Prediction of Novel Proteins for causing Cardiovascular Disease: Protein-Protein Interaction Analysis

Nihitha Vemulapalli[1], Korivi Baby Naina Tara[1] Velagapudi Sravani[1], Dr. K. Srinivas[2]

[1]Dept. of CSE, V R Siddhartha Engineering College, Vijayawada, India, nihithavemulapalli@gmail.com

[1]Dept. of CSE, V R Siddhartha Engineering College, Vijayawada, India,

[1]Dept. of CSE, V R Siddhartha Engineering College, Vijayawada, India,

[2]Dept. of CSE, V R Siddhartha Engineering College, Vijayawada, India vrdrks@gmail.com

Abstract--Cardiovascular disease (CVD) is typically caused by the accumulation of fatty deposits inside the arteries, which causes the flow of blood to the heart to be severely reduced or blocked, eventually leading to a heart attack. According to the World Health Organization (WHO), 17.9 million people die each year from CVDs, accounting for an estimated 32% of all deaths worldwide. The *In-Vivo* and *In-Vitro* approaches are costly and take a long time to achieve the desired result, whereas *In-Silico* approach gives the result more quickly. In this study the *In-Silico* approaches like Evolutionary Analysis and Protein-Protein Interactions are used to provide new insights into cardiovascular disease and allow for the prediction of new Cardiovascular Disease-causing proteins. Here, Evolutionary analysis is initially used to identify the similarities at the cellular and molecular levels of CVD-causing proteins and Protein-Protein Interaction (PPI) is used to build a network that identifies existing associations and predicts new ones between human proteins. This study reveals that Receptor tyrosine-protein kinase (RET), Neurotropic Receptor Tyrosine Kinase 2 (NTRK2), Butyrylcholinesterase (BChE) and Brain-derived neurotropic factor (BDNF) are the new proteins directly and indirectly associated with CVD.

Keywords—Evolutionary analysis, Cardiovascular Disease (CVD), Protein-Protein interaction analysis, In Vivo, In Vitro, In-Silico, RET, NTRK2, BChE, BDNF.

I. Introduction

Cardiovascular disease is the leading cause of death globally. Cardiovascular disease kills nearly 19 million people each year, accounting for an estimated 32% of all deaths worldwide. The major risk factors are physical inactivity, an unhealthy diet, consumption of alcohol and tobacco [1]. One of the largest global burdens of cardiovascular disease (CVD) is seen in India. Compared to 15.2% and 6.9%, respectively, in 1990, CVDs in India amounted to 28.1% of all fatalities and 14 % of all disability-adjusted life years (DALYs) in 2016 [2].

Cardiovascular disease, which continues to be the leading cause of death in women, manifests 7 to 10 years later in women than in males. The diagnosis and management of coronary heart disease in women involve a number of significant challenges. Studies have demonstrated that premenopausal women who experience hormonal disruption have a higher risk of atherosclerosis and CHD incidents. Pregnancy-related hypertension increases the chance of developing hypertension and early cardiovascular disease later in life in women. Although they present a unique

opportunity for better cardiovascular risk assessment and prevention, the peculiarities of illnesses associated to pregnancy have not yet been taken into account in the most recent recommendations for CHD prevention in women [3].

Cardiovascular disease and stress have been connected. Epidemiological studies show that persistent stress is a risk factor for coronary heart disease. Cardiovascular illnesses can be brought on by short-term stress, while long-term stress can result in death.[4]

Cardiovascular disease is a condition that affects the heart or blood arteries (CVD). It is frequently linked to atherosclerosis, which is the accumulation of fatty deposits in the arteries and increases the risk of blood clots. Coronary artery disease (CAD), cerebrovascular disease (CVD), peripheral artery disease (PAD), and aortic atherosclerosis are the four subtypes of cardiovascular disease. Angina, myocardial infarction (MI), and/or heart failure are all symptoms of coronary artery disease (CAD), also known as coronary heart disease (CHD), which develops as a result of decreased myocardial perfusion. One-third to fifty percent of CVD cases are attributed to it. The cerebrovascular disease is connected to strokes, also known as cerebrovascular accidents, and transient ischemic attacks (TIAs). Peripheral artery disease (PAD), an arterial condition that can lead to claudication and impair the limbs. Thoracic and abdominal aneurysms are among the symptoms of aortic atherosclerosis. Fig 1 shows the types of Cardiovascular Disease [5].

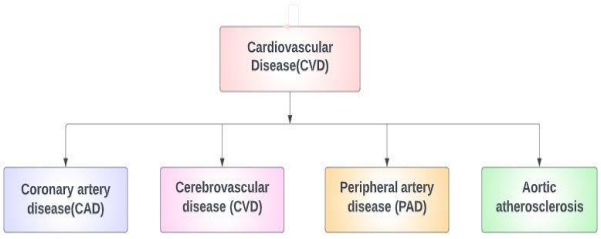

Fig.1: Types of Cardiovascular Disease

Diagnostic tools including electrocardiography, echocardiography, and radio diagnostics are used in modern clinical practice [6]. However, the above-mentioned techniques have significant limitations in terms of cost and time. So by using bioinformatics, one can overcome the limitations. The use of computation and analysis tools to collect and understand biological data is known as bioinformatics. For the administration of data in contemporary biology and medicine, bioinformatics is crucial. Computer software applications like BLAST and Ensemble, which rely on the availability of the internet, are part of the bioinformatics toolkit. One of the most significant accomplishments of bioinformatics to date is the study of genome sequence data, particularly the analysis of the human genome project. Future contributions to the functional understanding of the human genome, which will improve the identification of pharmacological targets and enable individualized treatment, are among the opportunities in the field of bioinformatics [7].

II. Literature Survey

Roman Fando et al [8] discussed the evolution of bioinformatics, genetic engineering, proteomics and biological databases. Bioinformatics techniques are used in the analysis of mutation-spectrum complexity and its dependence on nucleotide context, genome structure studies. They have discussed some recent developments in bioinformatics, specifically how bioinformatics and molecular engineering bring us to the threshold of macromolecular and cell component synthesis. But some potential risks of using a bioinformatics approach are sequencing error and human error.

Dr. Lakshmi Prasad Koyi et al [9] discussed the existing heart disease prediction systems, the Cleveland dataset, machine learning algorithms (Naive Bayes, Decision Trees, Neural Networks and Support Vector Machines), supporting tools (WEKA, Rapid Miner, Apache Mahout and Matlab), considerable

metrics and some of the research challenges including input dataset attributes modelling, attribute risk factor calculation, correlations mining, threshold determination and achieving high accuracy in disease prediction.

Long H. Ngo et al. [10] discussed that blood protein concentrations can be clinically useful and predictive biomarkers for cardiovascular disease. The aim of the study is to find out the relation between plasma protein levels and CVD incidence and to assess the influence of religiosity on significant protein-CVD associations, in South Asians from the MASALA study. They used SOMAscan to look for protein expression in plasma samples. They identified 36 proteins that were significantly expressed in CVD patients.

Man Li et al. [11] went into great detail about the relationship between growth differentiation factor-15 and cardiovascular disease. GDF-15 is a marker of oxidative stress, inflammation and it is associated with CVD. They concluded that patients with GDF-15 levels greater than 1800 ng/L were at an increased risk of major adverse cardiovascular events and death.

Luc Rochette et al. [12] discussed that GDF-15 acts as an inflammatory marker and plays a role in the pathogenesis of cardiovascular diseases, metabolic disorders, and neurodegenerative processes. They also stated that GDF-15 is highly expressed in cardiomyocytes after ischemia and in the heart within hours after myocardial infarction. Therefore, GDF-15 is identified as a cardiac hormone, and GDF-15 may be a predictive biomarker of cardiac events.

Fanaja Harianja Randriamahenintsoa et al. [13] discussed an expert system that is good at predicting the most suitable algorithm for both multiple sequence alignment and phylogenetic tree construction. Knowledge bases were designed using datasets whose instances are sets of protein sequences and whose attributes are various characteristics of the sequences. Decision trees were used and here the inference engine could predict the most suitable algorithm along with the most appropriate parameters.

V. Srinivasa Rao et al. [14] discussed protein-protein interaction methods. Protein-protein interactions handle a wide range of processes, which include cell-to-cell interactions, metabolic activities, biological processes and development control. Comparison of protein-protein interaction methods, computational analysis of protein-protein interaction networks. The role of protein-protein interactions in proteomics was also discussed.

Joy Dip Barua et al. [15] addressed how the condition's risk factors are related to it, various analyses were done. One of the tests they conducted was gene expression analysis, which is a highly effective way to acquire genetic data at the molecular level. To investigate how the illness interacted at the molecular level, they employed pathway-based analysis. This work uncovered the molecular links between CVD and associated risk factors. They used the Gene Expression Omnibus (GEO) datasets from the National Center for Biotechnology Information (NCBI). Using this information, they examined the genetics and cardiovascular disease risk variables.

III. Tools

STRING:

In molecular biology, STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource of known and predicted protein–protein interactions. The STRING database contains information from numerous sources, including experimental data, computational prediction methods and public text collections. It is freely accessible and it is regularly updated. The resource also serves to highlight functional enrichments in user-provided lists of proteins, using a number of functional classification systems such as GO, Pfam and KEGG. The latest version 11b contains information on about

24.5 million proteins from more than 5000 organisms. STRING has been developed by a consortium of academic institutions including CPR, EMBL, KU, SIB, TUD and UZH [16].

ClustalW:

ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods. In these, the most similar sequences, that is, those with the best alignment score are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained. This heuristic approach is necessary because finding the global optimal solution is prohibitive in both memory and time requirements [17].

ClustalX:

ClustalX is a windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analyzing the results. The sequence alignment is displayed in a window on the screen. A versatile coloring scheme has been incorporated allowing you to highlight conserved features in the alignment. The pull-down menus at the top of the window allow you to select all the options required for traditional multiple sequence and profile alignment [18].

Clustal Omega:

Clustal Omega is a multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. It produces biologically meaningful multiple sequence alignments of divergent sequences. Evolutionary relationships can be seen via viewing Cladograms or Phylograms [19].

Following are the steps involved in Clustal Omega:

1. **Sequence:** Three or more sequences to be aligned can be entered directly into this box. Sequences can be in GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR/NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only) format.
2. **Set Your Parameters:** Remove any existing alignment (gaps) from input sequences, Format for generated multiple sequence alignment.
3. **Submission:** It's possible to identify the tool result by giving it a name. This name will be associated to the results and might appear in some of the graphical representations of the results. Running a tool is usually an interactive process, the results are delivered directly to the browser when they become available. Depending on the tool and its input parameters, this may take quite a long time. It's possible to be notified by email when the job is finished by simply ticking the box "Be notified by email". An email with a link to the results will be sent to the email address specified in the corresponding text box. Email notifications require valid email addresses.

IV. Methodology

The proteins that cause cardiovascular disease are studied using Homology Analysis, Evolutionary Analysis, and Protein-Protein Interaction (PPI). And finally, new proteins causing cardiovascular disease can be identified. Fig 4.1 illustrates the proposed mechanism.

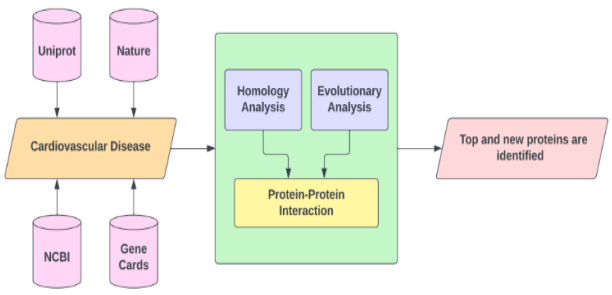

Fig 4.1: Proposed System

A. Data Collection

The dataset related to this project is all about the proteins related to Cardiovascular Disease. For this project we require a limited and main proteins that cause the disease. Identifying the top and main proteins by studying number of literature survey related to that disease. By collecting the main proteins for the disease through literature survey. The FASTA format of main proteins are collected from NCBI (National Center for Biotechnology Information), Uniprot, Genecards. In this website having proteinname, protein id, FASTA format of protein etc taking FASTA format as input. The top proteins responsible for cardiovascular disease (CVD) are listed in Table 1.

SI NO.	PROTEIN NAME	PROTEIN ACRONYM	ACCESS NUMBER
1	Troponin	TNNT2	P45379
2	Natriuretic Peptide B	NPPB	P16860
3	Lipoprotein(a)	LPA	P08519
4	Growth Differentiation Factor 15	GDF 15	Q99988
5	Cystatin	CST3	P01034
6	Lamin-B1	LMNB1	P20700
7	Tumor necrosis factor ligand superfamily member 18	TNFSF18	Q9UNG2
8	SLIT and NTRK-like protein 5	SLITRK5	Q810B7
9	insulin-like growth factor binding protein 1	IGFBP1	P08833
10	Leptin	LEP	P41159
11	Adipsin/ Complement factor D	CFD	P00746
12	Apolipoprotein-A1	APO-A1	P02647
13	BAG family molecular chaperone regulator 3	BAG3	O95817
14	Desmin	DES	P17661
15	Myosin-7	MYH7	P12883
16	Cardiac Phospholamban	PLN	P26678
17	RNA-binding protein 20	RBM20	Q5T481
18	Troponin C slow skeletal and cardiac muscles	TNNC1	P63316
19	Desmoglein-2	DSG2	Q14126
20	Junction plakoglobin	JUP	P14923
21	Desmoplakin	DSP	P15924
22	Transmembrane protein 43	TMEM43	Q9BTV4
23	GDNF family receptor alpha-1	GFRA	P56159

Table 1. List of identified top proteins responsible for CVD

B. Collecting FASTA format of proteins:

In this project the FASTA format of the proteins related to CVD are collected through various websites like NCBI (National Center for Biotechnology Information), Genecards, Uniprot and from Survey papers. The FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes.

Input: Cardiovascular Disease.
Output: FASTA format of proteins.
Methodology:
Step1. Open the website like NCBI[20], Uniport[21] and Genecards [22].
Step2. Enter the Cardiovascular Disease in search bar
Step3. A lot of proteins are displayed.
Step4. From that proteins select homo sapiens.
Step5. Collect FASTA format for related proteins.

C. Homology Analysis

Homology analysis is a concept that considers similarities between nucleic acid or protein sequences from two different organisms. In this model, FASTA protein format is used to find homology analysis of proteins for CVD. The homology analysis is identified by using the Clustal Omega tool.

Input: FASTA format of proteins.
Output: Phylogentic Analysis of proteins.
Methodology:
Step1. Open the website Clustal Omega[19].
Step2. In that website the input is either protein name or protein id or FASTA format of protein.
Step3. Click continue after the input is taken. It will display a text which contain the alignment of protein sequence.
Step4. Click on Phylogenetic tree construction. It will display the tree structure with Proteins.
Step5. By looking at the structure analyze the proteins. The proteins with similar sequences are all descended from the same parent branch.

D. Protein-Protein Interaction

The data collected from Module 1 is taken as input by the String tool for Protein-Protein Interaction. Protein-protein interaction plays a key role in predicting the protein function of the target protein and the drug ability of molecules. For this PPI, network, CVD-causing proteins are taken, and a network is constructed. The network describes the interactions it has had with other proteins, either directly or indirectly.

Input: FASTA format of proteins or protein name.
Output: Protein-Protein Interaction network.
Methodology:
Step1. Open the tool for Protein-Protein Interaction that is STRING [16].
Step2. To give multiple proteins click on multiple proteins option.
Step3. Enter the protein names for PPI network.
Step4. Click on continue will get a large network of interactions.
Step5. Change the confidence value according to the Interactions needed.
Step6. By looking at the structure analyze the proteins. The proteins with similar sequences are all descended from the same parent branch.

V. Result

The result includes collection of FASTA-format proteins, Homology Analysis and Protein-Protein Interactions of the cardiovascular disease-causing proteins.

A. Data Collection: Data is collected from NCBI, GeneCards, and Uniprot. A total of 23 proteins, along with their FASTA format related to CVD, are collected.

Fig 5.1 shows the data of the proteins taken from the Uniprot when searching for a protein related to CVD.

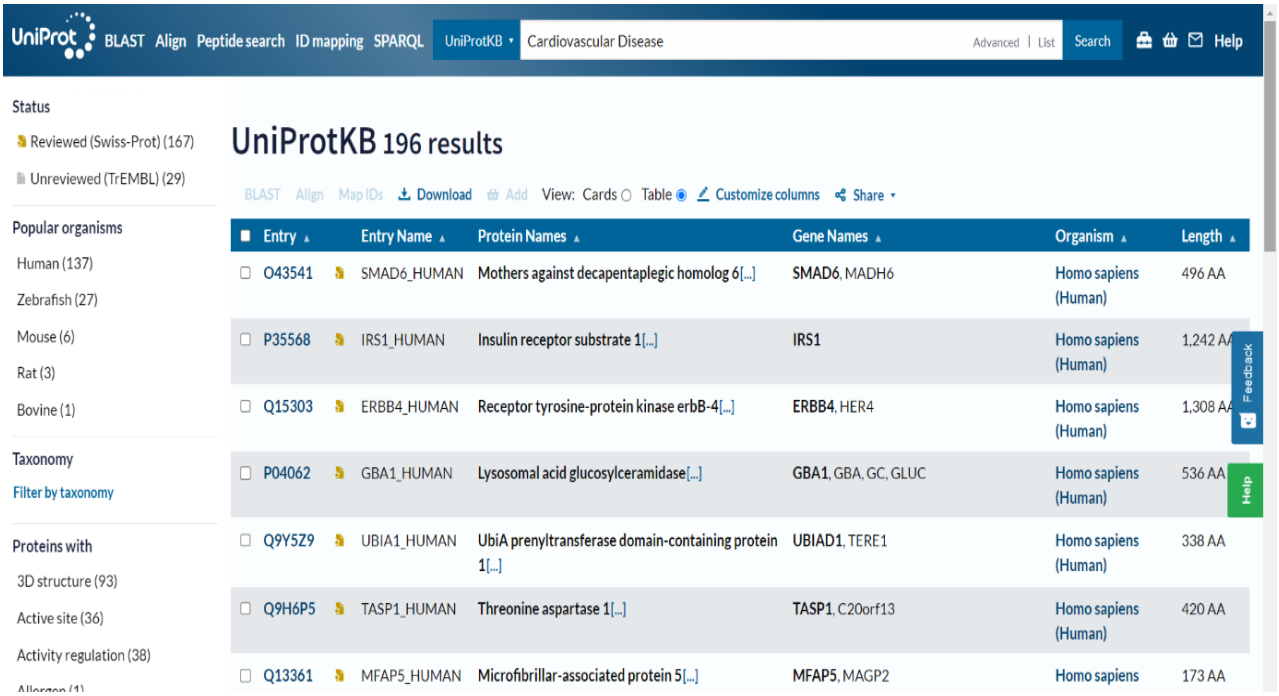

Fig 5.1: UniPort proteins sample image.

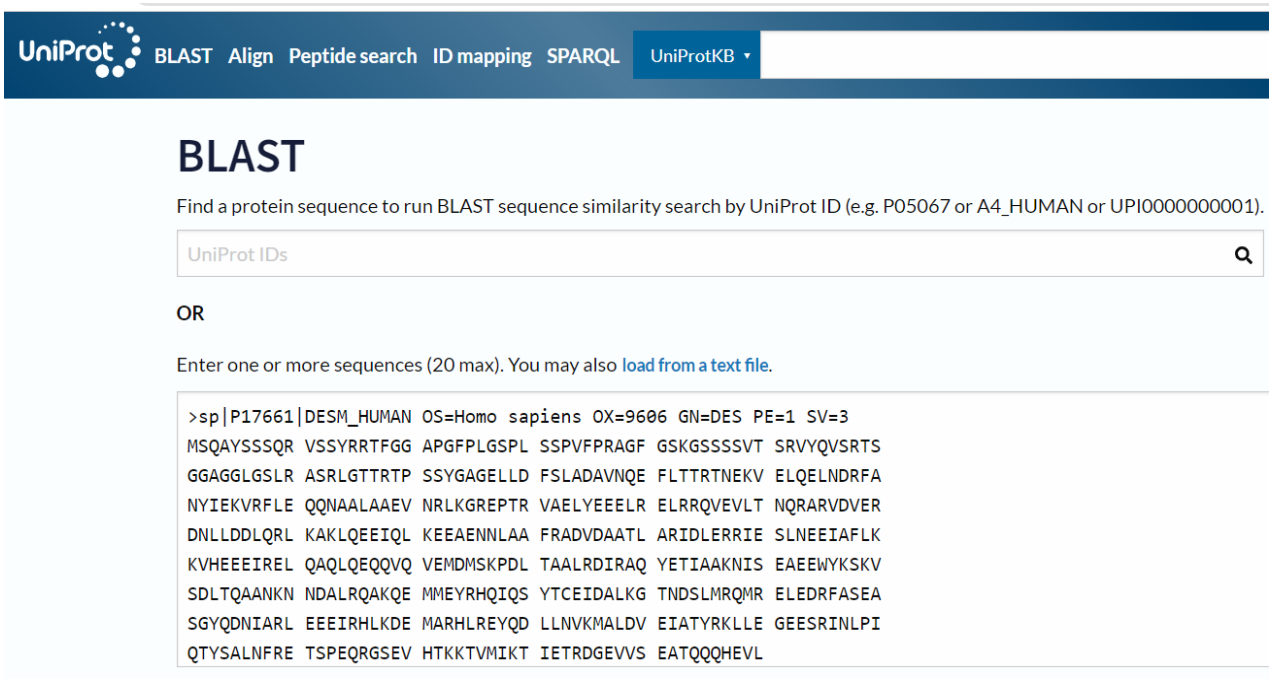

Fig 5.2 FAST data of the Desmin protein from UniPort

B. Homology Analysis

A phylogenetic tree is constructed by taking 23 CVD-causing proteins, as mentioned above in III.A, of which 10 have a lot of similarities to the CVD. The 10 proteins are BAG family molecular chaperone regulator 3 (BAG3), Growth Differentiation Factor 15 (GDF15),GDNF family receptor alpha-1 (GFRA1), Desmin (DESM), Junction plakoglobin (JUP),Lamin-B1 (LMNB1), Desmoglein-2 (DSG2), Desmoplakin (DESP), Myosin-7 (MYH7), and Troponin T2 (TNNT2). The 10 proteins share 0.20132, 0.22944, 0.21332, 0.30583, 0.34371, 0.32323, 0.42492, 0.39373, 0.39953, and -0.08736 percent similarity in their amino acid sequences with CVD. Fig 5.1 shows the phylogenetic analysis of the given 23 proteins.

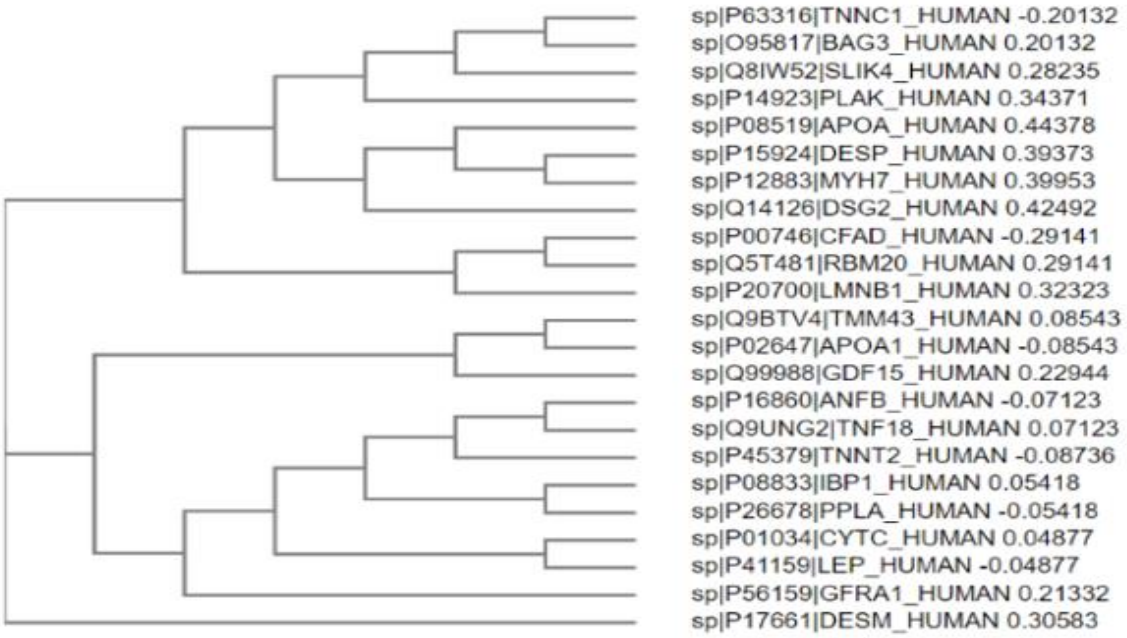

Fig 5.3: Phylogenetic Analysis – Cardiovascular Disease

C. Protein-Protein Interaction Network

Applying Protein-Protein Interaction (PPI) gives us an idea of how all 10 proteins have an interaction among themselves, and this is illustrated in the Fig 7.1 .PPI is applied to the 10 proteins: BAG family molecular chaperone regulator 3 (BAG3), Growth Differentiation Factor 15 (GDF15), GDNF family receptor alphaptor alpha-1 (GFRA1), Desmin (DESM), Junction plakoglobin (JUP), Lamin-B1 (LMNB1), Desmoglein-2 (DSG2), Desmoplakin (DESP), Myosin-7 (MYH7), and Troponin T2 (TNNT2) and the PPI network is constructed using the string tool. The confidence is considered to be 0.7, and the connections with no interactions are removed.

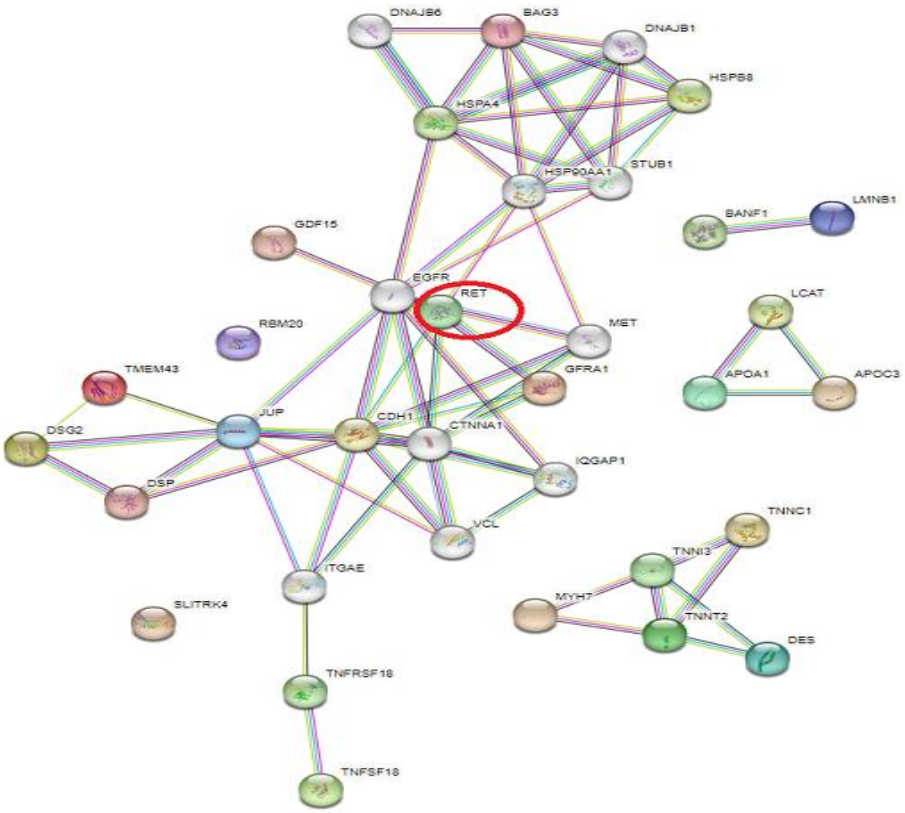

Figure 5.4: PPI network-Cardiovascular Disease

From the PPI network, we can observe that RET is interacting directly with the CVD-causing proteins. Therefore, RET might be one of the new proteins responsible for causing CVD. PPI is also used on the first ten proteins, along with RET, with a confidence level considered to be 0.4. From the results, it is clear that RET, BCHE, BDNF, and NTRK2 have a direct and indirect relationship with CVD-causing proteins. This is illustrated in the Fig 5.5.

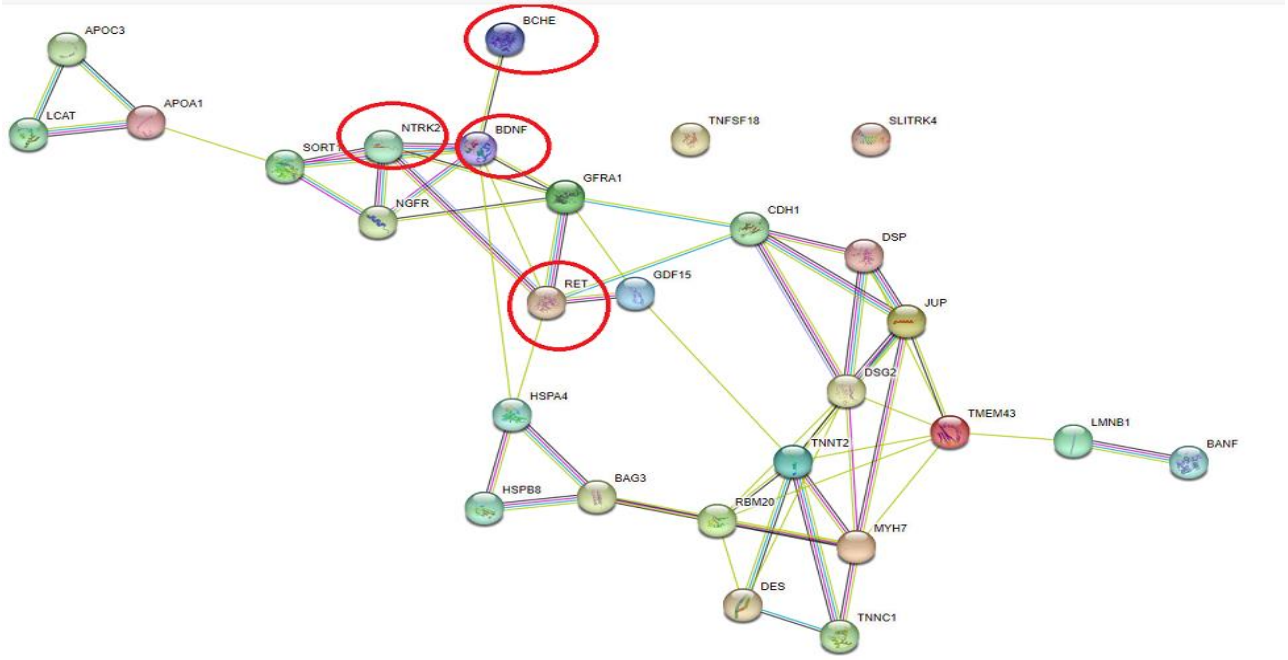

Figure 5.5: PPI network-Cardiovascular Disease

VI. Conclusion and Future Work

The common proteins that have an influence on CVD are identified in this study by constructing a polygenetic tree for the proteins, which are taken from the Uniport and NCBI gene cards, and identifying the interaction between these proteins by constructing a Protein-Protein Interaction (PPI) network. New proteins are found that can also be responsible for causing CVD. The identified new proteins are BCHE, BDNF, NTRK2, and RET, which can also be related to CVD. For this study the Future work include the Domain-Domain analysis which includes the detailed analysis of the common proteins by evaluating the Protein Family, Gene structure, Amino acids sequences, Biological process, Molecular function, Cellular Component, etc. The Future work also includes the drug target for the disease.

VII. References

[1] World Health Organization (WHO).[Updated 2021 June 11].”Cardiovascular diseases (CVD)”. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] Sreeniwas Kumar A, Sinha N. Cardiovascular disease in India: A 360 degree overview. Med J Armed Forces India.2020,Jan;76(1):1-3.doi: 10.1016/j.mjafi.2019.12.005. Epub 2020 Jan 13. PMID: 32020960; PMCID: PMC6994761.

[3] Maas AH, Appelman YE. Gender differences in coronary heart disease. Neth Heart J. 2010 Dec;18(12):598-602. doi: 10.1007/s12471-010-0841-y. PMID: 21301622; PMCID: PMC3018605.

[4] Steptoe A, Kivimäki M. Stress and cardiovascular disease. Nat Rev Cardiol. 2012 Apr 3;9(6):360-70. doi: 10.1038/nrcardio.2012.45. PMID: 22473079

[5] Olvera Lopez E, Ballard BD, Jan A. Cardiovascular Disease. [Updated 2022 Aug 8]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-Available from: https://www.ncbi.nlm.nih.gov/books/NBK535419/

[6] Khan S, Rasool ST, Ahmed SI. Role of Cardiac Biomarkers in COVID-19: What Recent Investigations Tell Us? Curr Probl Cardiol. 2021 Oct;46(10):100842. doi: 10.1016/j.cpcardiol.2021.100842. Epub 2021 Mar 19. PMID: 33994028; PMCID: PMC7977033.

[7] Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.

[8] R. Fando and M. Klavdieva, "Bioinformatics: Past and Present," 2018 International Conference on Engineering Technologies and Computer Science (EnT), 2018, pp. 34-36, doi: 10.1109/EnT.2018.00013.

[9] L. P. Koyi, T. Borra and G. L. V. Prasad, "A Research Survey on State of the art Heart Disease Prediction Systems," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 799-806, doi: 10.1109/ICAIS50930.2021.9395785.

[10] Ngo, L.H., Austin Argentieri, M., Dillon, S.T. *et al.* Plasma protein expression profiles, cardiovascular disease, and religious struggles among South Asians in the MASALA study. *Sci Rep* 11, 961 (2021). https://doi.org/10.1038/s41598-020-79429-1

[11] Li, M., Duan, L., Cai, YL. *et al.* Growth differentiation factor-15 is associated with cardiovascular outcomes in patients with coronary artery disease. *Cardiovasc Diabetol* 19, 120 (2020). https://doi.org/10.1186/s12933-020-01092-7

[12] Rochette L, Dogon G, Zeller M, Cottin Y, Vergely C. GDF15 and Cardiac Cells: Current Concepts and New Insights. Int J Mol Sci. 2021 Aug 18;22(16):8889. doi: 10.3390/ijms22168889. PMID: 34445593; PMCID: PMC8396208

[13] F. H. Randriamahenintsoa, T. H. Raboanary, H. A. Raboanary and J. A. Raboanary, "A novel expert system for the prediction of accurate multiple sequence alignment and phylogenetic tree construction algorithms," 2017 IEEEAFRICON, 2017,pp.94-99, doi:10.1109/AFRCON.2017.8095462.

[14] V. Srinivasa Rao, K. Srinivas, G. N. Sujini, G. N. Sunand Kumar, "Protein-Protein Interaction Detection: Methods and Analysis", International Journal of Proteomics, vol. 2014, Article ID 147648, 12 pages, 2014.(https://doi.org/10.1155/2014/147648)

[15] Joy Dip Barua, Shudeb Babu Sen Omit, Humayan Kabir Rana, Nitun Kumar Podder, Utpala Nanda Chowdhur and Md Habibur Rahman, Bioinformatics and System Biological Approaches for the Identification of Genetic Risk Factors in the Progression of Cardiovascular Disease,09August2022,
Available at: https://doi.org/10.1155/2022/9034996

[16] "Welcome to string," *STRING*. [Online].
Available: https://string-db.org/. [Accessed: 23-Dec-2022]

[17]"Multiple Sequence Alignment-CLUSTALW," Genome.jp,2019.
Available at: https://www.genome.jp/tools-bin/clustalw [Accessed: 23-Dec-2022]

[18] "ClustalW / Clustal X, Clustal W and Clustal X Multiple Sequence Alignment." Available at: http://www.clustal.org/clustal2/ (Accessed: December 23, 2022).

[19] "Clustal Omega Multiple Sequence Alignment EMBL-EBI."
Available at: https://www.ebi.ac.uk/Tools/msa/clustalo/ (Accessed: Dec.20, 2022).

[20] "Home - protein - NCBI," Nih.gov. [Online].
Available: https://www.ncbi.nlm.nih.gov/protein/.
[Accessed: 20-Dec-2022].

[21] "Uniprot website fallback message," UniProt. [Online].
Available: https://www.uniprot.org/.[Accessed:20-Dec-2022]

[22] Cardiovascular related genes - GeneCards Search Results," www.genecards.org.
https://www.genecards.org/Search/Keyword?queryString=Cardiovascular
(Accessed Dec. 20, 2022).