EVOLUTIONARY STUDY ON COMMON BIOMARKERS - ASSOCIATED WITH MMAD, ESPD AND MS APPLYING GRAPH THEORETICAL MODELLING

MASTER OF COMPUTER APPLICATION
in
Faculty of Engineering & Technology

By

MEGHDIPA GHOSH

Exam Roll No**: MCA186016**
Registration No**: 133678 of 15 - 16**

Under the Guidance of

Dr. Ujjwal Maulik

Department of Computer Science & Engineering

Jadavpur University

Kolkata - 700 032

2018

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

I hereby recommend that the thesis entitled **“Evolutionary Study On Common Biomarkers Associated With MMAD ESPD And MS Applying Graph Theoretical Modeling”** prepared under my supervision by **MEGHDIPA GHOSH,** Exam- Roll No: MCA186016, be accepted for the degree of **Master of Computer Application** of Jadavpur University, Kolkata.

Prof. Ujjwal Maulik (Project Supervisor)
Head, Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Prof. Ujjwal Maulik
Head, Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Computer Science and Engineering
Jadavpur University, Kolkata-700032

JADAVPUR UNIVERSITY

FACULTY OF ENGINEERING & TECHNOLOGY

CERTIFICATE OF APPROVAL*

The foregoing thesis **"Evolutionary Study On Common Biomarkers Associated With MMAD ESPD And MS Applying Graph Theoretical Modeling"** at instance is hereby approved as a creditable study of an engineering subject carried out and presented in a manner of satisfactory to warrant its acceptance as pre-requisite to the degree for which it has been submitted. It is to be understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed and conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

Final Examination for the Evaluation of Thesis

Board of Examiners

____

____

(Signature of Examiners)

(*)Only in case this thesis is approved

JADAVPUR UNIVERSITY

FACULTY OF ENGINEERING & TECHNOLOGY

DECLARATION OF ORIGINALLY COMPLIMANCE OF ACADEMIC ETHICS

I hereby declare that this thesis entitled **"Evolutionary Study On Common Biomarkers Associated With MMAD ESPD And MS Applying Graph Theoretical Modeling"** contains literature survey and original research work by the undersigned candidate, as part of her Degree of Master of Computer application.

All information in this document has been obtained and presented according to the academic rules and ethical conduct.

I also declare that as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name : **Meghdipa Ghosh**
Registration No : **133678 of 2015-2016**
Examination Roll No : **MCA186016**

Thesis Title : **Evolutionary Study on Common Biomarkers Associated With MMAD, ESPD and MS Applying Graph Theoretical Modeling**

Signature:

____

Date:

____

ACKNOWLEDGEMENT

"Knowledge is in the end based on acknowledgement" – Ludwig Wittgenstein

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my postgraduate experience has been one that I will cherish forever.

Foremost, I would like to express my profound gratitude and sincere thanks to my adviser, **Dr. Ujjwal Maulik, Head of the Department, Department of Computer Science & Engineering**, **Jadavpur University**, for his valuable suggestions, guidance, constant encouragement and intent supervision at every stage of my work. I have been amazingly fortunate to have him as my project Supervisor who gave me the freedom to explore on my own, and at the same time guided me to recover when my steps faltered. It has been a great learning process for me.

I am grateful to **Mr. Sagnik Sen**, Research Scholars, Department of Computer Science & Engineering, Jadavpur University, for his encouragement and practical advice. I am thankful to him, for introducing me to the world of Machine learning and Bioinformatics which helped in understanding the viewpoint of this dissertation. I am deeply grateful to him for the long discussions that helped me sort out the technical details of my work. I am thankful to him for his insightful comments and constructive criticisms at different stages of my research, which were thought provoking and helped me focus on my ideas.

I would like to thank Jadavpur University for providing me the opportunity to work in such a nice working environment. My sincere gratitude to all my professors who were always there to extend a hand whenever I was in need of guidance. Many friends have helped me stay sane through these difficult years. Their support and care helped me overcome setbacks and stay focused on my postgraduate study. I greatly value their friendship and I deeply appreciate their belief in me.

Most importantly, none of this would have been possible without the love and patience of my family. My family, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength all these years. I would like to express my heart-felt gratitude to my family for bearing with me and providing me with constant motivation during the entire time.

Date:
Place:

Meghdipa Ghosh
MCA(C.S.E)
Exam Roll No. MCA186016

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Mild cognitive impairment (MCI) is an intermediate stage between the expected cognitive decline of normal aging and the more-serious decline of dementia. A person with MCI is at an increased risk of developing Alzheimer's or another dementia. Neurodegenerative diseases like Alzheimer's disease, Parkinson's disease or multiple sclerosis occurs largely because of protein misfolding, which occurs mainly due to the malfunction of intrinsically disordered proteins or IDPs.

Dysfunction of Intrinsically disordered proteins (IDP) or intrinsically disordered regions (IDR) can bring about serious pathological problem as they take part in crucial biological function. Biological markers or biomarkers are quantifiable indicator of pathological process. The goal of this work is to perform an evolutionary study of the common biomarkers of Alzheimer's disease, Parkinson's disease and multiple sclerosis that will help us in predicting the evolution of these neurodegenerative diseases with greater accuracy and the therapeutic responses that should follow up.

Chapter 1: Introduction

1. INTRODUCTION

1.1 Background

Mild cognitive impairment (MCI) is an intermediate stage between the expected cognitive decline of normal aging and the more-serious decline of dementia. Mild cognitive impairment (MCI) causes a slight but noticeable and measurable decline in cognitive abilities including memory, judgement, and language and thinking skills. A person with MCI, particularly MCI with memory impairment, is at an increased risk of developing Alzheimer's or another dementia.

Alzheimer’s disease (AD) is a neurodegenerative disease that develops gradually and is characterized by the decline in memory and other cognitive functions, as well as behavioural changes[1][2]. The diagnostic criteria for AD propose three stages of AD, consisting on preclinical AD, mild cognitive impairment (MCI) due to AD and dementia due to AD. Approximately 15% of adults older than 65 years old suffer from MCI and from these more than half progress to AD within5 years [1][3].

Parkinson Disease (PD) is the second most common neurodegenerative disease, currently affecting 1% of people over age 65 [5]. As the world population ages, PD incidence will continue to rise over the next decade [6]. Approximately 20% of patients with PD have mild cognitive impairment (MCI) [5], and over 40% of PD patients with normal cognition at baseline develop MCI within 6 years [6]. MCI is considered an intermediate state of cognitive dysfunction in PD that typically progresses to PD dementia (PDD). Over 80% of people with PD develop dementia by 20 years into the disease [7] [4]

Multiple sclerosis (MS) is yet another neurodegenerative disease in which the insulating covers of nerve cells in the brain and spinal cord are damaged.[8]. Effect of multiple sclerosis usually varies from person to person depending on how the central nervous system is affected. Some may lose the ability to walk independently or at all while other may experience long periods of remission without any new symptoms.

Intrinsically disordered proteins(IDP) or intrinsically disordered regions(IDR) lack a stable three-dimensional structure and instead exist as a dynamic ensemble of interconverting structure. Though lacking stable secondary and tertiary structure IDP /IDR takes part in crucial biological functions. IDPs often fold to carry out biological reaction thus interacting with various other proteins, nucleic acids in the process. Since protein is an integral part of human life cycle their dysfunction can bring about serious pathological problems. Neurodegenerative disease is a type of protein misfolding diseases that occurs mainly because of a failure of a protein or a peptide to adopt to its functional conformational state. [9] Biological markers, or biomarkers, are objective and quantifiable indicators of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. [10] The identification of common biomarkers altering neurodegeneration could lead to early diagnosis or new drugs targeting the management of AD, PD or MS.

1.2 Motivation

IDPs are attractive target for drugs that modulate protein-protein interaction. Identifying the common biomarkers thus will help in more accurate early diagnosis of MS AD or PD and developing any kind of drugs or treatment for prevention of the same. This study will help us predict the evolution of these neurodegenerative diseases with greater accuracy and the therapeutic responses that should follow up.

1.3 Implementation framework

For this article, we have used the aid of various online tools and databases. A brief glimpse of all the languages, tools and databases we have used is given below:

The languages we have used are:

- **R scripting language(R** is a free software that is commonly used for statistical, data analysis and machine learning. Nowadays R is widely accepted as the most common language used for bioinformatics related research).
- **Matlab** (Matlab is a programming language that was essentially developed for numerical computing environment. It is used for various purposed such as matrix manipulation, plotting of function and data, algorithm implementation etc. We have mainly used for obtaining the graph plot of DCA contacts as we can see in the later part of this chapter)

Apart from that, we have used various online databases in order in different stages of our work. A brief introduction of the databases we have used is given below:

- **UniProt** (Uniprot is a freely accessible widely used protein database that contain protein sequence and huge amount of information about the biological functionality of the protein that is derived from research studies.)

- **D2P2** [11] (While Pondr computes the disorder percentage for a given protein sequence D2P2 is a freely accessible online database that store pre-computed disordered predictions on a large library of proteins. It contains 10,429,761 sequences in 1,765 genomes from 1,256 distinct species, with disorder predicted from 9 disorder predictors, and SCOP domain predictions from SUPERFAMILY.)

- **PDB** [12] (PDB or Protein Data Bank is a structural database that contains information about experimentally determined structures of proteins, nucleic acids, and complex assemblies. We have used this tool for obtaining the three dimensional structure of the proteins as discussed previously.)

- **Pfam** [13] (Last but not the least Pfam is another online freely accessible database that we have used here. Essentially Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. The current version of Pfam, Pfam 31.0, contains a total of 16712 families and 604 clans.)

Finally, the tools used are:

- **PONDR** (Pondr or Predictor of Natural Disorder is a sophisticated tool that is used to predict disordered region and percentage of disorder in any single sequence using various prediction algorithm. Some of the algorithms used by Pondr predictor are VL-XT, VLS and so on.)

1.4 Organization of thesis work

Chapter 1: Introduction

In introduction, a brief discussion about the various concepts used in this article is given. We have discussed in brief about AD, PD and MS and their relation to MCI. In addition, the idea of IDP or IDR is given here along with their relation in disease production. The motivation behind this work and the implementation framework of the same is also given here.

Chapter 2: Related Work

This chapter includes discussion about some previous work regarding intrinsically disordered proteins. Here we have also provided a brief glimpse on works related to Alzheimer's, Parkinson's and multiple sclerosis disease.

Chapter 3: Proposed Methodology

This chapter contains the details of our implementation. The work done can be dived into four main parts-

Common autoantibodies identification: given in chapter 3.2

Study of Disordered region: given in chapter 3.3

Sequence based Analysis: given in chapter 3.4

Structure based analysis: given in chapter 3.5

Chapter 3.6 discuss about how to compare the result obtained to reach the conclusion. Finally, in chapter 3.7 we have provided a workflow diagram to provide a better understanding.

Chapter 4: Results and discussion

In this chapter, a detailed discussion of the results of our work is given.

Chapter 5: Conclusion and future scope

Finally, the thesis work is concluded in this chapter and the future scope of the work is specified.

Chapter 2:

Literature Survey

2. LITERATURE SURVEY

Biomarkers may be useful for predictive diagnosis of Alzheimer’s disease, Parkinson’s disease or multiple sclerosis. The combination of cerebrospinal fluid (CSF) biomarkers and imaging has been investigated extensively for a number of years. Contribution of classical biomarkers in predicting Alzheimer’s and the importance of novel candidates as potential biomarkers has already been discussed at length (N. El Kadmiri et al) [14]. Previous work has also been done to identify the Alzheimer’s diseases morphometric signatures using various machine learning techniques [3].

Since multiple sclerosis is a complex heterogeneous disease, diagnostic criteria are based on symptoms, biomarkers, and MRI data and so on. However, over the past few years the usefulness of biomarkers have progressively decreased with the development of new MRI criteria though dozens new biomarkers especially in CSF has been described. Large-scale studies validating some of these new biomarkers have also provided confirmation of a restricted set of biomarkers (presented here in this review) as having potential value for different stages of the disease, including as early as clinically isolated syndrome and radiologically isolated syndrome. [15]

. First described 200 years ago, Parkinson Disease (PD) exhibits considerable heterogeneity in clinical presentation, as well as trajectory of motor and non-motor decline. This heterogeneity, in turn, complicates the planning of clinical research, particularly trials of disease-modifying therapies, as well as the care of PD patients. Work has been done to review the present role of genetic and biochemical biomarkers in PD.[16]. There has been mention of use of biomarkers for clinical trial planning, as well as clinical care through the application of a “precision medicine” approach[16].

Intrinsically disordered proteins has already been identified as the main reason behind the various neurodegenerative diseases.[9]Substantial work has linked α-synucleic, a small highly conserved presynaptic protein with unknown function, as a intrinsically disordered protein as a major components of PD.[17].

Chapter 3: Proposed Methodology

3. Proposed Methodology

In this article, we perform an evolutionary study on the common biomarkers of MMAD, ESPD and MS by comparing and analyzing the effect on the active regions due to structure disorder. In this regard, we identify the differentially functioned autoantibodies common to the diseases and perform sequence-based analysis as well as structure-based analysis separately thus mapping the evolutionary coupled pair to the structure network. Finally, the disordered regions, active regions and the structure networks are analyzed to reach the conclusion. Each step is described in details below.

3.1 Differentially expressed protein identification

Here we use statistical test to identify the differentially expressed proteins for each of the disease MMAD,ESPD and MS. T-test, one of the most popular tests of the kind, is used for this purpose. It is one type of inferential statistics that is used to determine whether there is a significant difference between the means of two groups.

We can categorize t-test as:

i. *One-sample t-test*
ii. *Two-sample t-test*
iii. *Paired t-test*

One Sample t-test or single sample t-test determines whether the sample mean is statistically different from a known or hypothesized population.

Paired t-test is nothing but a variation of single sample t-test. It simply calculates the difference between paired observations (e.g., before and after) and then performs a 1-sample t-test on the differences.

3.1.1 Two sample t-test

In our problem, we are performing two-sample t-test on the normalized expression data. The **two-sample t-test** takes the sample data from two groups and boils it down to the t-value. Though being quite similar to one- sample t-test it differs in the sense that two independent group of data is required in 2-sample t-test.

In our experiment the two independent samples for each disease are:

(i)Controlled data

(ii)Diseased data

Two-sample t-test can be done using either equal variance or unequal variance depending on which is more appropriate for the dataset under consideration. Here we have assumed unequal variance to obtain our result.

When the two independent samples are assumed to be drawn from populations with unequal variances (i.e., $\sigma_1^2 \neq \sigma_2^2$), the test statistic '*t*' is computed as:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1} - \frac{{s_2}^2}{n_2}}}$$

Where,

t = test statistics or t-value

$\overline{x_1}$ = Mean of first sample

$\overline{x_2}$ = Mean of second sample

n_1 = Sample size of first sample

n_2 = Sample size of second sample

s_1 = Standard deviation of first sample

s_2 = Standard deviation of second sample

The **t-value** is simply the calculated difference represented in units of standard error. The greater the magnitude of t (it can be either positive or negative), the greater the evidence *against* the null hypothesis that there is no significant difference. The closer T is to zero, the more likely there is not a significant difference.

3.1.2 Filtering based on p-value

The **p-value** is the probability of observing a t-value as large as or larger than the actual observed t-value in which the null hypothesis is considered as true.

To determine whether the difference between the sample means is statistically significant, the p-value is compared to the significance level. Usually a standard significance level (denoted as alpha) of 0.05 is used. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

P-value ≤ α: The difference between the means is statistically significant (Reject H_0): If the p-value is less than or equal to 0.05 the decision is to reject the null hypothesis. Henceforth concluding the sample means are statistically significant. Therefore since we are searching for differently expressed proteins we select those with p-value<=0.05 indicating that they are differently expressed.

P-value > α: The difference between the means is not statistically significant (Fail to reject H_0): If the p-value is greater than the significance level, the decision is to fail to reject the null hypothesis. Similarly in our dataset we exclude those with p-value>0.05 since they are not differently expressed.

3.2 Identification of differentially functioned autoantibodies common to the diseases

The differently functioned autoantibodies common to the considered diseases can be found out quite simply by identifying the common proteins in the three datasets that is obtained after filtering using p-value.

Here the comparison between the three sets is done using Venn-diagram. A Venn diagram is a diagram that shows all possible logical relations between a finite collections of different sets. If we consider the set of DE for each disease as separate sets then simple intersection between the sets will produce the desired result.

Venny is used to perform this Venn diagram operation. VENNY is an interactive tool, which is used for comparing lists with Venn Diagrams.

Thus our required result is

$$X = ESPD \cap MMAD \cap MS$$

Where,

ESPD: differentially expressed proteins of ESPD

MMAD: differentially expressed proteins of ESPD

MS: differentially expressed proteins of MS

The dataset thus obtained will be henceforth considered for further analysis.

3.3 Study of Disordered Regions

An intrinsically disordered protein (IDP) is a protein that lacks a fixed or stable three-dimensional structure. IDPs constitute a very large and functionally important class of protein. In some cases, IDPs can binds with other macromolecules to adopt a three dimensional structure.

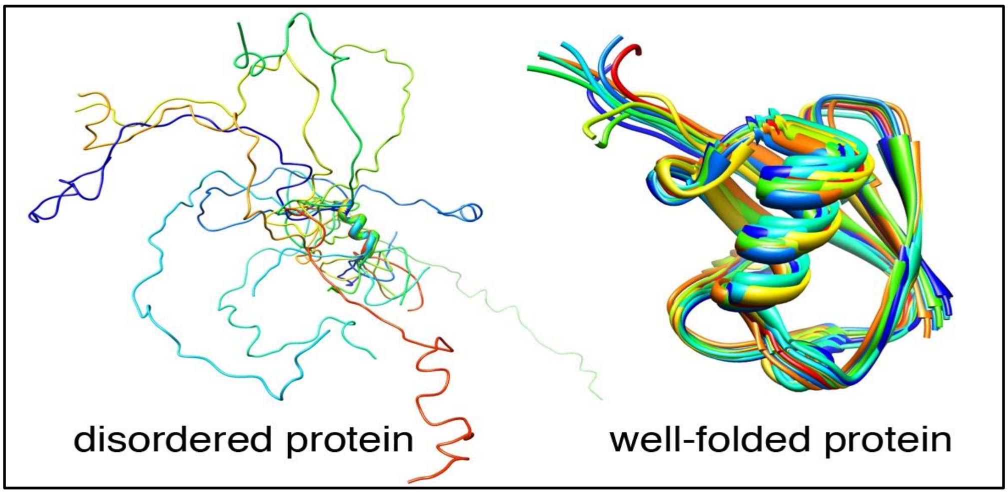


Fig 3.1: Structural difference between ordered and disordered protein

Disorder is mostly found in intrinsically disordered regions (IDRs) within an otherwise well-structured protein. Thus the term intrinsically disordered protein (IDP) therefore includes proteins that contain IDRs as well as fully disordered proteins.

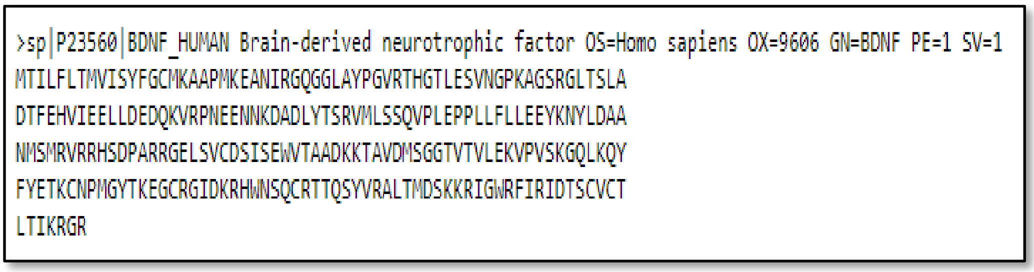
D2P2 (Database of Disordered Protein Prediction)[11] is a community resource for pre-computed disorder predictions on a large library of proteins from completely sequenced genomes. D2P2 use various predictors to list all the disordered region of a particular protein along with their start and end in the corresponding protein sequence. Among all the listed regions the maximum length disordered region is found and its corresponding start and end in the protein sequence is noted to identify the selected disordered region in the protein sequence. The disordered region thus obtained is used later for structural analysis using structure network, which is elaborated in later steps.

Another online tool, Pondr, is used to predict the overall percentage disorder of each of the IDP. **PONDR** (Predictor of Naturally Disordered Regions) is an online that works only on single sequence to produce an accurate percentage of disorder of the submitted protein. Here we make use of the VL-XT algorithm (PONDR VL-XT) which refers to the merger of three predictor, one trained on *V*ariously characterized *L*ong disordered regions and two trained on *X*-ray characterized *T*erminal disordered regions. VL-XT outputs are real numbers between one and zero, where 1 is the ideal prediction of disorder and 0 is the ideal prediction of order. VL-XT outputs are typically not ideal and a threshold is applied with disorder assigned to values greater than or equal to 0.5.

3.4 Sequence based analysis

Protein primary structure is the linear sequence of amino acids in a peptide or protein. Each protein can be represented using a FASTA sequence for analysis. In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The simplicity of FASTA sequence makes it easier to manipulate and parse using scripting language such as R, Perl, Ruby, python. We will be using Matlab for all the manipulation and analysis purpose.

In modern bioinformatics, the sequence header is preceded by '>'. Following the initial line (used for a unique description of the sequence) is the actual sequence itself in standard one-letter character string.



```
>sp|P23560|BDNF_HUMAN Brain-derived neurotrophic factor OS=Homo sapiens OX=9606 GN=BDNF PE=1 SV=1
MTILFLTMVISYFGCMKAAPMKEANIRGQGGLAYPGVRTHGTLESVNGPKAGSRGLTSLA
DTFEHVIEELLDEDQKVRPNEENNKDADLYTSRVMLSSQVPLEPPLLFLLEEYKNYLDAA
NMSMRVRRHSDPARRGELSVCDSISEWVTAADKKTAVDMSGGTVTVLEKVPVSKGQLKQY
FYETKCNPMGYTKEGCRGIDKRHWNSQCRTTQSYVRALTMDSKKRIGWRFIRIDTSCVCT
LTIKRGR
```

Fig 3.2: A sample FASTA sequence

3.4.1 Identification of seed alignment of protein families

Each disordered protein belongs to a specific protein family. Our goal here is to find the seed alignment from which the family is built. The seed alignment, which will then be used for evolutionary study to reach the conclusion.

3.4.1.1 Finding the disordered protein's families

The aid of various protein databases available online is used to identify the families of the various disordered proteins. Protein databases are used for generating huge amounts of data for protein structures, functions, and particularly sequences. Uniprot and Pfam are such databases, which we have used here to obtain the desired sequences.

UniProt is a freely accessible structural database of protein sequence and functional information where many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. In Uniprot, each entry is identified with a unique identification ID. For each disordered protein, the corresponding Uniprot entry gives us detailed knowledge about the protein structure and about the family to which it belong. From there we collect the Pfam entry ID which can then be used to collect further knowledge about the family and to identify the seed alignment of the said family.

3.4.1.2 Identifying the seed alignment using Pfam

Pfam is another popular freely accessible online protein database that intends to provide a complete and accurate classification of protein families and domains. Along with storing the full alignment, Pfam stores the seed alignment from which the families are built which is generated by searching the sequence database.

The seed alignment thus obtained is morphed into FASTA format for easier manipulation and evolutionary study using MATLAB.

3.4.2 Direct Coupled Analysis (DCA)

Direct coupling analysis or DCA is an umbrella term that comprised of various methods for analyzing sequence data. The general idea is to use statistical modelling that quantifies the strength of direct relationship between two positions of biological sequence. The effects from other positions are not taken into consideration.

DCA is not same as correlation since correlation could be high even if there is no direct relationship between the positions. Hence the name direct coupling analysis. It act as an evolutionary pressure for two positions to maintain mutual compatibility in bio molecular structure of the sequence, leading to molecular co evolution between the two positions. DCA has been used in the inference of protein residue contacts,[18][19][20][21] RNA structure prediction,[22][23] the inference of protein-protein interaction networks[24][25].

Here we will use a computationally efficient implementation of DCA [26] for further analysis.

3.5 Structure Based Analysis

Here we simply use the protein structure to develop a structured network which be used for further studies. Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. To be able to perform their biological function, proteins fold into one or more specific spatial conformations driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions etc.

There are four distinct level of protein structure which we will discuss here in short.

Primary Protein Structure: The primary structure of a protein refers to the sequence of amino acids in the polypeptide chain. The primary structure is held together by peptide bonds. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity.

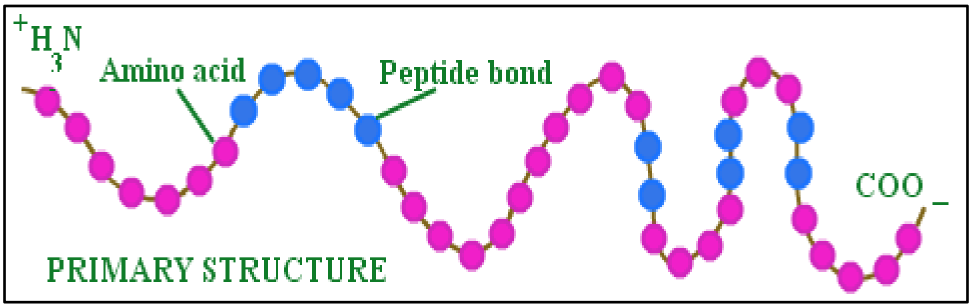


Fig 3.3: Primary protein structure consisting of peptide bonds

Secondary Protein Structure: Secondary structure refers to highly regular local sub-structures on the actual polypeptide backbone chain. Two main types of secondary structure, the α-helix and the β-strand or β-sheets, were suggested in 1951 by Linus Pauling and co-workers[27].These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. Both the α-helix and the β-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone

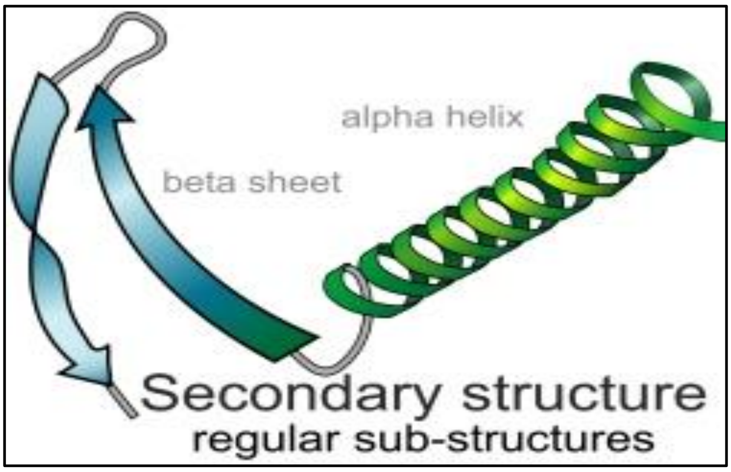


Fig 3.4: Secondary protein structure consisting of alpha helix and beta sheets

Tertiary Protein Structure: Tertiary structure refers to the three-dimensional structure of monomeric and multimeric protein molecules. The α-helixes and β-pleated-sheets are folded into a compact globular structure.

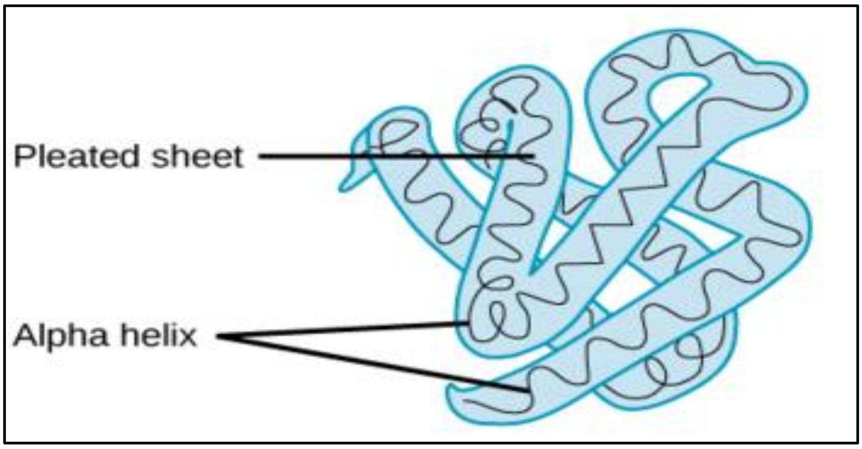


Fig 3.5: Tertiary or three dimensional protein structures

Quaternary Protein Structure: Quaternary structure is the three-dimensional structure consisting of the aggregation of two or more individual polypeptide chains (subunits) that operate as a single functional unit. In this context, the quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure.

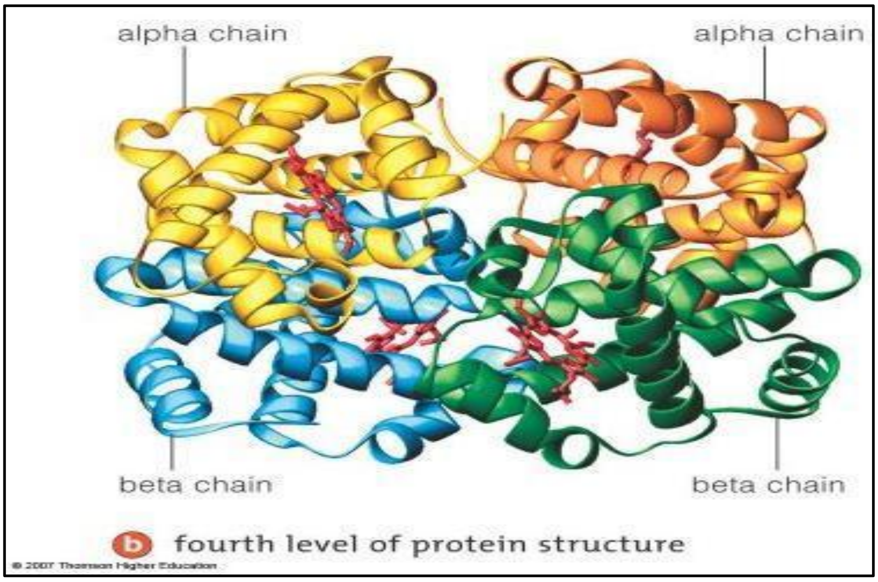

alpha chain
alpha chain
beta chain
beta chain
b fourth level of protein structure
© 2007 Thomson Higher Education

Fig 3.6: Quaternary protein structure

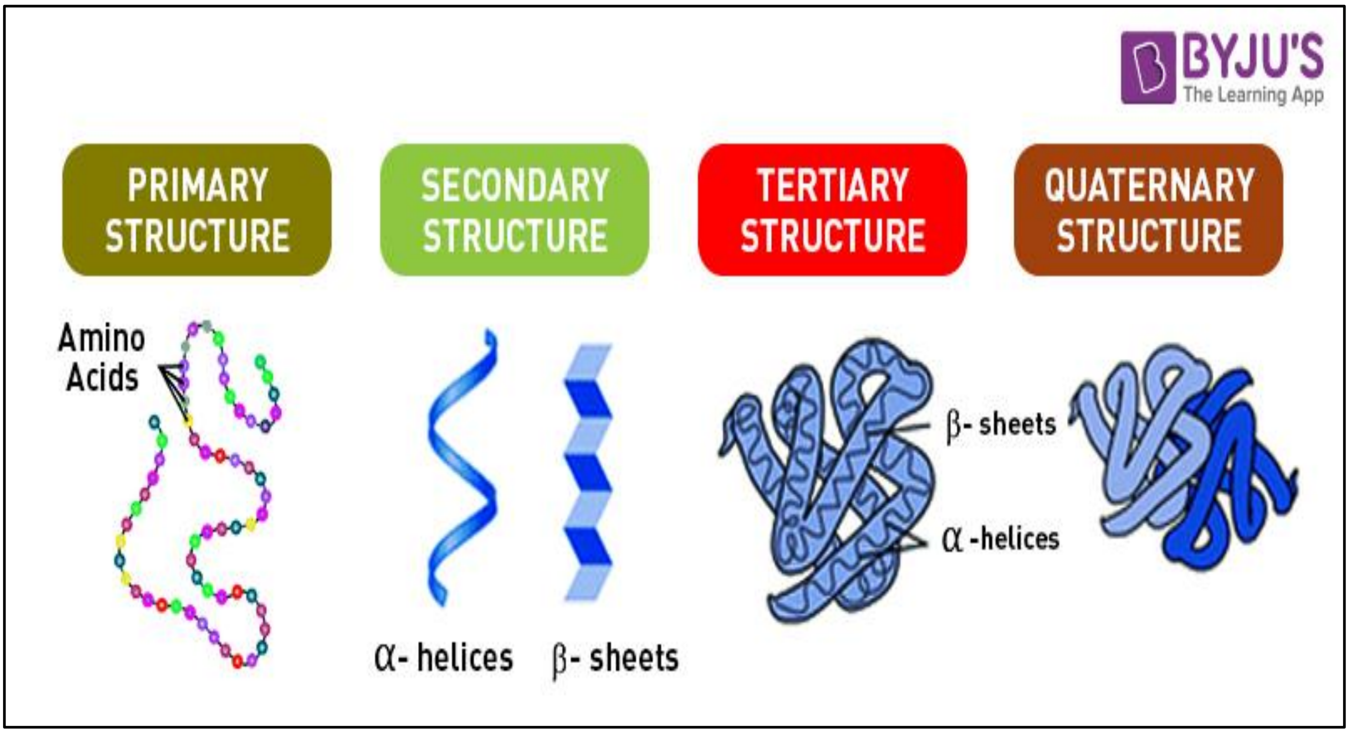


Fig 3.7: Showing the different protein structures

3.5.1 Obtaining three dimensional protein structure

Another freely accessible online database PDB or Protein Data Bank is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The **PDB** archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. We will be using the PDB ID to obtain the three dimensional structure of the disordered proteins which will be used to obtain the structure network of the same.

3.5.2 Structure Network Analysis

For creating and analysing the protein structure network, we use the Bio3D package available in R version 2.1 or above. To obtain the structural network we first perform a normal mode analysis using the nma( ) function followed by dynamic cross correlation analysis by dccm( ) and finally network analysis using cna( ).

The network thus obtained is then plotted to obtain an all residue network and a simplified community network. Finally, for each network obtained, the module number containing the disordered region is identified.

3.6 Analysis by comparing the regions specific to the effect of structural disorder

Active site of an enzyme is the region where the substrate molecules bind and undergo chemical region. Hence the active site, even though being small compared to the while volume of the enzyme, is easily the most important part, since it directly catalyzes the reaction. Active site generally consists of three to four amino acids while the other amino acids are used to maintain the tertiary or three-dimensional structure of the enzyme.

Hence, if the disordered regions directly or indirectly influence the active site, the effect would be huge. Here we will be performing a comparative study to analyze how the disordered region can affect the active site.

3.7 Workflow

A detailed workflow diagram is given below to provide a better visualization and summarization of the working methodology proposed above.

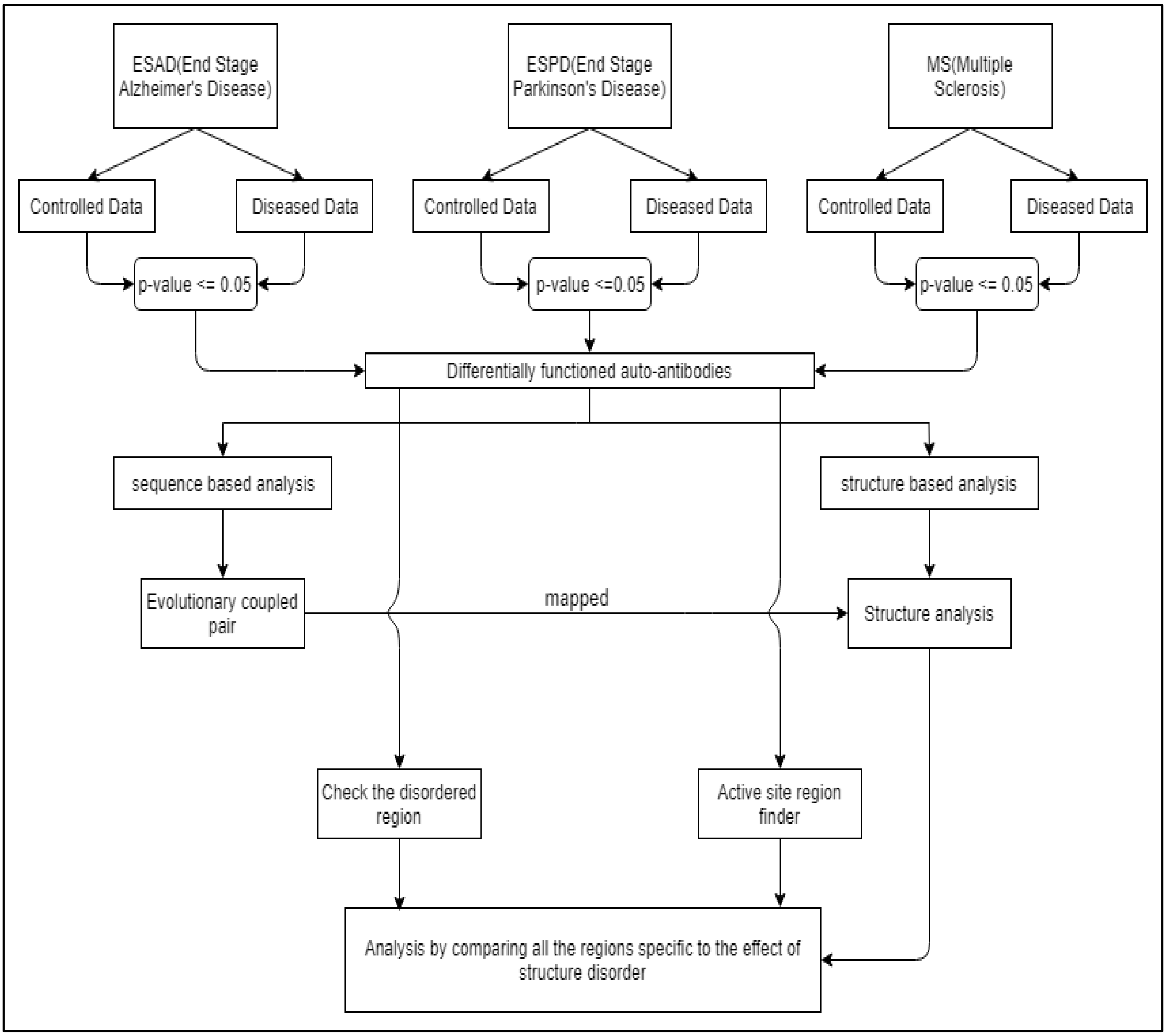

ESAD(End Stage Alzheimer's Disease)
ESPD(End Stage Parkinson's Disease)
MS(Multiple Sclerosis)
Controlled Data
Diseased Data
Controlled Data
Diseased Data
Controlled Data
Diseased Data
p-value <= 0.05
p-value <=0.05
p-value <= 0.05
Differentially functioned auto-antibodies
sequence based analysis
structure based analysis
Evolutionary coupled pair
mapped
Structure analysis
Check the disordered region
Active site region finder
Analysis by comparing all the regions specific to the effect of structure disorder

Fig 3.8: Workflow diagram

Chapter 4:

Results

&

Discussion

4. RESULTS AND DISCUSSION

As the active site of an enzyme directly catalyzes a chemical reaction, a disordered active site can have adverse affect on the functionality of a protein. Presence of an anomaly can be determined if any of following scenario occurs:

If the active site is present in the disordered cluster, it is evident that any reaction involving that site will have an even worse effect. However, even if the active site is not directly present in any of the disordered cluster there is still a chance of that active site be affected if there exists a direct relationship between the two.

In this article, we will use an example of six disordered proteins, to elaborate on the idea, with three having disordered active site and the other three having ordered active site.

4.1Extraction of common autoantibodies from 3 different datasets

Initially we have taken into consideration three separate dataset, each for one of the disease mentioned above. For all the three diseased data we will be using MCI as the controlled data and the respective disease as the diseased data for t-testing.

As mentioned earlier we perform two sample t- testing on each of the entry assuming unequal variance. Based on the p-value obtained each entry is either discarded or retained if p-value> 0.05 or p-value <= 0.05. Since it is not possible to provide the entire database, we will show using 2 example how we are filtering the data using p-value.

4.1.1. Dataset for MMAD

Number of entries: 9480

Example 1

Protein name: neural cell adhesion molecule 2 (NCAM2)

Controlled Data	Diseased Data
-1.629503361	103.8089788
-0.547586045	80.22555908
1.045518636	238.2094571

Controlled Data	Diseased Data
0.483219114	230.0085173
0	133.5957145
-1.414667878	-2.381184773
-1.6214874	-2.186430658
-1.250977791	0.79784808
-1.25687334	-0.410706708
-0.620759628	1.153338789
-2.797713257	0.968961174
-2.29711062	0
1.048635796	18.77023714
-1.44743922	-2.197322416
-2.275176781	-1.350164334
-1.056889552	1.101563389
-1.368781322	-1.340055047
-1.136510755	-1.46790729
-1.0188194	-1.689246969
-0.740536723	-0.783954498
-0.831924195	-0.64404913
-0.673451793	0
-0.625726756	-0.680942524
-0.559583112	-0.733656047
-1.230737611	145.0644704

TABLE 1: Showing controlled and diseased data for NCAM2

After t testing the obtained **p-value** is **0.01557622** which is less than 0.05. Therefore, it is discarded.

<u>Example 2</u>

Protein name: leucine-rich repeats and IQ motif containing 2 (LRRIQ2)

Data:

Controlled Data	Diseased Data
402.0635638	766.89282
791.6289157	890.85241
755.2593367	480.12031
498.7298869	1167.603
1242.237061	412.61526
616.1318158	1270.9743
251.7226506	1001.7443
606.6492422	3116.4574
209.877834	658.43808
995.1069432	1035.0692

Controlled Data	Diseased Data
488.4203622	1470.2857
325.7810346	582.89548
276.3308396	740.18297
707.8809678	548.437
567.2752605	754.87936
437.0994194	558.04657
1145.954448	939.16562
440.3917301	797.01635
487.074834	432.69764
1178.984346	345.72243
379.0751264	777.94495
537.9388964	280.41852
358.4026807	427.22255
1214.483094	1800.8112
215.0854489	374.31578

TABLE 2 : Showing controlled and diseased data for LRRIQ2

After t-testing the obtained **p-value** is **0.063260099** which is greater than 0.05. Therefore, it is retained.

In this way entire database is filtered. After filtration is complete, the number of proteins obtained is: 1863

4.1.2. Dataset for ESPD

Number of entries: 9480

Here also we apply the same protein filtering process using p-value. After filtration is complete, the number of proteins obtained is: 1628

4.1.3. Dataset for MS

Number of entries: 9480

Here also we apply the same protein filtering process using p-value. After filtration is complete, the number of proteins obtained, is 517.

4.1.4. Common protein identification using Venny

Finally, we extract the common protein among all the three datasets after filtering, for further analysis. Venny, an online tool for Venn diagram drawing is used for the purpose.

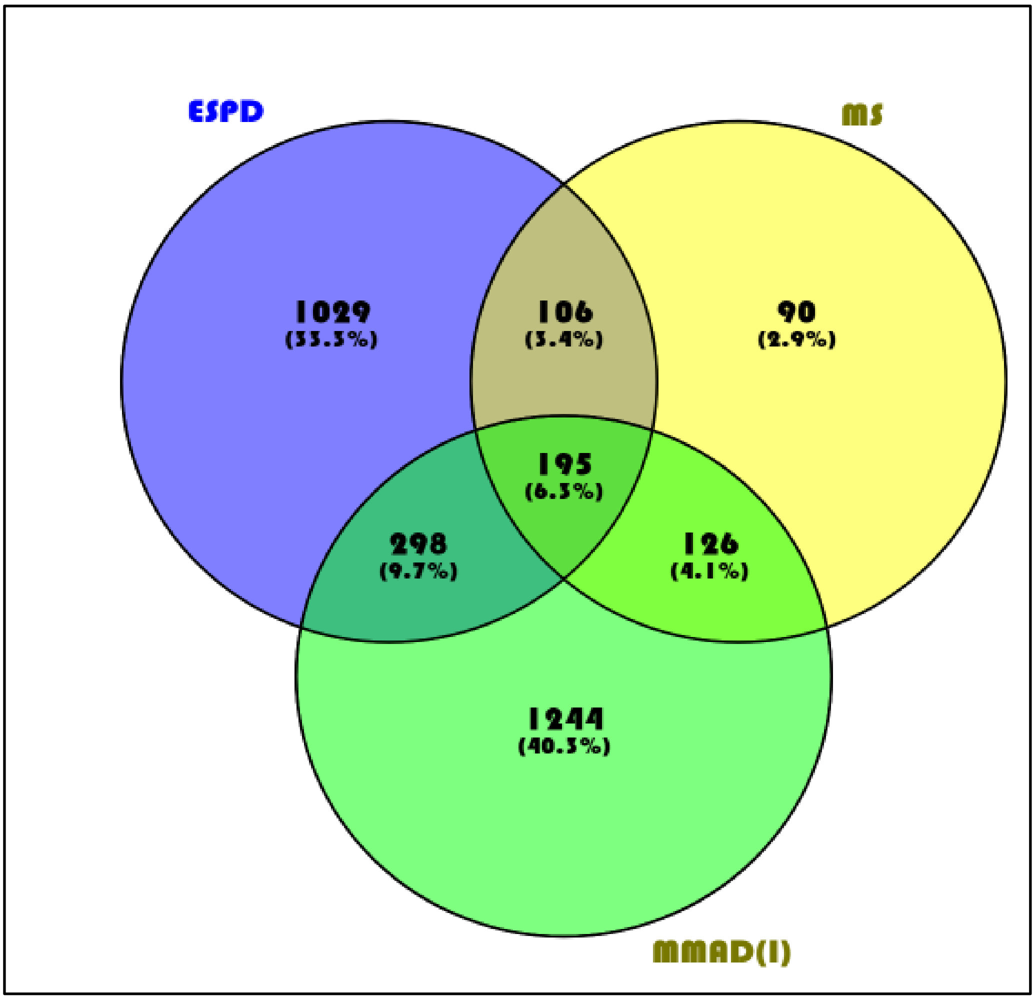


Fig 4.1: Common proteins between ESPD MMAD and MS

The common proteins thus obtained are given below:

195 COMMON ELEMENT IN ESPD, MS AND MMAD
TG F-beta receptor type-2
Up stream stimulatory factor 2
Aldose reductase
brain-derived neurotrophic factor (BDNF)
POU domain, class 5, transcription factor 1
Transaldolase
Bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 1

195 COMMON ELEMENT IN ESPD, MS AND MMAD
leucine rich repeat containing 8 family, member D (LRRC8D)
tocopherol (alpha) transfer protein (TTPA)
Isovaleryl-CoA dehydrogenase, mitochondrial
Centromere protein L
SFRS protein kinase 1 (SRPK1)
Death-associated protein kinase 2
ubiquitin-conjugating enzyme E2 variant 1 (UBE2V1)
hematopoietic SH2 domain containing (HSH2D)
Granulocyte-macrophage colony-stimulating factor
Serine/threonine-protein kinase Pim-2
tec protein tyrosine kinase (TEC)
Reticulon-4-interacting protein 1
Bifunctional polynucleotide phosphatase/kinase
neural cell adhesion molecule 2 (NCAM2)
Ephrin type-A receptor 2
Phosphorylase b kinase gamma catalytic chain, testis/liver isoform
translocase of inner mitochondrial membrane 8 homolog B (yeast) (TIMM8B)
THAP domain containing 4 (THAP4)
tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator) (TNFRSF14)
v-raf murine sarcoma viral oncogene homolog B1 (BRAF
Serine/threonine-protein kinase 36
glutathione S-transferase omega 2 (GSTO2)
fms-related tyrosine kinase 3 ligand (FLT3LG)
Ral GEF with PH domain and SH3 binding motif 1 (RALGPS1)
mitogen-activated protein kinase kinase 6 (MAP2K6), transcript variant 2; see catalog number for detailed information on wild-type or point mutant status
carbonic anhydrase I (CA1)
Ig gamma-1 chain C region
Tubulin-specific chaperone D
coiled-coil domain containing 100 (CCDC100)
glycogen synthase kinase 3 beta (GSK3B)
albumin (ALB)
glucagon receptor (GCGR)
Cell division cycle 2-related protein kinase 7
proteasome (prosome, macropain) subunit, beta type, 1 (PSMB1)
COP9 signalosome complex subunit 1
aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II) (AKR1C3)
ribosomal protein L11 (RPL11)
Serine/threonine-protein kinase PAK 1
LATS, large tumor suppressor, homolog 1 (Drosophila) (LATS1)
MAP/microtubule affinity-regulating kinase 4
T-cell immunoglobulin and mucin domain containing 4 (TIMD4)
haptoglobin (HP)

195 COMMON ELEMENT IN ESPD, MS AND MMAD
prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase) (PTGS2)
proteasome (prosome, macropain) 26S subunit, ATPase, 1 (PSMC1)
mitogen-activated protein kinase kinase kinase 5 (MAP3K5)
zinc finger, CCHC domain containing 6 (ZCCHC6)
RING finger and WD repeat domain-containing protein 3
Small EDRK-rich factor 1
F-box and leucine-rich repeat protein 16 (FBXL16)
coiled-coil domain containing 89 (CCDC89)
vang-like 1 (van gogh, Drosophila) (VANGL1)
serine active site containing 1 (SERAC1)
chromosome 16 open reading frame 5 (C16orf5)
interleukin 20 (IL20)
Interleukin-19
Protein jagunal homolog 1
G antigen 7 (GAGE7)
Olfactory receptor 3A1
transmembrane protein 43 (TMEM43)
immunoglobulin lambda variable 2-14 (IGLV2-14)
homeobox C4 (HOXC4)
RNA pseudouridylate synthase domain containing 2 (RPUSD2)
Uncharacterized protein C20orf96
neurogenic differentiation 6 (NEUROD6)
isoleucyl-tRNA synthetase 2, mitochondrial (IARS2)
cyclin B2 (CCNB2)
uroplakin 1A (UPK1A)
Serine/threonine-protein kinase Sgk2
maelstrom homolog (Drosophila) (MAEL)
stathmin-like 3 (STMN3)
homeobox B5 (HOXB5)
RAD23 homolog A (S. cerevisiae) (RAD23A)
immunoglobulin kappa variable 1-5 (IGKV1-5)
PRKR interacting protein 1 (IL11 inducible) (PRKRIP1)
gastrokine 2 (GKN2)
periphilin 1 (PPHLN1)
sciellin (SCEL)
Fc fragment of IgG, low affinity IIIa, receptor (CD16a) (FCGR3A)
homeobox B7 (HOXB7)
pseudouridylate synthase-like 1 (PUSL1)
UL16 binding protein 1 (ULBP1)
mitochondrial ribosomal protein S18B (MRPS18B)
butyrophilin, subfamily 2, member A1 (BTN2A1), transcript variant 2
Protein FAM113B

195 COMMON ELEMENT IN ESPD, MS AND MMAD
V-set and immunoglobulin domain containing 1 (VSIG1)
putative homeodomain transcription factor 2 (PHTF2)
four and a half LIM domains 1 (FHL1)
Rab9 effector protein with kelch motifs (RABEPK)
purinergic receptor P2Y, G-protein coupled, 2 (P2RY2)
CaM kinase-like vesicle-associated (CAMKV)
Ets2 repressor factor (ERF)
Neuroplastin
zinc finger protein 44 (ZNF44)
leucine rich repeat containing 42 (LRRC42)
1-acylglycerol-3-phosphate O-acyltransferase 4 (lysophosphatidic acid acyltransferase, delta) (AGPAT4)
interferon-induced protein with tetratricopeptide repeats 3 (IFIT3)
phosphodiesterase 7B (PDE7B)
ATPase, class II, type 9B (ATP9B)
dystrophia myotonica-protein kinase (DMPK), transcript variant 2
minichromosome maintenance complex component 5 (MCM5)
Putative E3 ubiquitin-protein ligase SH3RF2
cleavage and polyadenylation specific factor 3, 73kDa (CPSF3)
glycogen synthase 1 (muscle) (GYS1)
neuroligin 4, Y-linked (NLGN4Y)
signal transducer and activator of transcription 6, interleukin-4 induced (STAT6)
ADAM metallopeptidase domain 2 (fertilin beta) (ADAM2)
SLIT and NTRK-like protein 3
ADAMTS-like 1 (ADAMTSL1)
Dynein heavy chain 6, axonemal
immunoglobulin lambda constant 1 (Mcg marker) (IGLC1)
Ig lambda chain C regions
neuroblastoma breakpoint family, member 1 (NBPF1)
anthrax toxin receptor 1 (ANTXR1), transcript variant 3
RAB40B, member RAS oncogene family (RAB40B)
DEAD (Asp-Glu-Ala-Asp) box polypeptide 20 (DDX20)
hypothetical protein MGC31957 (MGC31957)
centrosomal protein 72kDa (CEP72)
tumor suppressing subtransferable candidate 4 (TSSC4)
septin 4 (SEPT4), transcript variant 1
EMG1 nucleolar protein homolog (S. cerevisiae) (EMG1)
polypyrimidine tract binding protein 1 (PTBP1), transcript variant 3
N-acylsphingosine amidohydrolase (acid ceramidase) 1 (ASAH1), transcript variant 1
neuro-oncological ventral antigen 1 (NOVA1), transcript variant 1
septin 10 (SEPT10), transcript variant 1
potassium voltage-gated channel, shaker-related subfamily, beta member 2 (KCNAB2), transcript variant 1
ERO1-like beta (S. cerevisiae) (ERO1LB)

195 COMMON ELEMENT IN ESPD, MS AND MMAD
XIAP associated factor-1 (XAF1), transcript variant 1
RNA/RNP complex-1-interacting phosphatise
immunoglobulin heavy constant mu (IGHM)
Ig mu chain C region
piccolo (presynaptic cytomatrix protein) (PCLO)
Trans-2-enoyl-CoA reductase, mitochondrial
peroxisomal trans-2-enoyl-CoA reductase (PECR)
olfactory receptor, family 51, subfamily E, member 1 (OR51E1)
YTH domain family, member 2 (YTHDF2)
immunoglobulin superfamily, member 1 (IGSF1), transcript variant 2
centaurin, delta 2 (CENTD2), transcript variant 1
Wolf-Hirschhorn syndrome candidate 1 (WHSC1), transcript variant 5
hypothetical protein BC014011 (LOC116349)
homeobox A1 (HOXA1), transcript variant 1
cDNA clone MGC:22645 IMAGE:4700961, complete cds
PHD finger protein 17 (PHF17), transcript variant S
unkempt homolog (Drosophila)-like (UNKL)
ADP-ribosylation factor GTPase activating protein 1 (ARFGAP1), transcript variant 2
mitochondrial GTPase 1 homolog (S. cerevisiae) (MTG1)
SLAIN motif family, member 1 (SLAIN1), transcript variant 2
immunoglobulin lambda locus (IGL@)
PREDICTED: Homo sapiens hypothetical LOC150371 (LOC150371)
cDNA clone MGC:31944 IMAGE:4878869, complete cds
mitochondrial ribosomal protein L19 (MRPL19), nuclear gene encoding mitochondrial protein
sorcin (SRI), transcript variant 1
cDNA clone MGC:31936 IMAGE:4765518, complete cds
cDNA clone MGC:27152 IMAGE:4691630, complete cds
cDNA clone MGC:27376 IMAGE:4688477, complete cds
nudix (nucleoside diphosphate linked moiety X)-type motif 16-like 1 (NUDT16L1)
potassium voltage-gated channel, shaker-related subfamily, beta member 1 (KCNAB1), transcript variant 1
mesenchyme homeobox 1 (MEOX1), transcript variant 1
immunoglobulin heavy constant gamma 3 (G3m marker) (IGHG3)
potassium voltage-gated channel, shaker-related subfamily, member 1 (episodic ataxia with myokymia) (KCNA1)
non-SMC condensin II complex, subunit H2 (NCAPH2), transcript variant 1
mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase, isozyme B (MGAT5B)
vacuolar protein sorting 53 homolog (S. cerevisiae) (VPS53)
activin A receptor, type IB (ACVR1B), transcript variant 1
DEAD (Asp-Glu-Ala-Asp) box polypeptide 47 (DDX47), transcript variant 1
cDNA clone IMAGE:4903661, complete cds
Recombinant human CTLA-4/Fc
hypothetical protein MGC26647 (MGC26647)
chromosome 2 open reading frame 47 (C2orf47)

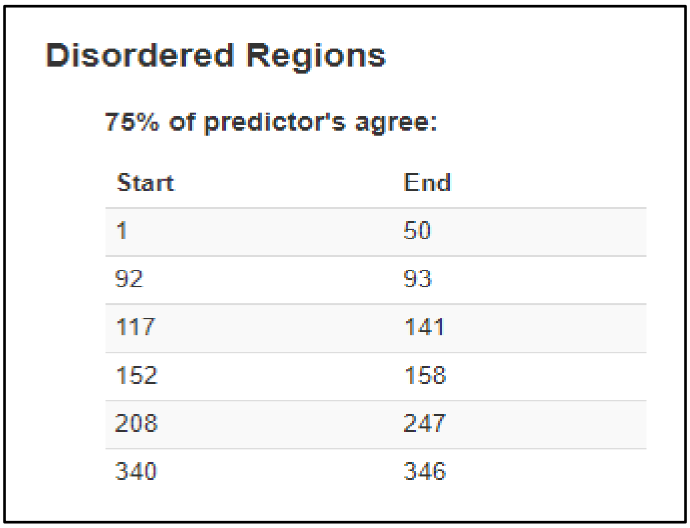
195 COMMON ELEMENT IN ESPD, MS AND MMAD
neutrophil cytosolic factor 4, 40kDa (NCF4), transcript variant 1
NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa (NDUFB6), nuclear gene encoding mitochondrial protein, transcript variant 2
cyclin B1 interacting protein 1 (CCNB1IP1), transcript variant 1
deleted in a mouse model of primary ciliary dyskinesia (RP11-529I10.4)
cDNA clone MGC:32654 IMAGE:4701898, complete cds
glutaryl-Coenzyme A dehydrogenase (GCDH), nuclear gene encoding mitochondrial protein, transcript variant 1
RAB43, member RAS oncogene family (RAB43)
RAP1, GTP-GDP dissociation stimulator 1 (RAP1GDS1), transcript variant 5, mRNA.
sphingosine kinase 1 (SPHK1), transcript variant 1
PREDICTED: Homo sapiens hypothetical protein LOC145842 (LOC145842)
CSAG family, member 3A (CSAG3A)
processing of precursor 7, ribonuclease P/MRP subunit (S. cerevisiae) (POP7)
olfactory receptor, family 11, subfamily L, member 1 (OR11L1), mRNA
Cellular nucleic acid-binding protein
cDNA clone IMAGE:3351130, complete cds
CAP-GLY domain containing linker protein family, member 4 (CLIP4)
cDNA clone MGC:23888 IMAGE:4704496, complete cds
mitochondrial ribosomal protein S16 (MRPS16), nuclear gene encoding mitochondrial protein
v-akt murine thymoma viral oncogene homolog 1 (AKT1), transcript variant 3

TABLE 3: Common proteins between ESPD MMAD and MS

4.2. Identification of maximum disordered sequence

To identify the disordered sequence of maximum length D2P2 is used. For a particular protein, D2p2 provides all the agreed upon disordered region as given by the various predictor algorithm used. Among all those regions the maximum length disordered sequence is identified is using its sequence start and end length.

For better understanding, let us consider the protein ‘*Upstream stimulatory factor 2’* with the Uniprot ID ‘Q15853’, the agreed upon various disordered region as found out by the predictors are:



Disordered Regions

75% of predictor's agree:

Start	End
1	50
92	93
117	141
152	158
208	247
340	346

Fig 4.2: Showing the agreed upon disordered regions of '*Upstream stimulatory factor 2'*

Among these, it can be easily spotted that the sub-sequence starting at '1' and ending at '50' is the required maximum disordered length sequence. In this way, all the other proteins are similarly evaluated to find their corresponding maximum length disordered sequence.

Protein name	UNIPROT ID	SEQUENCE LENGTH	DISORDER SEQUENCE START	DISORDER SEQUENCE END
Aldose reductase	P15121	316	221	225
Bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 1	P52848	882	50	58
SFRS protein kinase 1 (SRPK1)	Q96SB4	655	235	353
tec protein tyrosine kinase (TEC)	P42680	631	157	179
glutathione S-transferase omega 2 (GSTO2)	Q9H4Y5	243	1	15
mitogen-activated protein kinase kinase 6 (MAP2K6), transcript variant 2	P52564	334	1	35

TABLE 4: Showing the maximum disordered length of the considered proteins

4.3 Structure based analysis

4.3.1. Obtaining structure network from three dimensional structure

The first step for performing structure is to obtain the three dimensional structure of the protein. The structure thus obtained is then used to produce a full residue structure network.

We will be using the PDB ID to obtain the three dimensional structure of the disordered proteins which will then be used to obtain the structure network of the same.

UNIPROT ID	PDB ID
P15121	1AZ2
P52848	1NST
Q96SB4	1WBP
P42680	2LUL
Q9H4Y5	3QAG
P52564	3VN9

TABLE 5: Retrieving the PDB ID of the proteins

The PDB structure and the corresponding structure network for each of the 10 proteins has is given below.

a) Aldose reductase

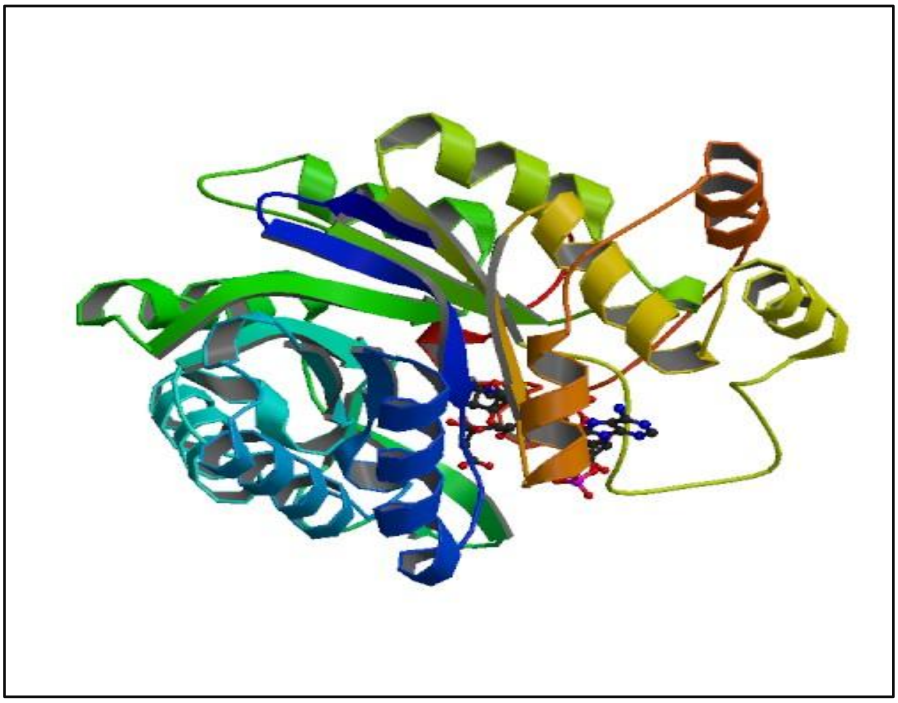


Fig 4.3a : Three dimensional structure of aldose reductase with PDB ID 1AZ2[28]

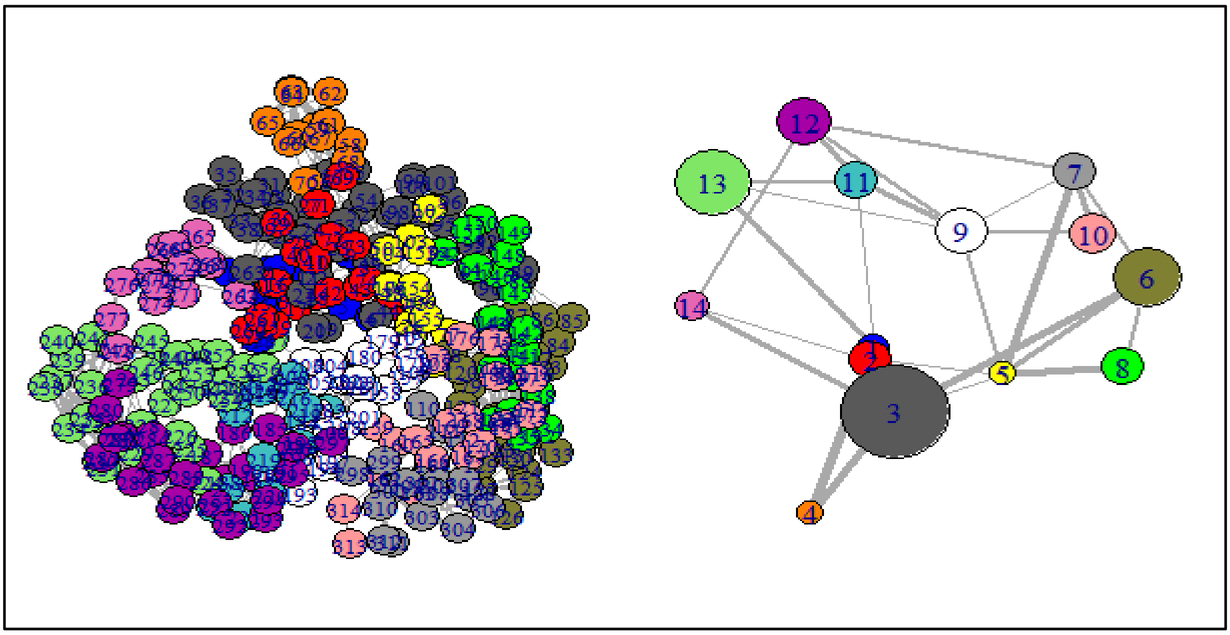


Fig 4.3b : All residue network and simple community network of 1AZ2

b) SFRS protein kinase 1 (SRPK1)

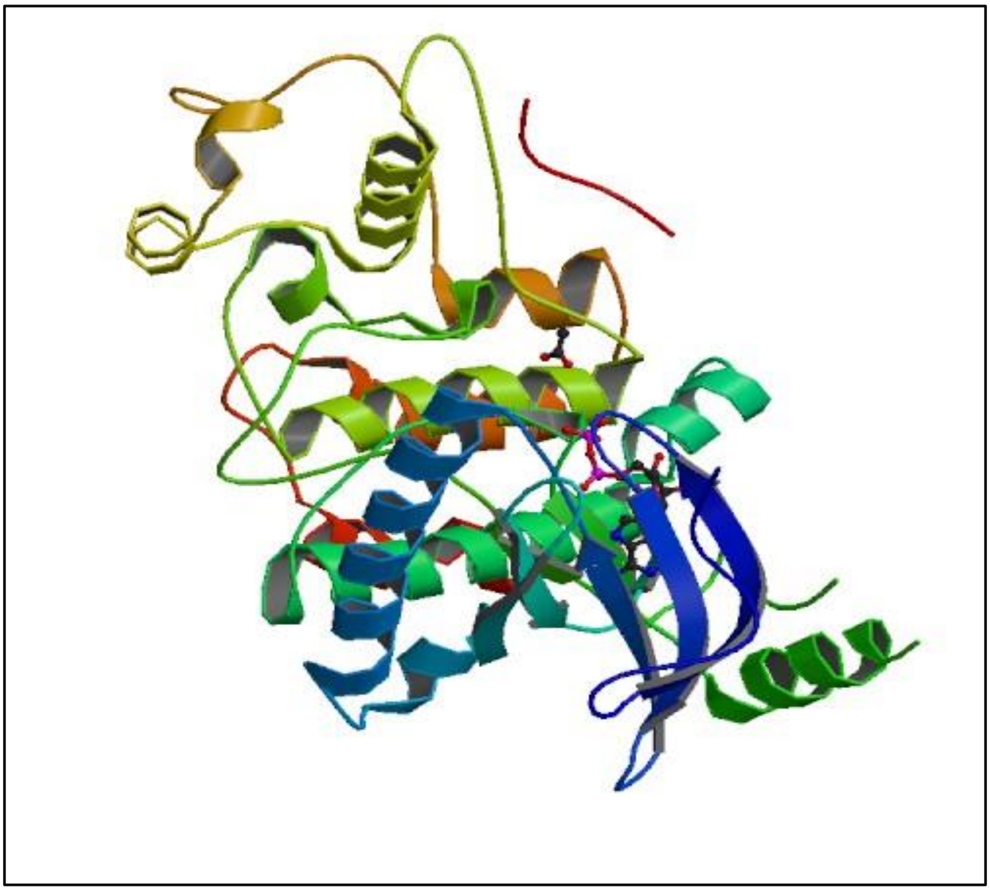


Fig 4.4a: Three dimensional structure of SFRS protein kinase 1 with PDB ID 1WBP[29]

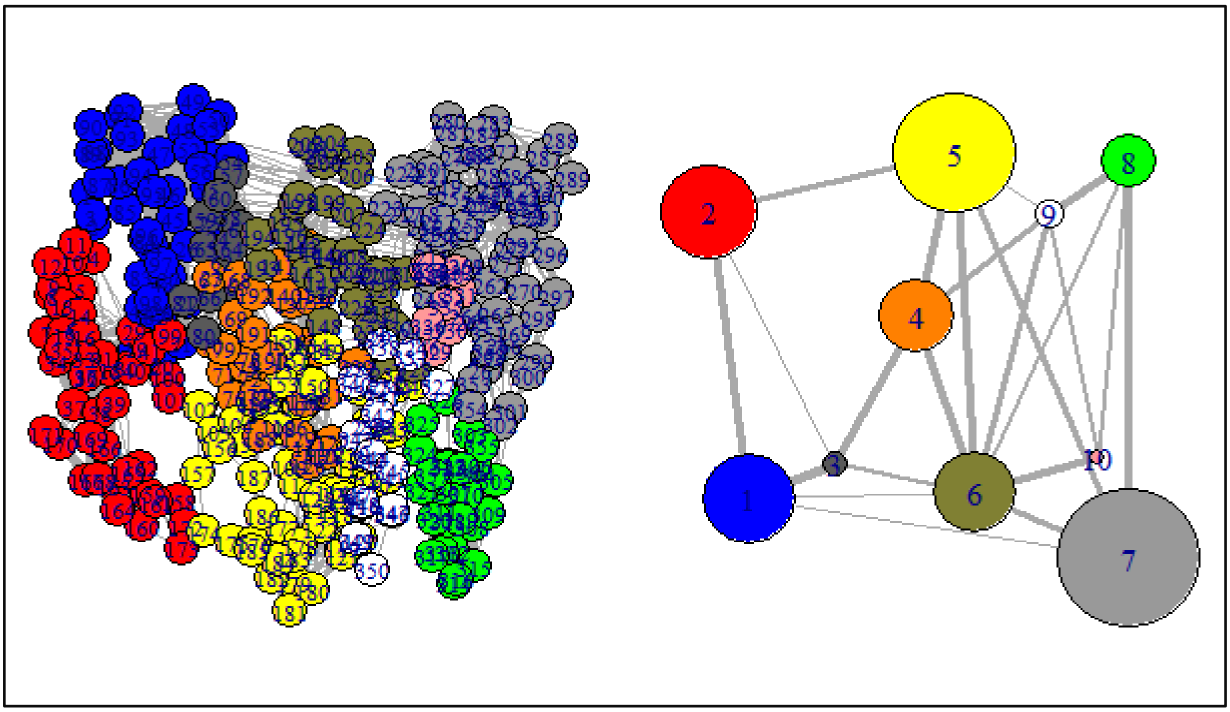


Fig 4.4b: All residue network and simple community network of 1WBP

c) tec protein tyrosine kinase (TEC)

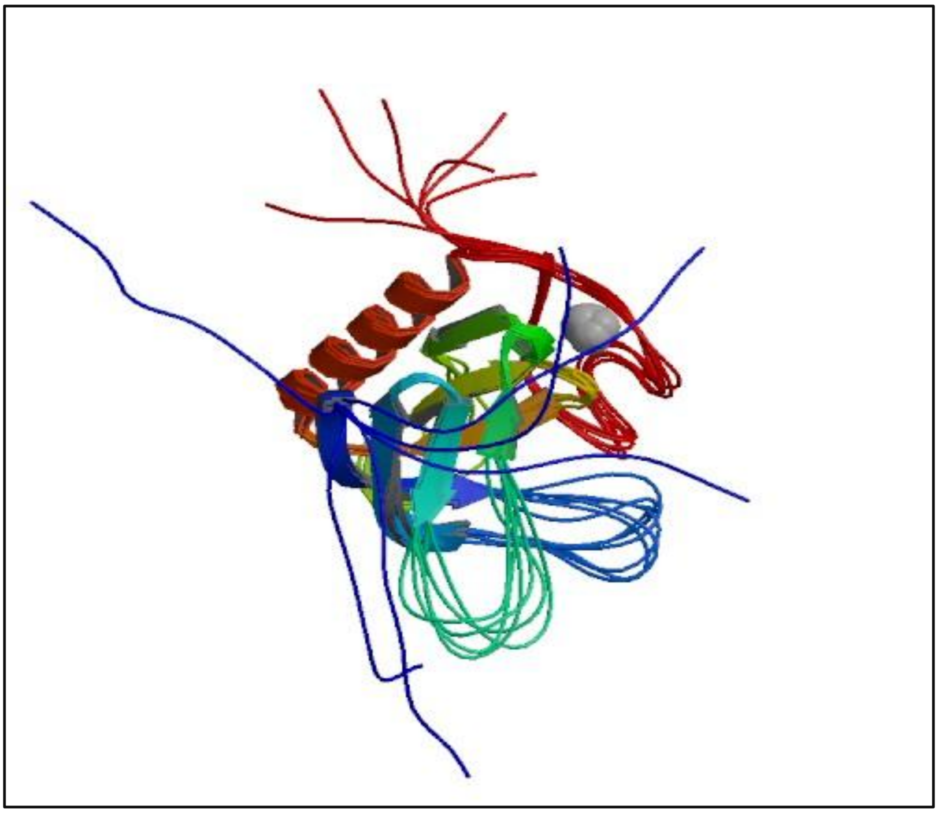


Fig 4.5a: Three dimensional structure of tec protein tyrosine kinase with PDB ID 2LUL[30]

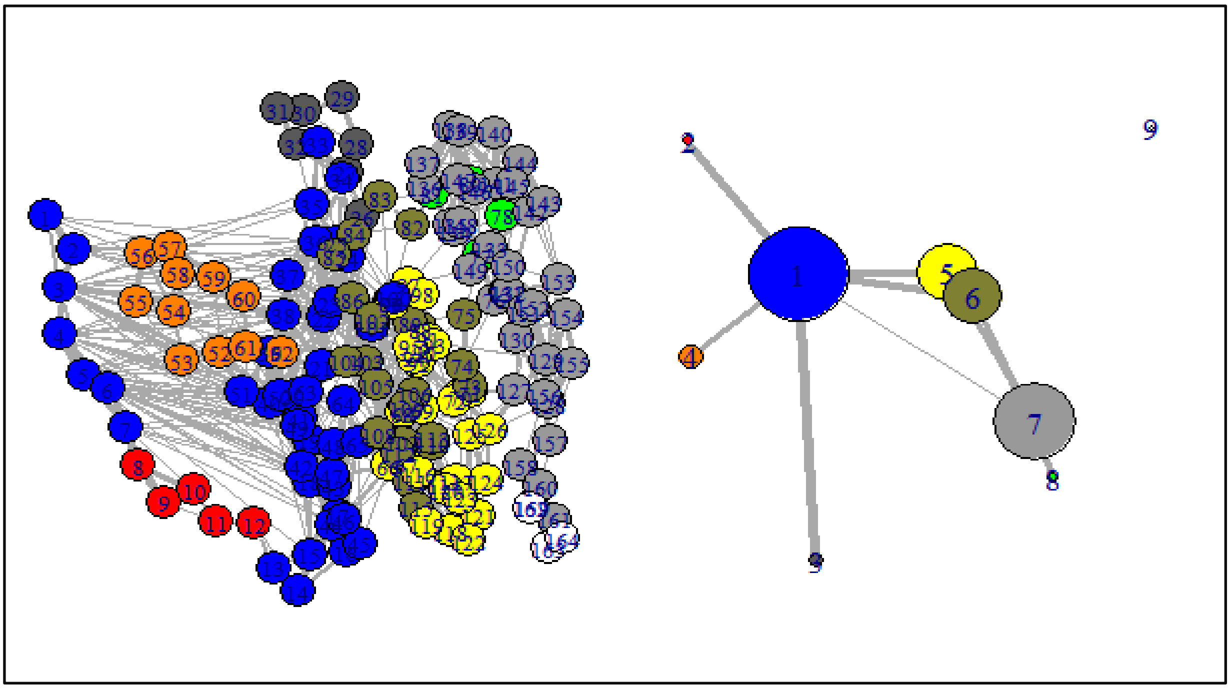


Fig 4.5b: All residue network and simple community network of 2LUL

d) mitogen-activated protein kinase kinase 6 (MAP2K6), transcript variant 2

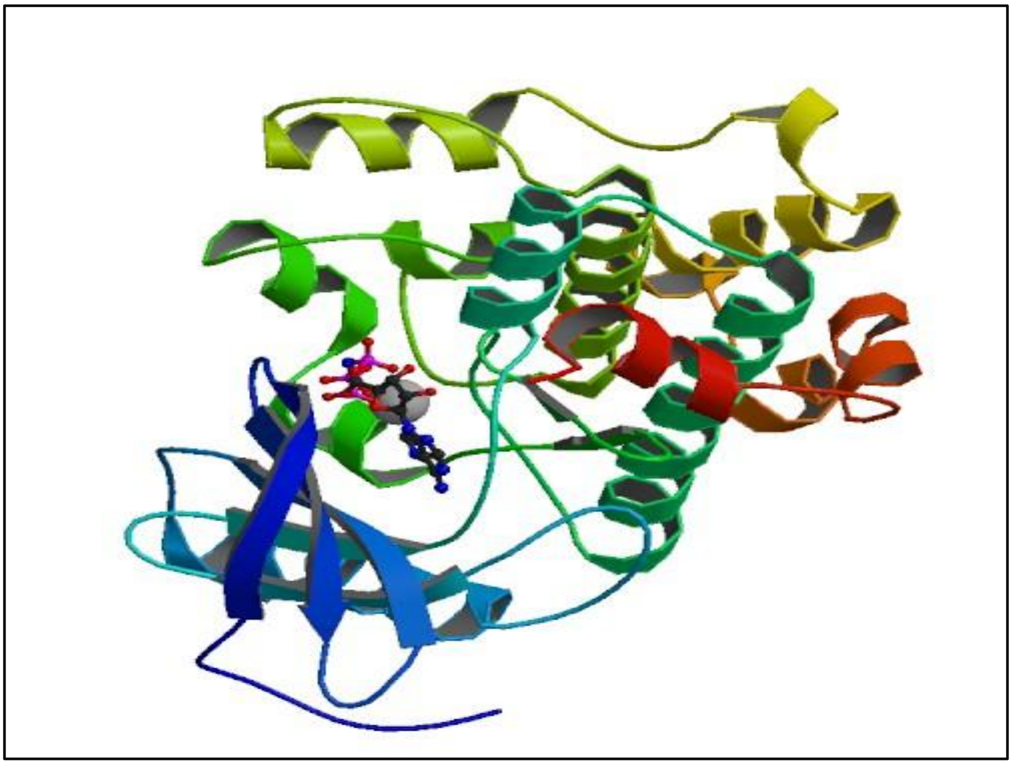


Fig 4.6a: Three dimensional structure of mitogen-activated protein kinase kinase 6 (MAP2K6), transcript variant 2 with PDB ID 3VN9[31]

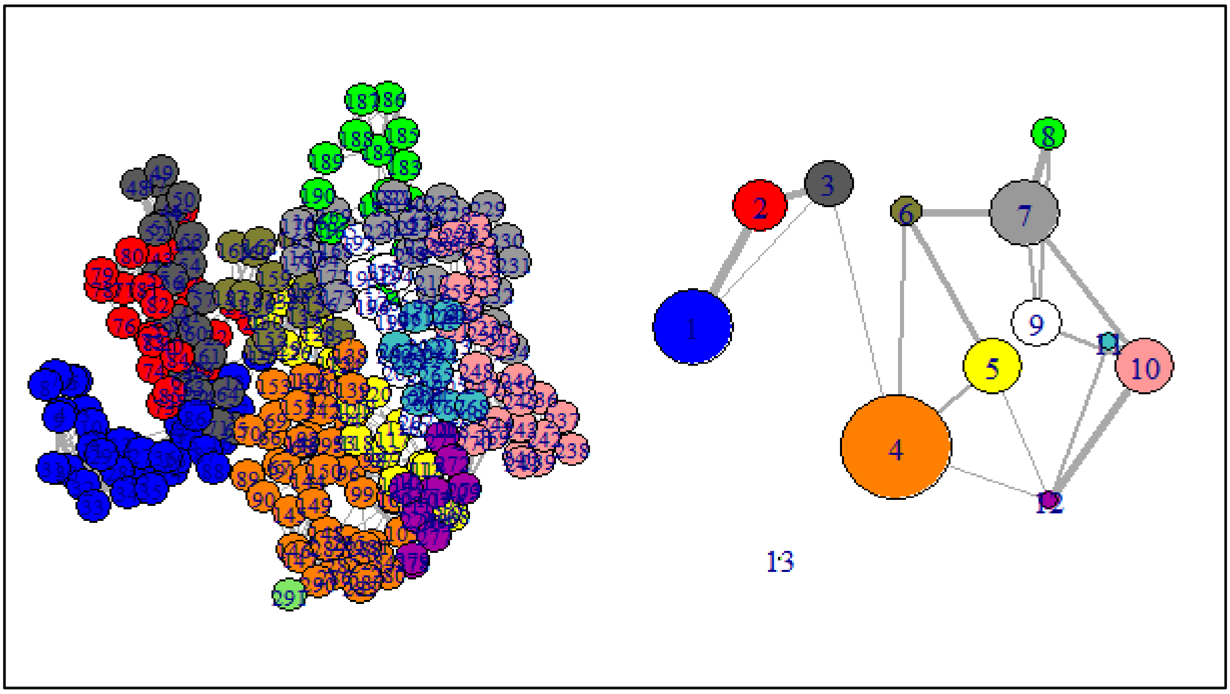


Fig 4.6b: All residue network and simple community network OF 3VN9

e) Bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 1

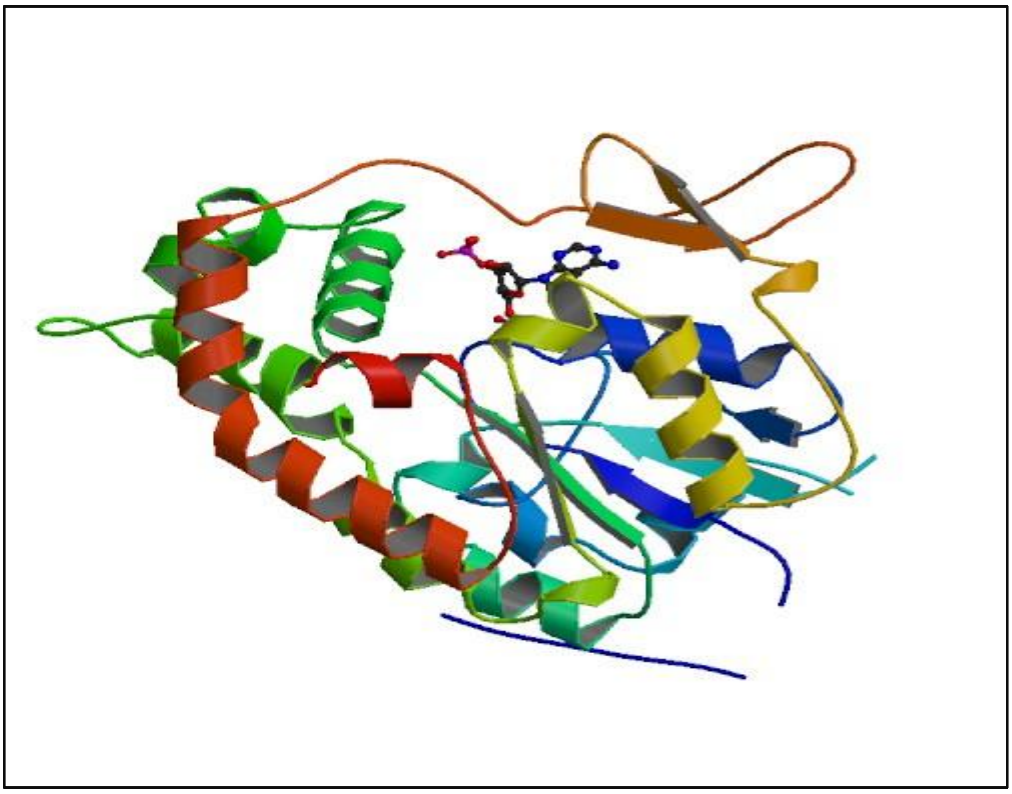


Fig 4.7a: Three dimensional structure of Bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 1 with PDB ID 1NST[32]

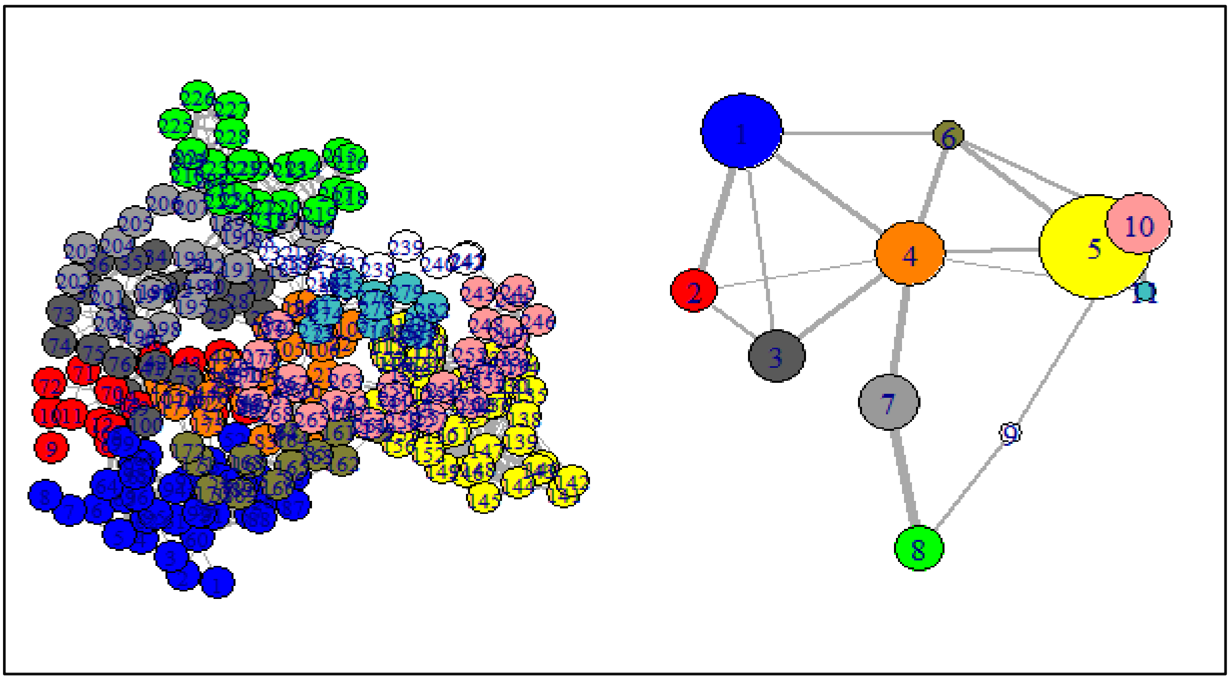


Fig 4.7b: All residue network and simple community network of 1NST

f) glutathione S-transferase omega 2 (GSTO2)

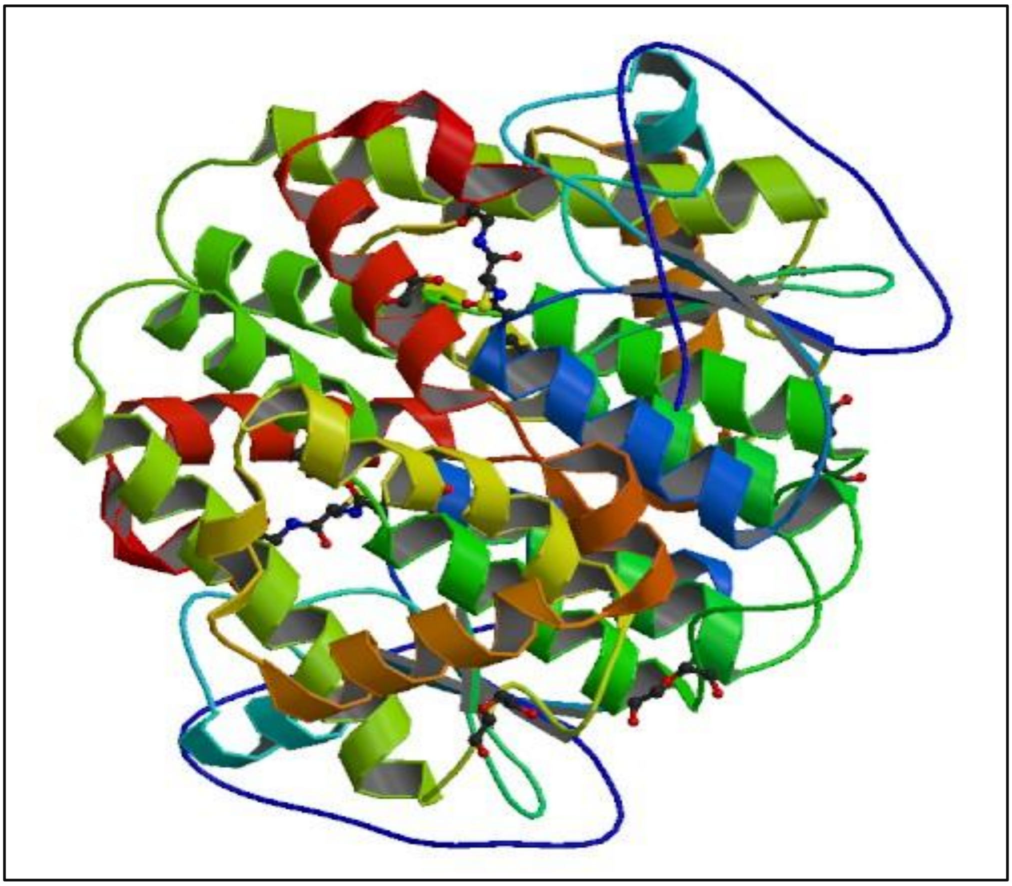


Fig 4.8a: Three dimensional structure of glutathione S-transferase omega 2 with PDB ID 3QAG[33]

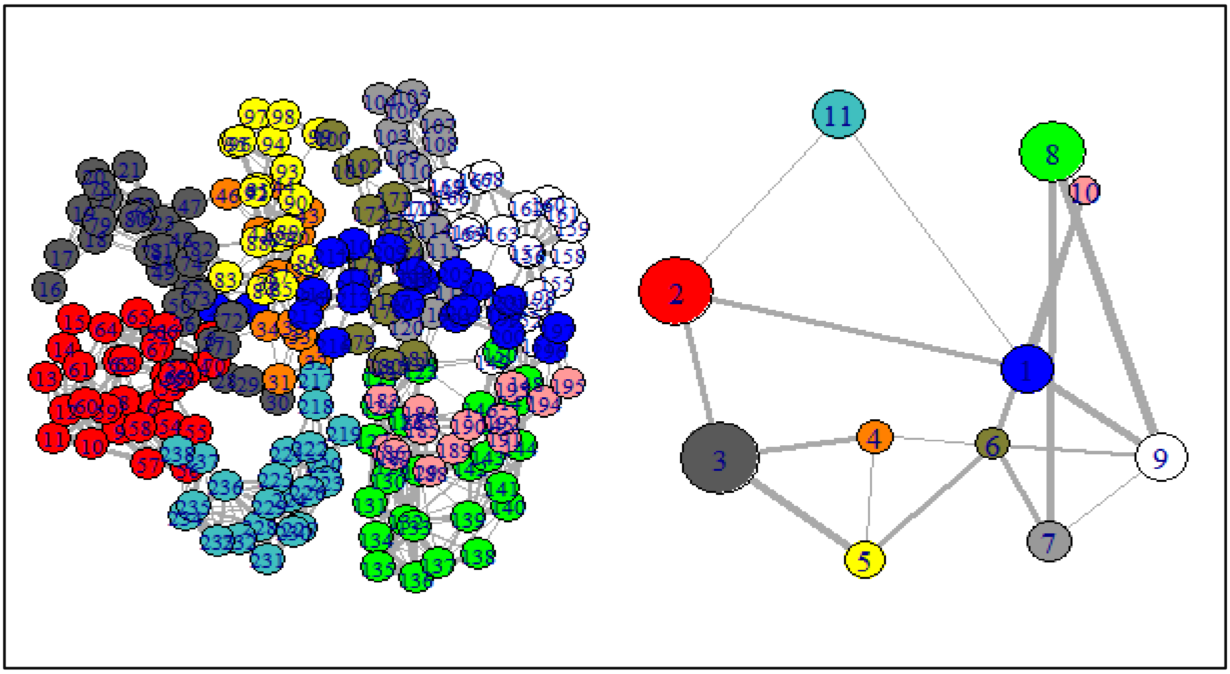


Fig 4.8b: All residue network and simple community network of 3QAG

4.3.2. Identification of modules containing the disordered regions from the structure network

After obtaining the structure network, the modules containing the disordered regions are identified for later steps. The module numbers can be obtained by performing a network analysis. For each of the above given protein the module numbers thus acquired are given below:

PDB ID	MODULE NUMBERS
1AZ2	11,13
1NST	1,2
1WBP	4,5,6,7,8,9,10
2LUL	7
3QAG	1,2
3VN9	1,2

TABLE 6: Showing the module numbers containing the disordered regions

4.4 Sequence Based Analysis

4.4.1. Identifying the protein family and obtaining seed alignment

Uniprot, alongside protein sequence stores detailed functional information about the proteins. Hence, we can easily obtain information about which family a particular proteins belong to and the corresponding Pfam ID is noted. The Pfam ID is then used to acquire its seed alignment.

The Pfam ID thus obtained for the proteins mentioned above is given below:

UNIPROT ID	PFAM ID
P15121	PF00248
P52848	PF00685
Q96SB4	PF00069
P42680	PF07714
Q9H4Y5	PF13417
P52564	PF00069

TABLE 7: Retrieving the Pfam ID required for seed alignment

4.4.2. Plotting the predicted contact map

The aligned sequence is then used to perform direct coupling analysis to obtain the DCA contact points[22].The contact map thus obtained for each protein corresponding to the family it belongs to, is given below:

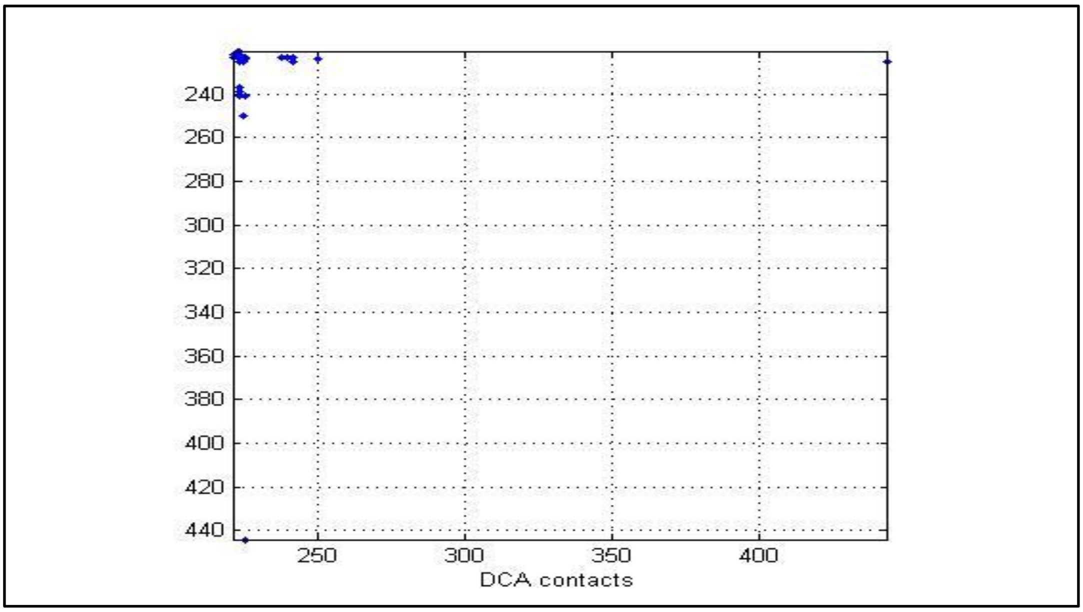


Fig 4.9: Contact map for 1AZ2

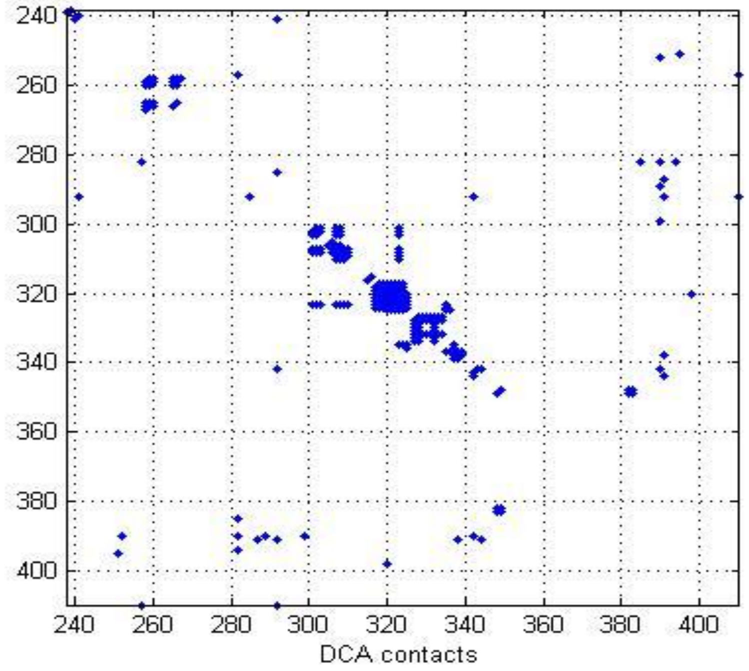


Fig 4.10: Contact map for 1WBP

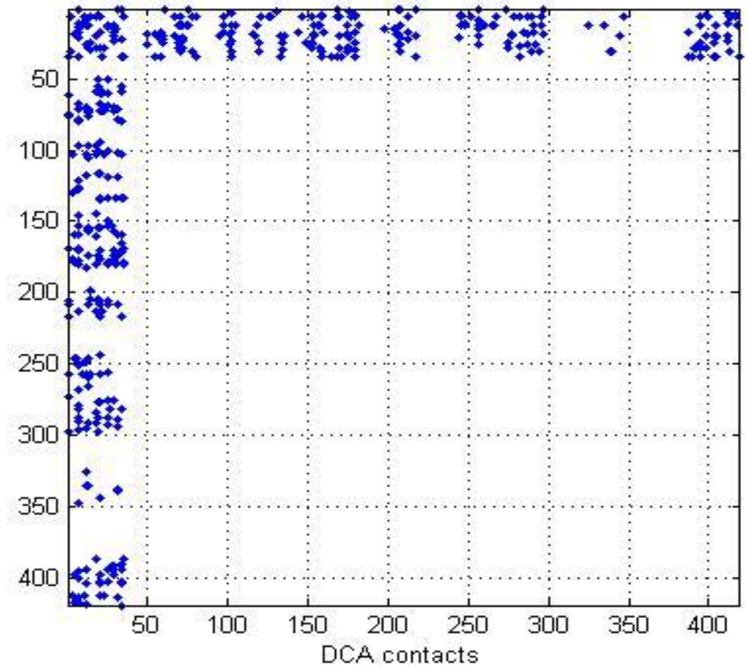


Fig 4.11: Contact map for 3VN9

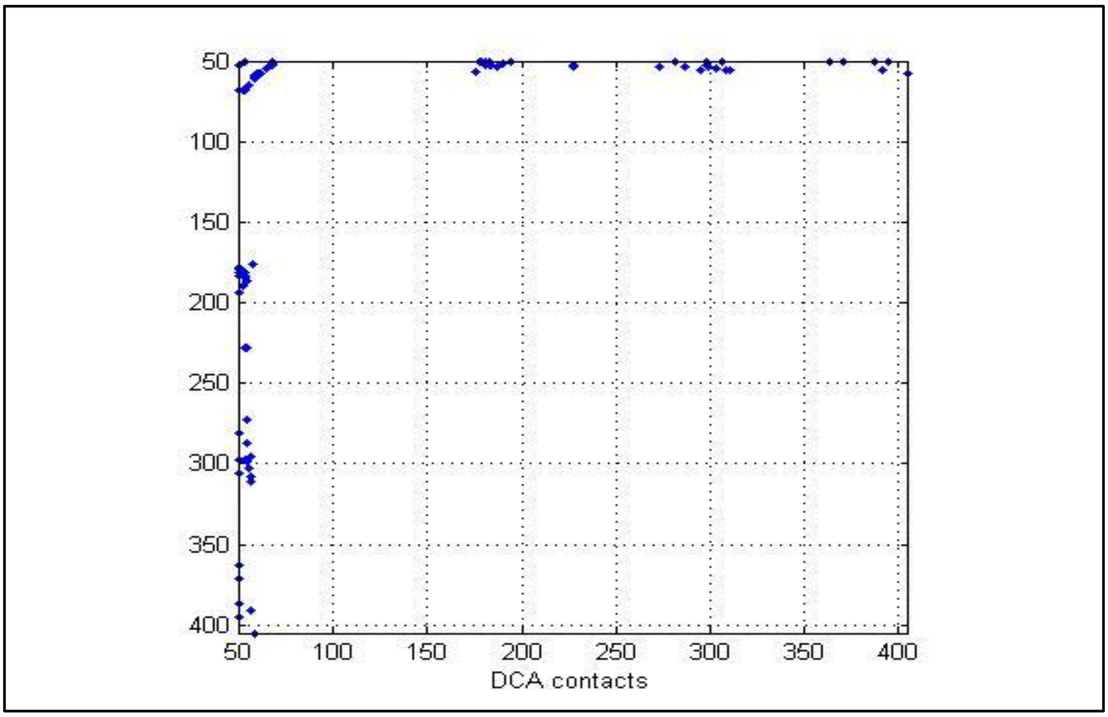


Fig 4.12: Contact map for 1NST

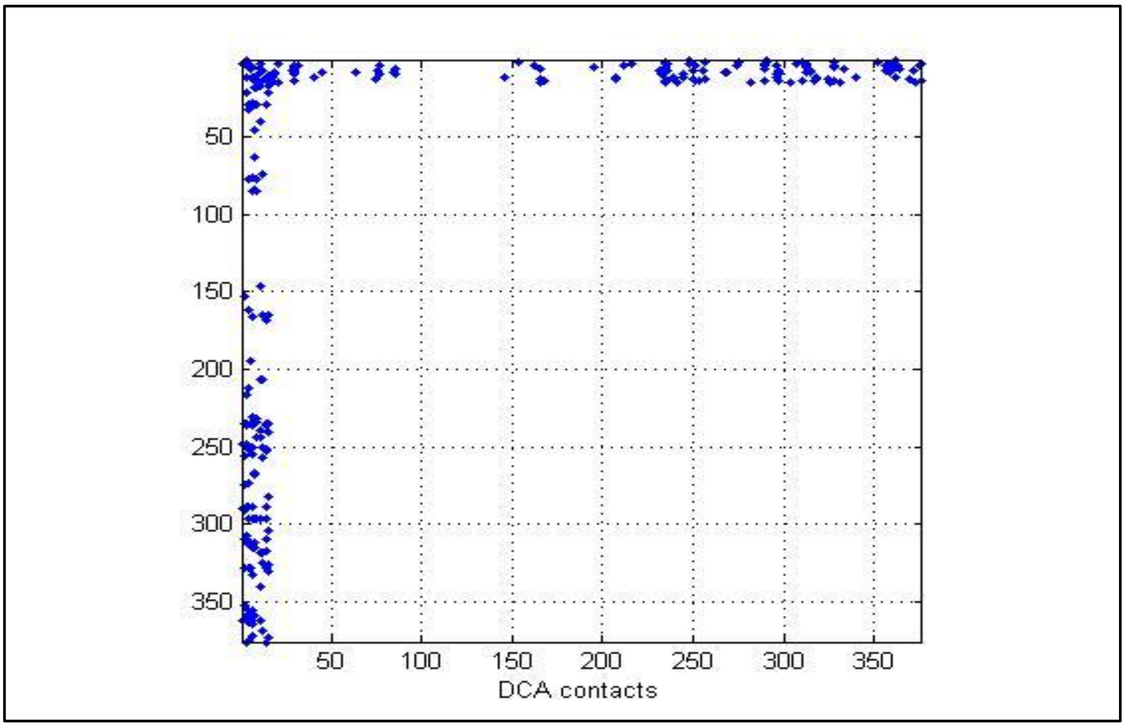


Fig 4.13: Contact map for 3QAG

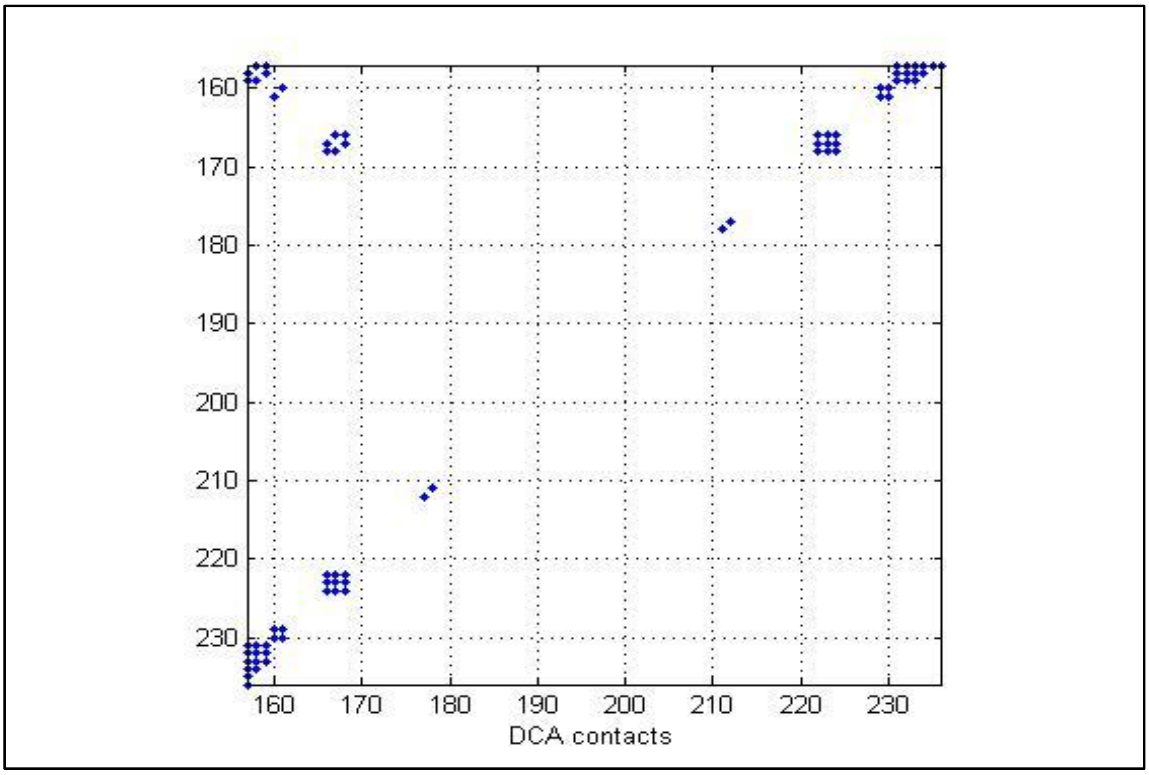


Fig 4.14: Contact map for 2LUL

4.5 Comparative Study and Discussion

For each of the proteins, the active site and the corresponding module number of the network structure in which the active site resides in, is noted. Even if module number is absent in some cases we can still conclude using the direct relationship theory.

PDB ID	ACTIVE SITE	MODULE NUMBER FOR ACTIVE SITE
1AZ2	49	3
1NST	614	
1WBP	213	6
2LUL	489	
3QAG	32	4
3VN9	179	7

TABLE 8: Showing the active site and the module containing the active site

For each of the above proteins a detailed study is performed to reach the conclusion

1AZ2: From table 6, it can be seen that the network modules containing the disordered residues are 11 and 13. In addition, from table 8, it can be seen that the active site not in any of disordered module. From the DCA contact map, it can be inferred that no disordered residues is in a direct relationship with the active site. In addition, no residue, which is in direct contact with the active site, is in the same module with any of the disordered residues. Thus, it can be concluded that the disordered regions have no effect on the active.

1NST: From table 6, it can be seen that the network modules containing the disordered residues are 1 and 2. Since the network module containing the active site is not present the analysis is done based on the DCA contact map. From the DCA contact map, it can be seen that no disordered residues is in a direct relationship with the active site. In addition, no residue, which is in direct contact with the active site, is in the same module with any of the disordered residues. Thus, here also it can be concluded that the disordered regions have no effect on the active.

1WBP: From table 6, it can be seen that the network modules containing the disordered residues are 4,5,6,7,8,9 and 10. Moreover, from table 8, the active site reside in network module 6. Thus the disordered residue has a direct affect on the active site as both share a common network module. So any chemical reaction the protein takes part will result in abnormality.

2LUL: From table, 6 it can be seen that the network module 7 contain the disordered residues. Since the network module containing the active site is not present, the analysis is done based on the DCA contact map. From the DCA contact map, it can be seen that no disordered residues is in a direct relationship with the active site. In addition, no residue, which is in direct contact with the active site, is in the same module with any of the disordered residues. Thus, here also it can be concluded that the disordered regions have no effect on the active.

3VN9: From table 6, it can be seen that the network modules containing the disordered residues are 1 and 2. Moreover, from table 8, the active site reside in network module 7.From the DCA contact map it can be seen that the active site is directly related to the disordered residues 25,5,35 and 30. Hence even though the active site does not reside in any of the disordered modules abnormality is still present.

3QAG: From table 6, it can be seen that the network modules containing the disordered residues are 1 and 2. Moreover, from table 8, the active site reside in network module 4. From the DCA contact map it can be seen that the active site is directly related to the disordered residues 25,5,35 and 30. Hence, even though the active site does not reside in any of the disordered modules abnormality is still present.

Chapter 5:

Conclusion

&

Future Scope

5. CONCLUSION AND FUTURE SCOPE

Biomarkers research continues to be a very active area of interest for AD PD or MS. They help in early diagnosis of this kind of neurodegenerative diseases and also acts as an major stepping stone in distinguishing these diseases from other diseases with increased accuracy. The present work involves the identification of common biomarkers altering neurodegeneration that could lead to early diagnosis or new drugs targeting the management of AD, PD or MS. Since IDP or IDR are attractive targets of drugs for protein-protein interaction, accurate identification of those would open a new era in therapeutic intervention.

The use of biomarkers in clinical research is still new and hence somewhat limited. Nevertheless better and progressive approach will help in further strengthening the study and opening a new dimension of therapeutic approach. An example would be the use of the concept of betweenness centrality to better understand the dependency among the residues which will in turn help in achieving a more accurate result.

In future, these approaches are likely to represent powerful tools for improving how these neurodegenerative patients are phenotyped, thereby expanding future considerations for managing patients with MCI and making personalized medicine possible.

REFERENCES

1. *Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst,A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies,B., Phelps, C.H., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic guidelines forAlzheimer's disease. Alzheimer's Dement. 7 (3), 270–279, http://dx.doi.org/10.1016/j.jalz.2011.03.008.*

2. *Farlow, M.R., 2009. Treatment of mild cognitive impairment (MCI). Curr. AlzheimerRes. 6 (4), 362–367, http://dx.doi.org/10.2174/156720509788929282.*

3. *Patricio Andres Donnelly-Kehoe, Guido Orlando Pascariello, Juan Carlos Gómez, (for the Alzheimers Disease Neuroimaging Initiative.) Looking for Alzheimer's Disease morphometric signatures using machine-learning techniques. Journal of Neuroscience Methods 302 (2018) 24–34 https://doi.org/10.1016/j.jneumeth.2017.11.013*

4. *Thomas F. Tropea, Alice S. Chen-Plotkin. Department of Neurology, Perelman School of Medicine at the University of Pennsylvania, United States Unlocking the mystery of biomarkers: A brief introduction, challenges and opportunities in Parkinson Disease. Parkinsonism and Related Disorders 46 (2018) S15eS18*

5. *L.M. de Lau, M.M. Breteler, Epidemiology of Parkinson's disease, Lancet Neurol. 5 (2006) 525e535, http://dx.doi.org/10.1016/S1474-4422(06)70471-9.*

6. *E.R. Dorsey, R. Constantinescu, J.P. Thompson, K.M. Biglan, R.G. Holloway, K. Kieburtz, F.J. Marshall, B.M. Ravina, G. Schifitto, A. Siderowf, C.M. Tanner, Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030, Neurology 68 (2007) 384e386, http://dx.doi.org/ 10.1212/01.wnl.0000247740.47667.03.*

7. *P. Martinez-Martin, C. Rodriguez-Blazquez, M.M. Kurtis, K.R. Chaudhuri, The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease, Mov. Disord. 26 (2011) 399e406, http://dx.doi.org/10.1002/mds.23462.*

8. *NINDS Multiple Sclerosis Information Page". National Institute of Neurological Disorders and Stroke. 19 November 2015.*

9. *Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker , 1Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 2Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Russia 3Molecular Kinetics, Inc., Indianapolis, Indiana. Intrinsically Disordered Proteins in Human Diseases:Introducing the D2 Concept. 10.1146/annurev.biophys.37.032807.125924*

10. *Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, Clin. Pharmacol. Ther. 69 (3) (2001 Mar) 89e95.*

11. *Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztányi,S., Uversky,V.N., Obradovic,Z., Kurgan,L., Dunker,A.K., Gough,J. (2013). "D2P2: Database of Disordered Protein Predictions." Nucleic Acids Research 41(D1):D508-D516 DOI:10.1093/nar/gks1226*

12. *H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.*

13. *The Pfam protein families database: towards a more sustainable future: R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman. Nucleic Acids Research (2016) Database Issue 44:D279-D285*

14. *N. El Kadmiri et al. Neuroscience 370 (2018) 181–190 Biomarkers for Alzheimer Disease: Classical and Novel Candidates' Review*

15. *E. Thouvenot. Multiple sclerosis biomarkers: Helping the diagnosis? https://doi.org/10.1016/j.neurol.2018.04.002*

16. *Thomas F. Tropea, Alice S. Chen-Plotkin. Unlocking the mystery of biomarkers: A brief introduction, challenges and opportunities in Parkinson Disease.*

17. *Leonid Breydo, Jessica W. Wu , Vladimir N. Uversky. α-Synuclein misfolding and Parkinson's disease*

18. *Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. (21 November 2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". Proceedings of the*

National Academy of Sciences. 108 (49): E1293–E1301. doi:10.1073/pnas.1111471108. PMC 3241805 . PMID 22106262.

19. *Kamisetty, H.; Ovchinnikov, S.; Baker, D. (5 September 2013). "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era". Proceedings of the National Academy of Sciences. 110 (39): 15674–15679. doi:10.1073/pnas.1314045110. PMC 3785744 PMID 24009338.*

20. *Ekeberg, Magnus; Lövkvist, Cecilia; Lan, Yueheng; Weigt, Martin; Aurell, Erik (11 January 2013). "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models". Physical Review E. 87 (1). arXiv:1211.1281 . doi:10.1103/PhysRevE.87.012707.*

21. *Marks, Debora S.; Colwell, Lucy J.; Sheridan, Robert; Hopf, Thomas A.; Pagnani, Andrea; Zecchina, Riccardo; Sander, Chris; Sali, Andrej (7 December 2011). "Protein 3D Structure Computed from Evolutionary Sequence Variation". PLoS ONE. 6 (12): e28766. doi:10.1371/journal.pone.0028766. PMC 3233603 . PMID 22163331.*

22. *De Leonardis, Eleonora; Lutz, Benjamin; Ratz, Sebastian; Cocco, Simona; Monasson, Rémi; Schug, Alexander; Weigt, Martin (29 September 2015). "Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction". Nucleic Acids Research. 43: gkv932. doi:10.1093/nar/gkv932. PMC 4666395 PMID 26420827.*

23. *Weinreb, Caleb; Riesselman, Adam J.; Ingraham, John B.; Gross, Torsten; Sander, Chris; Marks, Debora S. (May 2016). "3D RNA and Functional Interactions from Evolutionary Couplings". Cell. 165 (4): 963–975. doi:10.1016/j.cell.2016.03.030.*

24. *Ovchinnikov, Sergey; Kamisetty, Hetunandan; Baker, David (1 May 2014). "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information". eLife. 3. doi:10.7554/eLife.02030.*

25. *Feinauer, Christoph; Szurmant, Hendrik; Weigt, Martin; Pagnani, Andrea; Keskin, Ozlem (16 February 2016). "Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon". PLOS ONE. 11 (2): e0149166. doi:10.1371/journal.pone.0149166. PMC 4755613 . PMID 26882169*

26. *.F Morcos, A Pagnani, B Lunt, A Bertolino, DS Marks, C Sander, R Zecchina, JN Onuchic, T Hwa, M Weigt (2011), Direct-coupling analysis of residue co-evolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. 108:E1293-1301.*

27. *Pauling L, Corey RB, Branson HR (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain". Proc Natl Acad Sci USA.* ***37*** *(4): 205–211. doi:10.1073/pnas.37.4.205. PMC 1063337. PMID 14816373.*

28. *The alrestatin double-decker: binding of two inhibitor molecules to human aldose reductase reveals a new specificity determinant. Harrison, D.H.***,** *Bohren, K.M.***,** *Petsko, G.A.***,** *Ringe, D.***,** *Gabbay, K.H. (1997) Biochemistry 36***:** *16134-1614. PubMed: 9405046 .DOI: 10.1021/bi9717136*

29. *Interplay between Srpk and Clk/Sty Kinases in Phosphorylation of the Splicing Factor Asf/Sf2 is Regulated by a Docking Motif in Asf/Sf2 Ngo, J.C., Chakrabarti, S., Ding, J.-H., Velazquez-Dones, A., Nolen, B., Aubol, B.E., Adams, J.A., Fu, X.-D., Ghosh, G. (2005) Mol.Cell 20: 77. PubMed: 16209947 DOI: 10.1016/j.molcel.2005.08.025*

30. *Solution NMR Structure of PH Domain of Tyrosine-protein kinase Tec from Homo sapiens, Northeast Structural Genomics Consortium (NESG) Target HR3504C .Liu, G., Xiao, R., Janjua, H., Hamilton, K., Shastry, R., Kohan, E., Acton, T.B., Everett, J.K., Lee, H., Pederson, K., Huang, Y.J., Montelione, G.T.*

31. *Crystal structure of non-phosphorylated MAP2K6 in a putative auto-inhibition state. Matsumoto, T., Kinoshita, T., Matsuzaka, H., Nakai, R., Kirii, Y., Yokota, K., Tada, T. (2012) J.Biochem. 151: 541-549. PubMed: 22383536 . DOI: 10.1093/jb/mvs023*

32. *Crystal structure of the sulfotransferase domain of human heparan sulfate N-deacetylase/ N-] sulfotransferase 1. Kakuta, Y., Sueyoshi, T., Negishi, M., Pedersen, L.C. (1999) J.Biol.Chem. 274: 10673-10676. PubMed: 10196134*

33. *Structural insights into the dehydroascorbate reductase activity of human omega-class glutathione transferases. .Zhou, H., Brock, J., Liu, D., Board, P.G., Oakley, A.J.(2012) J.Mol.Biol. 420: 190-203. PubMed: 22522127. DOI: 10.1016/j.jmb.2012.04.014*