* Align the query sequences to the graph representation using GraphAligner, considering various alignment parameters.

1. **Analyzing Results and Identify Hidden Homologies:**

\* Collecting the output from the K\_mismatch algorithm and GraphAligner for each query sequence.

\* Evaluating the alignments and identify potential hidden homologies based on predefined criteria.

\* Analyzing the aligned regions and patterns to determine the significance and functional implications of the detected hidden homologies.

1. **Validate and Refine the Results:**

\* Validating the detected hidden homologies through comparison with existing literature or experimental data.

\* Iterate and refine the analysis parameters, such as K\_mismatch thresholds and alignment parameters, to improve the accuracy of hidden homology detection.

1. **Results report:**

\* Summarizing the findings and conclusions from the hidden homology detection analysis.

\* Providing a detailed description of the identified hidden homologies and their potential implications.

\* Presenting visualizations or diagrams to aid in the interpretation and understanding of the results.

\* Discussion of the limitations and challenges encountered during the analysis process.

\* Incorporate feedback and suggestions from the advisor.

\* Ensure the report is well-organized and formatted according to the designated guidelines.

Software Engineering: Detailed Implementation Steps

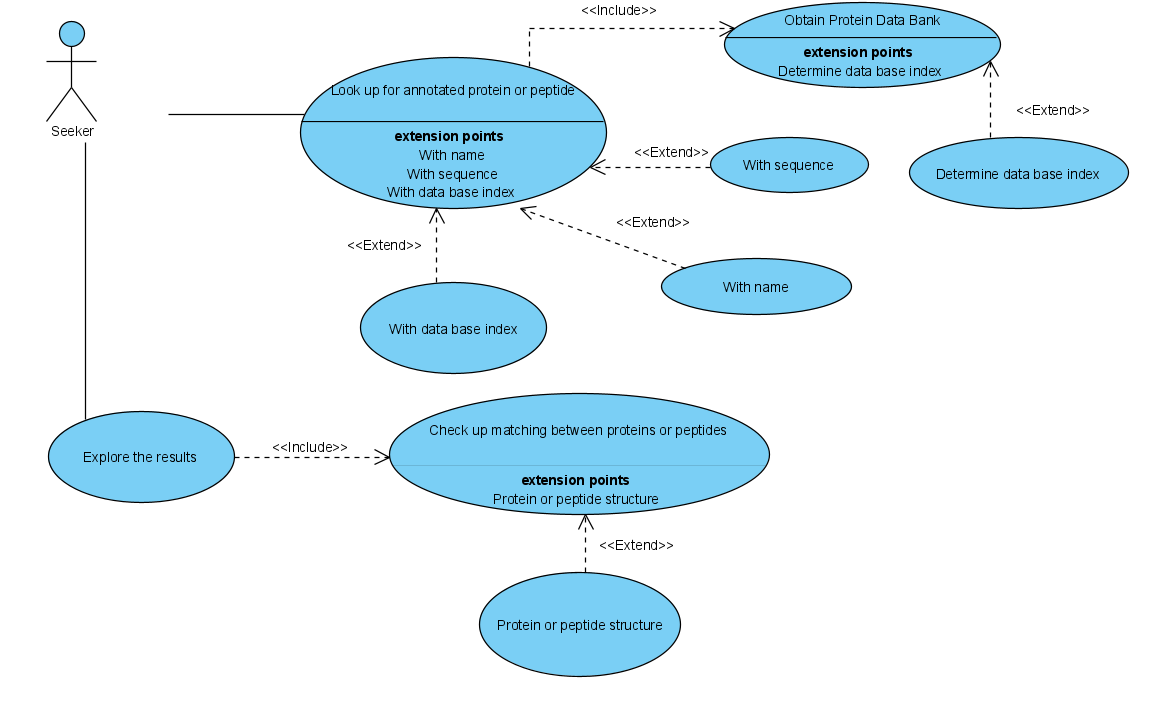
The following actions should be taken to create a weighted relatedness protein network according to [16]:

1. Obtaining a protein network.
2. Generating training data.
3. Generating a weighting function derived from training data values.
4. Applying said weighting function to a protein network.

**8.3. System steps - flow chart**

**//**

**8.4. Use Case diagram.**

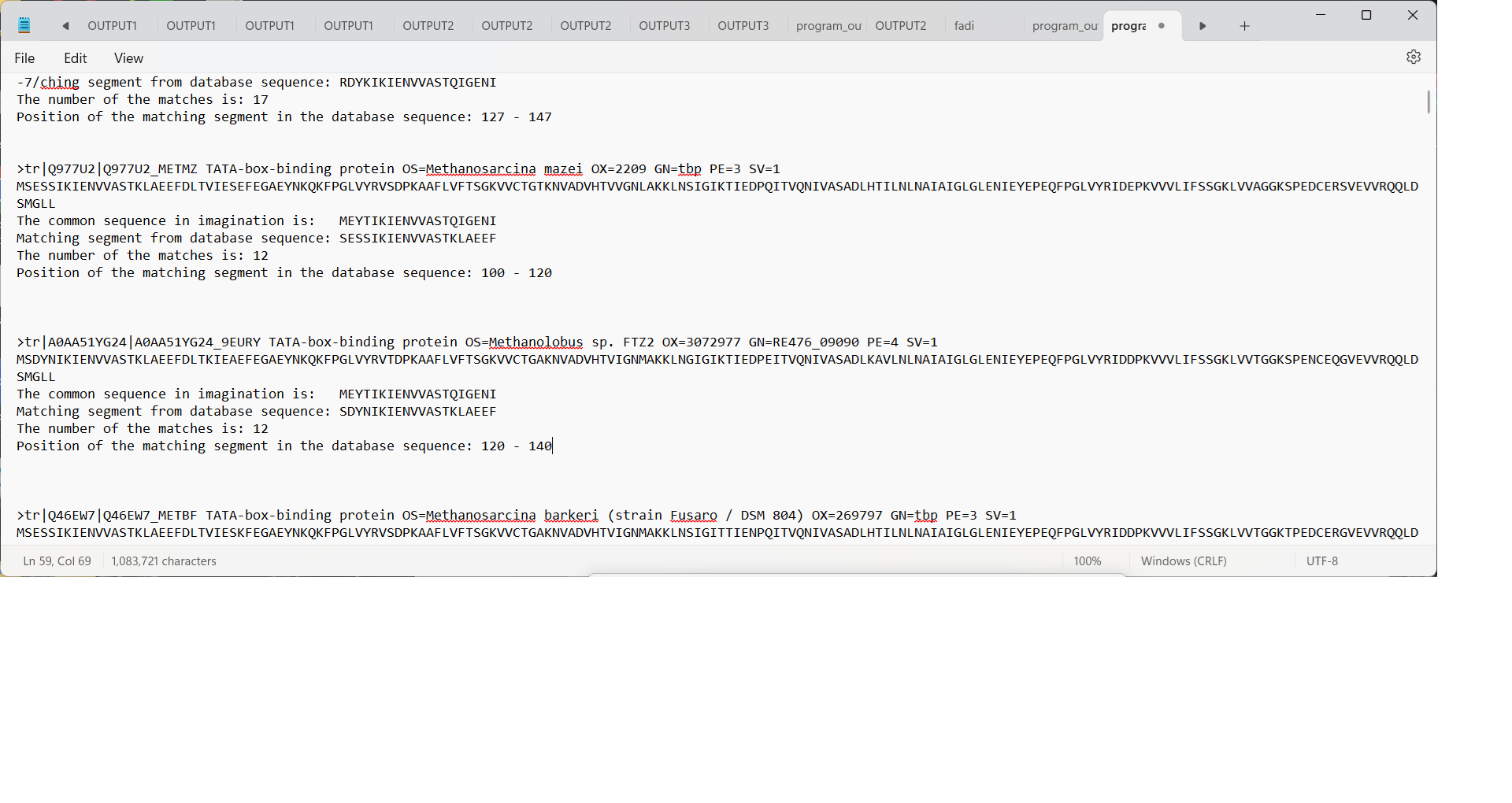
****

**7.5. Class diagram**

A picture containing text, diagram, screenshot, plan

Description automatically generated

**7.6. Results**



**Here we can see the search results for a certain protein, with a level of similarity up to 8 mismatches, after that we will filter the results on 3 levels, for example, on the first level we will take the first result (first sequence) and shoot all the sequences that have up to one mismatch , after that a second step with two inconsistencies, after that with 3 inconsistencies, and then we filtered the results.**

**7.7. Testing plan**

To evaluate the system's performance, we will execute the program using substantial input

|  |  |  |
| --- | --- | --- |
| **Test No.** | **Test subject** | **Expected result** |
| Protein DB tests | | |
| 1. | Retrieve protein/peptide for UniProt DB | Create protein class with the expected protein details |
| 2. | Division to 20 amino acid long fragments with overlapping | Protein divided to fragments correctly |
| Training data tests | | |
| 3. | Generating pairs of fragments between two proteins | Receive all possible combinations of pairs. |
| 4. | Calculate Hamming distance between two mismatch fragments | Return 20 |
| 5. | Calculate Hamming distance between two match fragments | Return 0 |
| 6. | Remove all pairs with hamming distance under the threshold | A table with pairs that has a hamming distance equal or above threshold |
| 7. | Check weight function result between two fragments with a previously calculated result | Weight function result equals to the previously calculated result |
| User interface tests | | |
| 8. | Enter invalid protein/peptide sequence | Error message appear. |
| 9. | Enter invalid protein/peptide name | Error message appear. |
| 10. | Enter valid protein/peptide name | Return a correct protein sequence |
| 11. | Enter invalid protein/peptide index | Error message appear. |
| 12. | Enter valid protein/peptide index | Return a correct protein sequence |
| 13. | Load an empty protein/peptide DB | Error message appear. |
| 14. | Try to search when protein/peptide DB isn't ready | Error message appear. |
| 15. | Try to search without input | Warning message appear. |
| 16. | Try to research the same protein/peptide twice | Receive the same result |
| 17. | Enter a known valid protein/peptide sequence | List of expected results and their similarity score |

**User help :**

**In this project it is about finding similarities between proteins and identifying homology, to find proteins like the desired protein,**

**you need to open the database that contains the protein sequences that exist in the world and run the code for the desired protein,**

**the size of the database files is over 128 gigabytes, from the UNIPROT website.**

**https://www.uniprot.org/uniprotkb?query=mitochondria-derived+peptides.**

**8.reference**

1. Frenkel Z, Trifonov, E., Walking through protein sequence space. *Journal of Theoretical Biology*, 2007 Jan 7; 244(1):77-80.
2. Frenkel, Z.M. and Trifonov, E.N. Evolutionary networks in the formatted protein sequence space. *Journal of Computational Biology*, 2007 Oct; 14(8):1044-57
3. K Grewal, R., & Roy, S. (2015). Modeling proteins as residue interaction networks. *Protein and peptide letters*, *22*(10), 923-933.
4. Z.M. Frenkel, Z.M. Frenkel, E.N. Trifonov, S.Snir, "Structural relatedness via flow networks in protein sequence space", *Journal of Theoretical Biology* 260: 438–444, 2009.
5. John-Marc Chandonia, "The ASTRAL Compendium in 2004", Nucleic Acids Research, Vol. 32, 2004.

,MN B ?<mB

1. MRFalign: Protein Homology Detection through Alignment of Markov Random Fields.
2. Daniels NM, Hosur R, Berger B, Cowen LJ (2012) SMURFLite: combining simplified Markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone. Bioinformatics 28: 1216–1222.
3. GraphAligner: rapid and versatile sequence-to-graph alignment .
4. Project in Software Engineering -NETWORK PARTITION BY CONDUCTIVITY FOR PCN-June 2014.
5. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., ... & Yeh, L. S. L. (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research*, *32*(suppl\_1), D115-D119.
6. Capstone Project Phase - Enhancement of the K-mismatch search algorithm Via tree data-structures – Spring-2022.
7. Esko Ukkonen. "*Approximate string-matching with q-grams and maximal matches*".
8. I. Kurfarseva, R. Abagyan, "Methods of protein structure comparison", *Methods Mol Biol.,* 857:231–257,2012.
9. M. Petitjean, "On the root mean square quantitative chirality and quantitative symmetry measures", *Journal of mathematical physics* vol.40, 9:4587-4597,1999.
10. Z.M. Frenkel, E.N. Trifonov "From protein sequence space to elementary protein modules", *Science Direct* ,65-68,2008.
11. Z. Frenkel, "A system and method for generating detection of hidden relatedness between proteins via a protein connectivity network", Application Number: IL2015/050489, 19/11/2015.
12. Kalaev, M., Smoot, M., Ideker, T., & Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, *24*(4), 594-596.
13. Daniels, N. M., Hosur, R., Berger, B., & Cowen, L. J. (2012). SMURFLite: combining simplified Markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone. *Bioinformatics*, *28*(9), 1216-1222.