

**PREDICTION OF DISORDERED REGIONS THROUGH THEIR  
HYDROPHOBICITY SCORES IN THE PROTEINS RESPONSIBLE  
FOR POLYGLUTAMINE REPEAT DISEASES BY NEURAL  
NETWORK REGRESSION ANALYSIS**

**A**

**THESIS**

Submitted in Partial Fulfillment of the  
Requirements for the Award of Degree of

**MASTER OF TECHNOLOGY  
IN  
BIOINFORMATICS**

**Submitted By  
AINA MARYAM  
Scholar No: 132104113**



**JUNE 2015  
DEPARTMENT OF BIOINFORMATICS  
MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY  
BHOPAL (M.P.)-462051**

**PREDICTION OF DISORDERED REGIONS THROUGH THEIR  
HYDROPHOBICITY SCORES IN THE PROTEINS RESPONSIBLE  
FOR POLYGLUTAMINE REPEAT DISEASES BY NEURAL  
NETWORK REGRESSION ANALYSIS**

**A  
THESIS**

Submitted in Partial Fulfillment of the  
Requirements for the Award of Degree of

**MASTER OF TECHNOLOGY  
IN  
BIOINFORMATICS**

**Submitted By  
AINA MARYAM  
Scholar No: 132104113**



**Under the Guidance of  
Dr. Chandan Kumar Verma,  
Department of Mathematics, Bioinformatics and Computer Application,  
MANIT Bhopal, India**

**JUNE 2015  
DEPARTMENT OF BIOINFORMATICS  
MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY  
BHOPAL (M.P.)-462051**



**मौलाना आज़ाद राष्ट्रीय प्रौद्योगिकी संस्थान, भोपाल, भारत, 462051**  
**MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY**  
**BHOPAL, INDIA, 462051**

---

## **CANDIDATE DECLARATION**

This is to certify that project report entitled “Prediction of disordered regions through their hydrophobicity scores in the proteins responsible for polyglutamine repeat diseases by neural network regression analysis”, which is submitted by me in partial fulfillment of the requirement for the completion of M.Tech. in Bioinformatics to Maulana Azad National Institute of Technology, Bhopal comprises only my original work and due acknowledgement has been made in the text to all other material used.

Aina Maryam



मौलाना आज़ाद राष्ट्रीय प्रौद्योगिकी संस्थान, भोपाल, भारत, 462051  
MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY  
BHOPAL, INDIA, 462051

---

## CERTIFICATE

This is to certify that Aina Maryam, a student of M.Tech of batch 2013-2015 has completed the Project titled “Prediction of disordered regions through their hydrophobicity scores in the proteins responsible for polyglutamine repeat diseases by neural network regression analysis”, being submitted for the fulfillment of degree of M.Tech. in Bioinformatics to Maulana Azad National Institute of Technology, Bhopal, is a bonafide record of research work carried out by him under my supervision.

Date:

Dr. Chandan Kumar Verma,  
Assistant Professor



## ACKNOWLEDGEMENT

Acknowledgement is not a mere formality or ritual but a genuine opportunity to express the indebtedness to all those without whose active support and encouragement this project would not have been possible. One of the most pleasing aspects in collecting the necessary information and compiling it is the opportunity to thank those who have actively contributed to it. Successful completion of a well cherished task in the scientific methodology has its own reward. I shall ever, remain thank fully indebted to all those learned souls, my present and former teachers, known and unknown hands who directly or indirectly motivated me to achieve my goal and enlightened me with the touch of their knowledge and constant encouragement. I feel this is an extremely significant and joyous opportunity bestowed upon me by the goddess of learning, to think about and thank all those persons.

Words are inadequate lexicon to avouch the guidance given by my advisor **Dr. Chandan Kumar Verma**, Assistant Professor, Department of Mathematics (Computational & Integrated Science (Bio Informatics)), Maulana Azad National Institute of technology, Bhopal. Her dedication to research, meticulous planning, consecutive counsel and unreserved help served as a beacon light throughout the course of study, research work and completion of this manuscript.

I also express my grateful gratitude to, **Dr. K. R. Pardesani**, Professor, Dept of Mathematics (Computational & Integrated Science (Bio Informatics)), Maulana Azad National Institute of technology, Bhopal for his consistent and invaluable inspirations, prolific and

introspective guidance with constructive suggestions, deliberative discussions and active persuasion encouragement and providing access world class infrastructure for successful completion of this research work throughout the course of my study.

I am also highly thankful to our HOD, **Dr. Sanjay Sharma**, for his valuable guidance with suggestions and providing access class infrastructure for successful completion of my research work throughout the whole course study.

Regarding this thesis work, I would like to heartily thank my coordinator, **Dr. Usha Chauhan**, for her able guidance.

I am also thankful to our director, **Dr. Appu Kuttan K. K.**, and the entire teacher community of mathematics department namely, **Dr. Neeru Adlakha, Mr. Alekh Gour, Mr. Ashok Kumar Diwedi, Mr. Praveen Kumar, Mr. Prakash Nemade, Mr. Sunil Kumar Suryawanshi** and **Ms. Shweta Katiyar** for their co-operation throughout the course of study.

It is like a drop in the ocean of words that can never reach its mark to acknowledge infinite love, blessings, sacrifices and constant encouragement of my beloved parents who brought me to this stage.

I am highly grateful to DBT, New Delhi and MPCST, Bhopal for providing me Bioinformatics Infrastructure Facility and center for Bioinformatics to carry out this work at MANIT, Bhopal.

I apologize for the faux pass of the persons who have extended the help in a way or other and deserve such thanks.

**Aina Maryam**

## **ABSTRACT**

---

Intrinsically Disordered Proteins (IDPs) are signalized by their structural flexibility and they have allured number of researchers as many proteins which are accountable for various diseases have been proved to be IDPs. Hence, it makes them of significant relevance as drug targets. Also due to their structural ranges they have some functional relevance too which serves for their momentousness. Although X-ray crystallography is not applicable for the study of IDPs but NMR spectroscopy has been successful in achieving this. This led to the prediction of IDPs through machine learning techniques. The past two decades have witnessed the prediction of IDPs through Machine Learning Algorithms and several advances and development have been done in that area. The machine learning is commonly done on the basis of various reliable parameters of disordered proteins namely amino acid composition, charge, hydrophobicity, complexity, bulkiness, entropy, aromaticity and various other parameters. In this study, we use only hydrophobicity scales of amino acids in a protein for the prediction of IDPs. For this, we used MATLAB Neural Network tool to perform Neural Network Regression Plot Analysis.

**Keywords:** Intrinsically Disordered Proteins, X-ray crystallography, NMR spectroscopy, Machine Learning, Hydrophobicity, Neural Network, Regression Plot.

## TABLE OF CONTENTS

<b>1. Biological background</b>	<b>1-8</b>
1.1 Overview.....	2
1.2 Intrinsically disordered proteins.....	2
1.3 CAG-Polyglutamine Repeat Diseases.....	4
1.4 Targeting IDPs in neurodegenerative diseases.....	6
1.5 Peculiarities of amino acid sequences of IDPs.....	8
<b>2. Computational approach</b>	<b>9-14</b>
2.1 Overview.....	10
2.2 Computational analysis of IDPs.....	10
2.3 Computer-aided search for IDPs.....	11
2.4 Machine learning algorithms.....	12
2.5 Comparison of disorder prediction methods.....	13
<b>3. Review of literature</b>	<b>15-20</b>
3.1 Overview.....	16
3.2 Literature review.....	16
3.3 Conclusion.....	20
<b>4. Materials and methods</b>	<b>21-34</b>
4.1 Objective of this study.....	22
4.2 Data Resources.....	22



4.3 Tools used.....	24
4.3.1 MATLAB R2014a.....	24
4.3.2 Neural Network Toolbox™.....	26
4.3.3 ExPASy-ProtScale.....	28
4.4 Methodology.....	30
4.4.1 Creation of the input and target datasets.....	30
4.4.2 Training the datasets using nftool.....	32
4.5 Conclusion.....	34
 <b>5. Results and discussions</b>	 <b>35-48</b>
5.1 Overview.....	36
5.2 Results.....	36
5.2.1 Datasets.....	36
5.2.2 Regression plots.....	38
5.3 Discussion.....	48
 <b>6. Conclusion and future scope</b>	 <b>49-51</b>
6.1 Conclusion.....	50
6.2 Related work.....	50
6.3 Future scope.....	50
 <b>References</b>	 <b>52-60</b>

## LIST OF FIGURES

Figure No.	Description	Page No.
1.1	Length of disordered regions in different diseases.	7
1.2	Neuro-degeneration linked proteins follow the trends expected for IDPs.	7
2.1	Increase in number of disorder predictors from 1979 to 2009.	12
4.1	The graphical interface to the MATLAB workspace.	25
4.2	Basic Neural Network	26
4.3	Neural Network Toolbox	27
4.4	Two layer feed-forward network used by nftool.	28
4.5	ExPASy-ProtScale Tool	29
4.6	Parameters to be set before calculating hydrophobicity scales.	30
4.7	Snapshot of excel file imported in the form of matrix to create .mat file.	31
4.8	Snapshot of inputs and targets being given in the form of Matrix rows.	32
4.9	Division of hydrophobicity scores for training, testing and validation.	33
4.10	Regression plot was obtained using plotregression.	33
5.1	Snapshot of the result of the hydrophobicity scores of a single protein.	37
5.2	Snapshot of dataset of 100 ordered proteins.	37
5.3	Snapshot of dataset of 100 disordered proteins.	38
5.4	(a) Snapshot of DRPLA protein Atrophin-1 against disordered proteins.	39
	(b) Snapshot of DRPLA protein Atrophin-1 against ordered proteins.	39
5.5	(a) Snapshot of HD protein Huntingtin against disordered proteins.	40
	(b) Snapshot of HD protein Huntingtin against ordered proteins.	40
5.6	(a) Snapshot of KD protein AR against disordered proteins.	41
	(b) Snapshot of KD protein AR against ordered proteins.	41
5.7	(a) Snapshot of SCA1 protein Ataxin-1 against disordered proteins.	42
	(b) Snapshot of SCA1 protein Ataxin-1 against ordered proteins.	42

5.8	(a) Snapshot of SCA2 protein Ataxin-2 against disordered proteins.	43
	(b) Snapshot of SCA2 protein Ataxin-2 against ordered proteins.	43
5.9	(a) Snapshot of SCA3 protein Ataxin-3 against disordered proteins.	44
	(b) Snapshot of SCA3 protein Ataxin-3 against ordered proteins.	44
5.10	(a) Snapshot of SCA6 protein $\alpha$ 1A against disordered proteins.	45
	(b) Snapshot of SCA6 protein $\alpha$ 1A against ordered proteins.	45
5.11	(a) Snapshot of SCA7 protein Ataxin-7 against disordered proteins.	46
	(b) Snapshot of SCA7 protein Ataxin-7 against ordered proteins.	46
5.12	(a) Snapshot of SCA17 protein TBP against disordered proteins.	47
	(b) Snapshot of SCA17 protein TBP against ordered proteins.	47

## LIST OF TABLES

Table No.	Description	Page No.
1.1	Types of CAG repeat diseases.	5
5.1	R values of resultant and validation regression plots of nine pathogenic proteins against 100 ordered and 100 disordered proteins datasets.	48

# **Chapter 1**

## **Biological Background**

## 1.1 Overview

Proteins are large biological molecules, or macromolecules that differ from each other primarily in the sequence of amino acids, which is responsible for folding of the protein into a specific 3D-structure that determines its activity and functions. Sometimes, if proteins are not properly folded their functional activity and structures gets affected. In this chapter such disordered proteins have been covered. Also their amino acid peculiarities have been studied in this chapter.

## 1.2 Intrinsically Disordered Proteins

It has been said that a protein must be folded to perform its functions properly. But in the beginning of 2000, it was found that not all proteins work in folded state, rather, they must be unfolded to perform their functions. While some of them fold when they form complex with targets. These proteins are called Intrinsically Disordered/Unstructured Proteins (IDPs/IUPs). As the word suggests IDPs are the proteins that lack fixed state or ordered 3D structure. IDPs cover a panorama from partially structured to fully unstructured and which embodies large multi-domain proteins which are connected by flexible linkers, random coils and pre-molten globules.

According to some statistics, around 10% of proteins are fully disordered, while 40% of eukaryotic proteins have at least one long i.e. greater than 50 amino acids, disordered loop. Intrinsic order and disorder are determined at the amino acid residue level [1]. Sometimes in the amino acids of the ordered regions, their backbone atoms undergo low amplitude, thermally-driven motions about their equilibrium positions. In some cases, ordered regions mutually swap between two or more specific states. Contrarily, intrinsically disordered proteins or disordered regions exist as energized aggregates and the Ramachandran angles of their backbone vary significantly over time with no specific symmetry values and typically involve non-mutual change of states. Hence, the disorder can be attributed to the protein's dynamical properties, and not to the presence or absence of local secondary structure. So an IDP is the one which consists of at least one disordered region.

IDPs carry important biological functions which gradually led to the growing interest in these proteins. Intrinsic disorder can exist *in vivo*, this has been revealed by whole-cell NMR

experiments [2]. Twenty-eight specific functions, grouped in four categories were assigned to 100 IDPs after being characterized through experimental methods, these functions were:

- (i) molecular recognition
- (ii) molecular assembly
- (iii) protein modification
- (iv) entropic chain activities

Mainly, IDPs wrap up the regulation, signalling and control pathways where interactions with multiple partners occur and hence the interactions have to be of high-specificity and low-affinity [3]. Hence, the functional spectrum of disordered regions accompanies the spectrum of ordered protein regions.

The ability of IDPs to recognize and bind is the feat of their unfolded nature. Their large hydrodynamic dimensions slow down their diffusion and hence they are a large target for earlier molecular encounters. Also, since they do not have rigid pockets to bind, there are many ways for a binding partner to bind in different adaptations, which increases their probability of having different significant interactions [4]. Additionally, IDPs also allow molecular elasticity by having more than one conformation and binding diversity by binding to several proteins and this is the reason many of the known hub proteins are IDPs. IDPs' accelerated mass exodus in the cell makes their tight regulation possible which occurs whenever needed in cell cycle and cell signalling [5].

Structurally, it was suggested that IDPs prevail in two forms, which are collapsed and extended i.e. like forms of molten globule and random coil respectively. Another type of disordered regions i.e. the pre-molten globules were suggested but it was not clarified whether they are truly distinct from the likes of random coils. It was also determined that disordered proteins model conformational changes which are function dependent. The alterations in environmental or cellular surroundings were said to be responsible for such phenomenon [6].

When in a progressive study of 28 protein families, having both ordered and disordered regions were studied, in twenty families it was determined, disordered regions evolved earlier than ordered ones, and three families had ordered regions that evolved faster. The amino acid replacements are also affected by the variance in amino acid composition that leads to

disorder [7]. These days scoring matrices are being designed using disordered protein sequences while usually matrices that show the anticipation of mutation rely on the ordered protein sequences. It was determined that scoring matrices which rely on disordered protein are more fruitful for the alignment of similar disordered protein sequences rather than the ordered proteins scoring matrices [8].

If sequence analysis is done of the data taken from Protein Data Bank (PDB), Swiss Protein (SwissProt) database and thirty four complete or nearly complete genomes, it testimonies the fact that;

- In a protein structure disorder is quite common.
- The complexity of the sequence is directly proportional to the firmness of the prediction of disorder.
- Evidence of eukaryotes having more proteins having disorder is there rather than eubacteria and archebacteria [9].

### 1.3 CAG-Polyglutamine Repeat Diseases

CAG repeat diseases belong to the category of trinucleotide repeat disorders which are the set off genetic disorders occur due to trinucleotide repeat expansion. It is a kind of mutation where trinucleotide repeats in certain genes exceed the normal, stable threshold. If the repeat is there in a normal gene, an aggressive mutation might extends the repeat count and result in an abnormal gene. Trinucleotide repeats are sometimes tabulated as insertion mutations [10].

Presently, nine neurologic disorders are known to be caused by an increased number of CAG repeats. These repeats are present in the exons of otherwise unrelated proteins. During protein synthesis, the aggregated CAG repeats are coded into an array of glutamine residues without any barrier, which hence models a polyglutamine tract ("polyQ"). These polyglutamine tracts might get projected to increased extension.

Various studies have determined that the CAG repeats are not always encoded in some toxic protein. Rather it was found that a protein called muscleblind (mbl) in *Drosophila* is proficient of binding CAG tracts. Additionally, when the CAG repeat was tailored into a CAACAG duplicating repeats (which are also encoded into polyQ tracts), toxicity scaled down effectively [11]. MBNL1 which is present in human beings and is homologous to mbl

was first identified as proficient of fastening with CUG tracts in RNA [12]. Now it has also been determined to fasten with CAG [13, 14] and CCG [14] tracts as well.

**Table 1.1: Types of CAG repeat diseases.**

Type	Protein	Normal PolyQ Repeats	Pathogenic PolyQ Repeats
DRPLA (Dentatorubral-pallidoluysian atrophy)	Atrophin-1	6-35	49-88
Huntington's disease	HTT (Huntingtin)	6-35	36-250
Kennedy's disease	AR (Androgen Receptor)	9-36	38-62
SCA1 (Spinocerebellar Ataxia Type 1)	Ataxin-1	6-35	49-88
SCA2 (Spinocerebellar Ataxia Type 2)	Ataxin-2	14-32	33-77
SCA3 (Spinocerebellar Ataxia Type 3)	Ataxin-3	12-40	55-86
SCA6 (Spinocerebellar Ataxia Type 6)	$\alpha$ 1A-voltage dependent calcium channel subunit	4-18	21-30
SCA7 (Spinocerebellar Ataxia Type 7)	Ataxin-7	7-17	38-120
SCA17 (Spinocerebellar Ataxia Type 17)	Tata Binding Protein	25-42	47-63



CAG repeat disorder exhibits autosomal-dominant mode of inheritance while only Kennedy's disease displays X-linked inheritance. They exhibit midlife age of onset which aggravates with time. The number of CAG repeats is directly proportional to the seriousness of the disease and the age at onset. Despite the fact that responsible genes are strongly encoded in all the polyglutamine diseases, all of them characterize a discriminatory arrangement of neuro-degeneration.

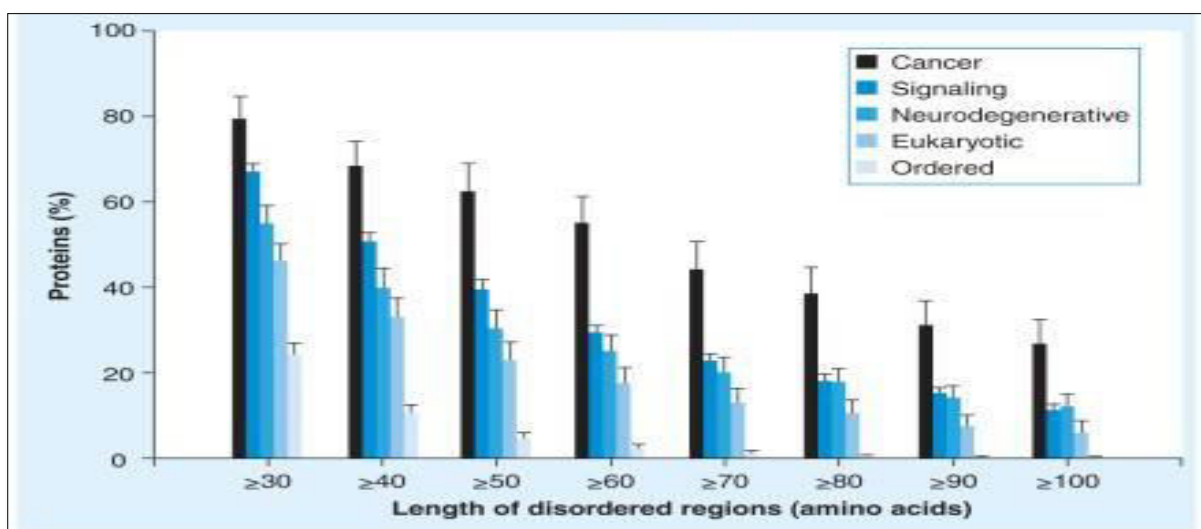
The revelation of each triplet repeat disorder is tremendously clinically beneficial. Because of this better allocation of the diseases is possible and promotes earlier diagnosis. It was determined that the pathological structure of trinucleotide repeat disorders is usually complicated and mostly includes more than one mode of operation of pathogenesis.

## 1.4 Targeting IDPs in neurodegenerative diseases

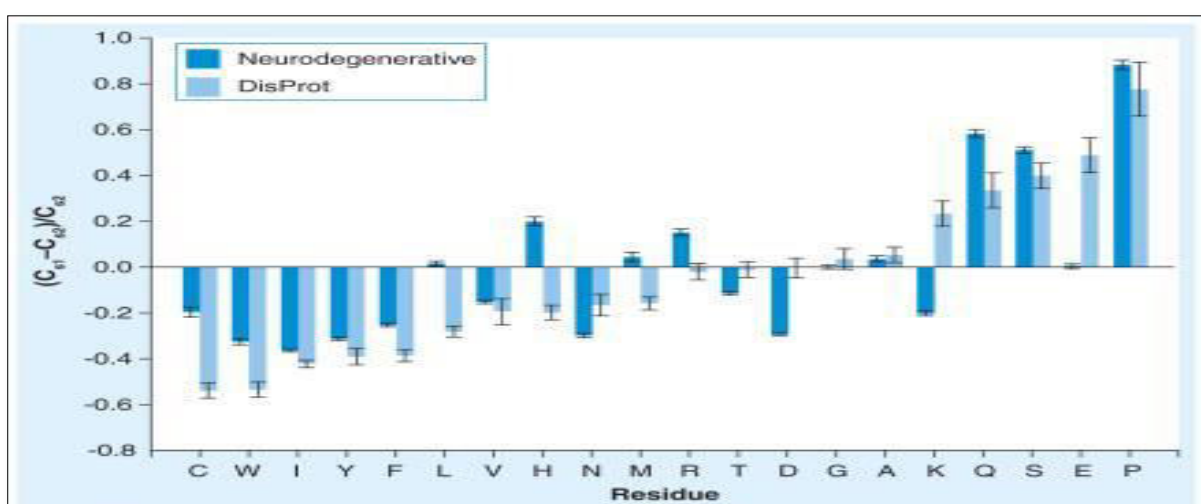
IDPs have been linked with various diseases by different bioinformatics tools. The results of study of figure 1 shows that there is no noticeable difference in the amount of disorderiness when the normal proteins responsible for cell signalling, and the mutated proteins responsible for other diseases like cancer, neurodegenerative diseases, cardiovascular disease and diabetes [15] were compared.

Through Figure 2 the amino acid compositions of different neurodegenerative diseases are IDPs are explained. The amino acid compositions of several neuro-degenerative disease related proteins and the compositions of ordered proteins from a protein data bank (PDB) are set side by side in this illustration [16]. Similar details of some important IDPs from the DisProt database [9] are also placed there for comparison. Normalization was done for the calculations which are explained for studying the IDPs [15, 17]. In the figure, negative values display the amino acids which are less in number in a given dataset, as compared to the amino acids in a set of ordered proteins. Whereas, the positive values show the residues that are present in high amount in a set.

Figure 2 shows that proteins related to neuro-degeneration usually chase the directions anticipated for IDPs. The lesser residues were order-promoting residues like C, I, Y, W, F, N and V while higher amount of disorder-promoting residues like P, R, S and Q were present [15]



**Figure 1.1: Length of disordered regions in different diseases [15].**



**Figure 1.2: Neuro-degeneration related proteins follow the trends expected for IDPs [15].**

Either complete disorderness or long stretches have been determined in the proteins responsible for CAG repeat diseases [16]. These proteins which are responsible are, androgen receptor in Kennedy's Disease [18], atrophin-1 in Dentatorubral-pallidoluyisian atrophy [19], ataxin-2 in spinocerebellar ataxia 2 [20], ataxin-3 in spinocerebellar ataxia 3 [21], voltage dependent calcium channel  $\alpha 1A$  subunit in spinocerebellar ataxia 6 [16], ataxin-7 in spinocerebellar ataxia 7 [16, 22] and TBP in spinocerebellar ataxia 17 [16, 23].

### 1.5 Peculiarities of amino acid sequences of IDPs

The combination of low hydrophobicity with high net charge is the basic requirement for unfoldedness of proteins because high net charge causes charge-charge repulsion, and low hydrophobicity causes limited protein compaction. The disordered proteins have smaller number of hydrophobic amino acids that are also bulky which are Val, Leu, and Ile and aromatic residues which are Phe, Tyr, and Trp. The hydrophobic centre of a folded macromolecular protein is formed by these amino acids. The content of Cys and Asn residues is also significantly low in IDPs. Cys is an important residue as it is important for the protein conformation and stability by the disulfide bond arrangements and is responsible for coordination of different attached groups [16].

Hence, the amino acids having lower levels like Tyr, Phe, trp, Ile, Val, Leu, Asn and Cys advocate orderness in a protein while, disorder advocating residues are Arg, Gly, Ala, Ser, Gln, Lys and Glu and also Pro which is hydrophobic as well as structure destroying amino acid [3, 15, 17, 24, 25].

Apart from amino-acid composition, the disordered segments are also compared with the ordered ones via attributes such as hydropathy, net charge, flexibility index, helix propensities, strand propensities, and compositions for groups of amino acids such as W + Y + F (aromaticity). Hence there are 265 propersty-based attribute scales [24] and more than 6,000 composition-based attributes [26].

Among these attributes it has been established that ten of the attributes, which include 14 Å contact number, hydropathy, flexibility,  $\beta$ -sheet propensity, coordination number, R+E+S+P, bulkiness, C+F+Y+W, volume, and net charge, give reliable results for discriminating ordered and disordered proteins [24].

# **Chapter 2**

## **Computational Approach**

## 2.1 Overview

With passing years, the amount of protein data has been increasing which has lead computational approaches come into picture to deal with such large amount of data. Apart from speed computer aided approaches are also more efficient and accurate. Since disordered proteins are unstable in nature the biological approach for their study is not always successful. Hence computational tools have been developed to study their characteristics and parameters responsible. In this chapter such approaches have been reviewed.

## 2.2 Computational analysis of IDPs

Three computational approaches have been put into picture for judging the abundance of IDPs:

- (1) In the first approach, many varieties of characteristic datasets of proteins are put in consideration and are linked with the query disease or syndrome. Distinct disorder predictors are used to do the analysis of these datasets computationally [27-30].
- (2) In the second approach, related proteins are interlinked within one disease and also between different diseases in a grid of genetic diseases [31].
- (3) In the third approach, the link of a specific protein function is evaluated with how much intrinsic disorder is present in a dataset of proteins that are responsible for that particular function [32-34].

The two step method used to assess the amount of intrinsic disorder in a query disease is as follow: first, a dataset of proteins related to that disease is found by searching different databases, and second, this assembly of proteins is assessed for intrinsic disorder [27-30, 35, 36]. This method has been established in being important to depict the high amount of disorderness in various proteins linked to cardiovascular disease, neurodegenerative diseases and cancer. It has been shown that such proteins have higher levels of disorder causing amino acids like Gln, Ser, Arg, Glu and Pro; and deficient in order causing amino acids like Tyr, Ile, Trp, Phe and Val.

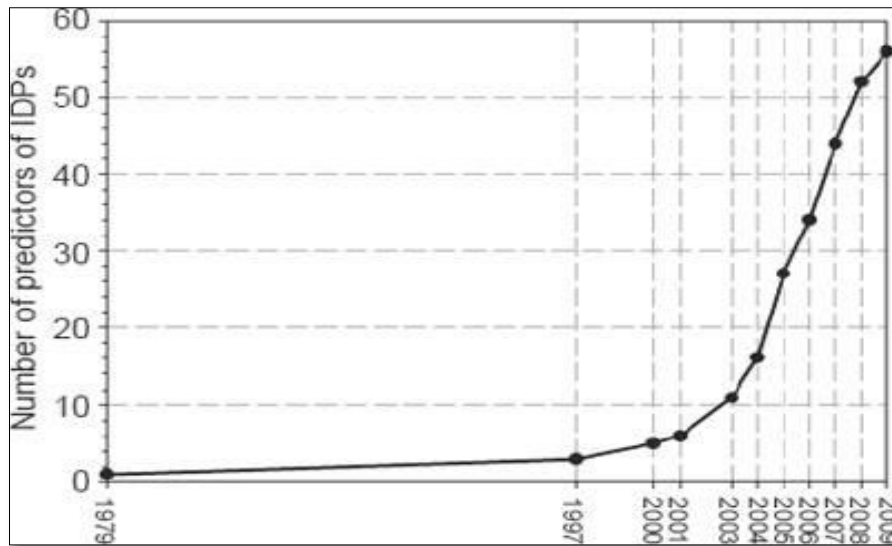
There is further advancement in the approaches used which is a computational tool described for assessment of proportionality between the functional additions in the data in the Swiss-Prot database and the anticipated amount of disorder in that particular protein [32-34]. There is also an advancement which relies on the hypothesis that if a function that is elaborated by some specific word relies on intrinsic disorder, than protein linked with that word is anticipated to have greater amount of disorder if comparison is done with proteins that are chosen randomly from the database of Swiss-Prot. When this tool was applied to 710 Swiss-Prot keywords in various datasets it was concluded that, 310 functional keywords were linked to ordered proteins, 238 functional keywords were linked to disordered proteins, and left over 162 keywords depicted uncertainty [32-34].

### 2.3 Computer-aided search for IDPs

Peculiarities of amino acid sequences of the IDPs are very important as many computational tools for disorder prediction rely on them. With over 70% success rate of initial predictors with passing years, the anticipation certainty is increasing. Recently 80% accuracy has been achieved by various machine-learning algorithms for predicting structure and disorder in proteins [37].

To date, above fifty diverse computational tools have come into picture for disorder prediction [38]. Given below is the tabulated form of the most widely recognised predictors and the parameters they rely on [39, 40].

Predictor	Principle
PONDR VSL2	SVMa with non-linear kernel
DISOPRED2	SVM, NNb for smoothing
IUPred	Estimated pairwise interaction energy
DisEMBL	Neural network
GlobPlot	Amino acid propensity, preference for ordered secondary structure
FoldUnfold	Amino acid propensity
FoldIndex	Amino acid propensity
NORSp	Secondary structure propensity
PreLink	Amino acid propensity, hydrophobic cluster analysis.



**Figure 2.1: Increase in number of disorder predictors from 1979 to 2009 [16].**

As the number of disorder predictors is increasing, diversity in prediction ideas is also increasing. And hence there is complexity in computing methods and attributes used for training which causes difficulty in selecting the perfect tool for disorder prediction. With different predictors providing different results for the query protein aggravates this complexity. To solve this complexity meta-predictors were introduced. In such method the anticipated results are taken from various other predictors and are used as inputs in meta-predictors. This advancement has shown to give more prediction accuracy and hence better results [37, 41, 42].

## 2.4 Machine Learning Algorithms

Machine learning is defined as the ways that can recognize particular patterns in the data, and then find the unspecified patterns for anticipating the query data. Also it performs some types of decision making during ambiguity. Machine learning is divided into two types:

- 1) Supervised learning-** In supervised learning approach, the learning is predictive. It aims to learn and get trained from inputs  $x$  to predict the target outputs  $y$  in a training dataset called  $D$  which contains both the inputs and targets data. The inputs are also termed as features or attributes.
- 2) Unsupervised learning-** In unsupervised learning, the learning is descriptive. In this approach only inputs are given to recognize patterns in a data and hence it is also called

knowledge discovery. Since the kinds of patterns one is looking for is untold one cannot compare prediction of  $y$  for a given  $x$ .

Also, since the most advanced approach to disorder prediction are machine learning (ML) algorithms, i.e. predictors trained to distinguished sequences that encode ordered or disordered structures, here, in this study supervised learning technique is used and neural network is the most widely used algorithm for that. Compared to the previous simpler approaches, disorder predictors use amino acid features and hidden sequence properties, which explain the superior performance. The most famous ML algorithm is PONDR (predictor of natural disordered regions), a neural network (NN) algorithm, which is based on local amino acid composition, flexibility and other sequence features [43]. PONDR has been developed into several variants, enabling prediction of disorder within the terminal regions of proteins [44]. It also helps in prediction of regions which have possibility to function like recognition motifs, for example, VL-XT [27]. Another variant is combined prediction of both short and long regions of disorder, for example, VSL2 [45]. The short disordered regions, lack of structure depends on their structural environment, whereas disorder of long regions stands on its own, hence, this combined approach results in one of the most powerful algorithms of disorder prediction.

Another computationally different ML approach is the application of support vector machines (SVMs), for example, DISOPRED2 [46]. This algorithm searches for a hyperplane in a feature space that separates ordered and disordered proteins. The hyperplane may either be linear or non-linear. In this unbalanced class frequencies of data from ordered (e.g. proteins in PDB) and disordered (e.g. proteins in DisProt database) proteins are taken into consideration. Sequence profiles generated by PSI-BLAST are also incorporated as an input.

## 2.5 Comparison of Disorder Prediction Methods

The performance of disorder predictors depends on their criteria of evaluation, and also on the datasets used for evaluation. Basically, there are two factors limiting direct comparison of predictors:

- (1) Comparisons cannot be based on simple percent values of predicted disorder, because the number of positive hits can be easily increased at the expense of false positives, i.e. predicting disorder for ordered regions.



(2) The amount of data on order and disorder differ significantly, which is difficult to handle when prediction accuracies are simply compared.

Accordingly, the performance of methods is compared by different measures, basically by calculating sensitivity and specificity, i.e. the ratio of correctly predicted disorder vs. the ratio of incorrectly predicted ordered regions. To reach the dependable assessment of disorder, it is recommended that several predictors based on different principles should be used.

Hence, different predictors perform at different levels, and it was demonstrated in the critical assessment of structure prediction algorithms experiments using CASP 6 [47] and CASP 7 [48] proteins.

# **Chapter 3**

## **Review of Literature**

### 3.1 Overview

It is clear from the elaboration of IDPs in previous chapters that their amino acid sequences and compositions are quite distinct from those of ordered proteins. This makes their recognition easier at protein level assessment. For this different tools have been developed. They predict and study the sequential parameters and characteristics of IDPs. For this lot of experimental investigations have been done and elaborated in IDPs which are discussed in this chapter.

### 3.2 Literature Review

**Tompa et al. [49]** studied that the success of the protein structure–function paradigm suggested that a protein could only function efficiently with a proper three-dimensional structure. This view is based on more than 60,000 high-resolution protein structures in the Protein Data Bank. However, many observations differ from the paradigm for a novel family of proteins, which apparently exist and function without a well-defined structure. They studied that among a range of physical techniques, NMR provides the most detail and insight into the structural ensemble of IDPs, and it is also potential of reporting the *in vivo* state and interactions of these proteins.

**Babu et al. [5]** studied the IDPs regulation and their altered expression association with many diseases. Recent studies show that IDPs are tightly regulated and that dosage-sensitive genes encode proteins with disordered segments. They found that the tight regulation of IDPs contributes to signaling loyalty and ensures that IDPs are available in proper amounts and not present after they are not needed. This altered availability of IDPs results in hiding of proteins through non-functional interactions involving disordered segments (i.e., molecular titration), thereby causing an imbalance in signaling pathways. They also discussed the address implications for signaling, disease and drug development, and also outlined the directions for future research.

**Uversky et al. [16]** described the family of intrinsically disordered proteins and their members who fail to form rigid 3-D structures under physiological conditions, either along their entire lengths or only in localized regions. They reported these proteins or regions exist as dynamic and their backbone Ramachandran angles exhibit extreme temporal fluctuations without specific equilibrium values. They discussed that the protein structure-function

paradigm has to be expanded to include intrinsically disordered proteins and alternative relationships among protein sequence, structure, and function. This shift in the paradigm represents a major breakthrough for biochemistry, biophysics and molecular biology, as it opens new levels of understanding with regard to the complex life of proteins.

**Xue et al. [50]** developed a meta-predictor of intrinsically disordered amino acids i.e. PONDR-FIT. They introduced a consensus artificial neural network (ANN) prediction method, which was developed by combining the outputs of several individual disorder predictors. They used eight-fold cross-validation to improve the prediction accuracy over a range of 3 to 20% with an average of 11%, depending on the datasets being used. Analysis of the errors showed that short disordered regions with less than ten residues are still error prone, as well as for the residues close to order/disorder boundaries. Deep learning of the underlying mechanism by which such meta-predictors give improved predictions will encourage for the further development of protein disorder predictors.

**Spath et al. [51]** found that the charge interactions can dominate the dimensions of IDPs. Quite a lot of eukaryotic proteins are disordered under physiological conditions, and fold into ordered structures only by binding to their cellular targets. Such IDPs often contain a large fraction of charged amino acids. They investigated the influence of charged residues on the dimensions of unfolded proteins and IDPs by using single-molecule Förster resonance energy transfer. They found that in contrast to the compact unfolded conformations that were observed for many proteins at low denaturant concentration, IDPs exhibit a significant expansion at low ionic strength that correlates with their net charge. Charge-balanced polypeptide exhibits an additional collapse at low ionic strength, as predicted by polyampholyte theory. This noticeable effect of charges on the dimensions of unfolded proteins has important significance for the cellular functions of IDPs.

**Uversky et al. [28]** demonstrated that the function of IDPs complements the functional of ordered proteins. IDPs are involved in regulation, signaling and control. IDPs are highly abundant in various human diseases, including neuro-degeneration and other protein dysfunction maladies and hence are attractive novel drug targets. They discussed the aspects of IDPs, as well as their roles in neuro-degeneration and protein dysfunction diseases. Also they discussed the peculiarities of IDPs as potential drug targets.

**Eliezer [52]** did biophysical characterization of IDPs, they studied hurdles of the structural characterization of disordered. The best suited technique for providing high-resolution structural information of IDPs is NMR spectroscopy. Other methods like Optical methods, solid state NMR, ESR and X-ray scattering can also provide valuable information regarding the conformations sampled by disordered states. He discussed recent advances in the applications of these methods to IDPs.

**Sickmeier et al. [9]** described the Database of Protein Disorder i.e. DisProt which links structure and functions information for intrinsically disordered proteins (IDPs). They defined IDPs as a protein that contains at least one experimentally determined disordered region. They described although lacking fixed structure, IDPs and regions carry out important biological functions. One of the major short-coming in the study of IDPs was the lack of organized information. DisProt was hence developed to facilitate the research on IDPs by collecting and organizing knowledge regarding the experimental characterization and the functional associations of IDPs at a single platform. It is a database for unique source of biological information on IDPs and it opens doors for significant bioinformatics studies.

**Linding et al. [53]** did comparative study of the relationship between protein structure and  $\beta$ -Aggregation in globular and IDPs. They used the algorithm TANGO to compare the  $\beta$  aggregation tendency of a set of globular proteins derived from SCOP with a set of 296 experimentally verified, non-redundant IDPs and also a set of IDPs predicted by the algorithms like DisEMBL and GlobPlot. Their analysis showed that the  $\beta$ -aggregation propensity of all- $\alpha$ , all- $\beta$  and mixed  $\alpha/\beta$  globular proteins as well as membrane-associated proteins is fairly similar. They also discussed that although IDPs have a much lower aggregation propensity than globular proteins, this does not mean that they have a lower potential for amyloidosis.

**Yang et al. [54]** developed the regional order neural network (RONN) software as an application of their algorithm based on 'bio-basis function neural network' pattern recognition. It was used for the detection of disordered regions in proteins. When 80 protein sequences were derived from PDB, were tested on different prediction tools based on property access measure, it was shown that RONN performed the best amongst the panel of nine disorder prediction tools.

**Mathura et al. [17]** investigated the determinants of protein order and disorder. They found as compared to the ordered protein, the intrinsically disordered segments in all four databases were significantly depleted in W, C, F, I, Y, V, L and N and rich in A, R, G, Q, S, P, E and K, and inconsistently different in H, M, T, and D, suggesting that the first set be called order-promoting and the second set disorder-promoting. They also ranked 265 amino acid properties by their ability to discriminate order and disorder and then pruned to remove the most highly correlated pairs. The highest-ranking properties after pruning consisted of residue contact scales, hydrophobicity scales, scales associated with  $\alpha$ -sheets and one polarity scale. These properties were used to suggest that disorder characterized by NMR and CD is the most similar, by CD and X-ray being next, and by NMR and X-ray is the least similar.

**Romero et al. [25]** designed a neural network based predictor for general disorder estimation and tested on data built from several public domain data banks through a nontrivial search, statistical analysis and data dimensionality reduction. Also they developed predictors for identification of family-specific disorder by extracting knowledge from databases generated through multiple sequence alignments of a known disordered sequence to other highly related proteins. They demonstrated that long disordered regions are common in nature, with an estimate that 11% of all the residues in the Swiss Protein data bank belong to disordered regions of length 40 or greater.

**Dunker et al. [24]** studied examples of five proteins i.e. fd phage, nucleosome, clusterin, calcineurin, calsequesterin and found disordered polyanion tails at the carboxy terminus bind many of calcium ions, perhaps without adopting a unique structure. They also studied disordered regions of 16 more proteins which included molecular recognition domains, protein folding inhibitors, flexible linkers, entropic springs, entropic clocks, and entropic bristles. Also, they studied the relationships between amino acid sequence and order/disorder, and from this information they predicted intrinsic order/disorder from amino acid sequence.

**Xie Q. et al. [55]** gave the sequence attribute method for determining relationship between sequence and protein disorder. They plotted of  $P(s|x)$  (conditional probability) versus  $x$  to provide information about the correlation between the given sequence attribute and disorder or order. These plots allowed quantitative comparisons between individual attributes for their ability to discriminate between order and disorder states. Using such quantitative comparisons, 38 different sequence attributes were rank-ordered. Attributes based on cysteine, the aromatics, flexible tendencies, and charge were found to be the best.

**Romero et al. [43]** worked on how to identify disordered regions in proteins from amino acid sequence. For that they developed a rule-based and several neural networks. The rule-based predictor was suitable only for very long disordered regions, whereas the neural network predictors were developed separately for short-, medium-, and long-disordered regions (S-, M-, and LDRs, respectively). When rule-based and LDR neural network predictors were applied to large databases of protein sequences it provided strong evidence that disordered regions are very common in nature.

### 3.2 Conclusion

From the discussed literature review done in this study, it can be concluded that, first, the structure and function of proteins depends on the degree of orderness and disorderness in them. Second, the disorder of proteins depends on few characteristic peculiarities in their sequences. Third, the disorder also depends on few parameters like hydrophobicity, propensity, charge, bulkiness etc. Lower the hydrophobicity more is the disorderness. Fourth, many disease related proteins have been linked with IDPs till date, and so it can be predicted that disorderness in the protein is the cause of alteration in the function of normal protein and hence the pathogenesis of the disease. And fifth, due to small success rate in the study of IDPs by biological approaches, computational approaches like machine learning algorithms mainly neural network are widely used with more accuracy in prediction of disordered regions.

# **Chapter 4**

## **Materials and Methods**



## 4.1 Objective of this study

The objective of this study done is to link the proteins responsible for the pathogenesis of the nine CAG repeat diseases with the IDPs and predict that they have regions of disorder in them by comparing them with a set of ordered and also with a set of disordered proteins. The results are analysed through regression plots because such plots are helpful in studying if a pattern or a trend of some property is taken under consideration. Here we are taking the hydropathy scales of the amino acid of proteins as our parameter of comparative analysis.

Efficiency and accuracy of any study done depends on the data collected for conducting that study and the method used for it. In this chapter the data resources and materials required to conduct this study has been covered.

## 4.2 Data Resources

For the present study data has been taken from the following resources:

### 1) NCBI

National Centre of Biotechnology Information houses different databases from where various different types of data can be retrieved. The protein sequences required in this study were fetched from here by selecting proteins in search drop down menu. The sequences were in FASTA format. Also it contains various tools to work on the retrieved sets of data. The Pubmed database of NCBI maintains various papers which contains studies conducted by different scientists in a particular field.

### 2) UniprotKB/Swiss-prot

It is a high quality, manually annotated database which is freely accessible for protein sequence and its functional information. Also it is the reviewed part of the Uniprot Knowledgebase in which many entries are derived from genome sequencing projects. It contains non-redundant protein sequence data and a vast amount of information about the biological function of proteins derived from the research literature. It also keeps various experimental findings, computed results and various conclusions on a single platform.

### 3) Disprot Database

The database of proteins disorder (Disprot) is a curated database which provides information about the proteins that lack fixed three-dimensional structure. It is maintained by Center for Computational Biology and Bioinformatics at Indiana University School of Medicine and Center for Information Science and Technology at Temple University. It contains lists of various regions which are disordered in FASTA or XML format, which are modified with passing years, the most recent being of 2013 [9].

### 4) PDB Select\_25

It is the list of ordered proteins or stable proteins having the range of chains with length of 25-10000. This list contains 3119 chains with 356088 residues having less than 25% similarity with each other. [56-58]

## 4.3 Tools used

### 4.3.1 MATLAB R2014a

The name MATLAB stands for MATrix LABoratory. MATLAB was developed originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects. MATLAB [59] is a high-performance language for technical computing.

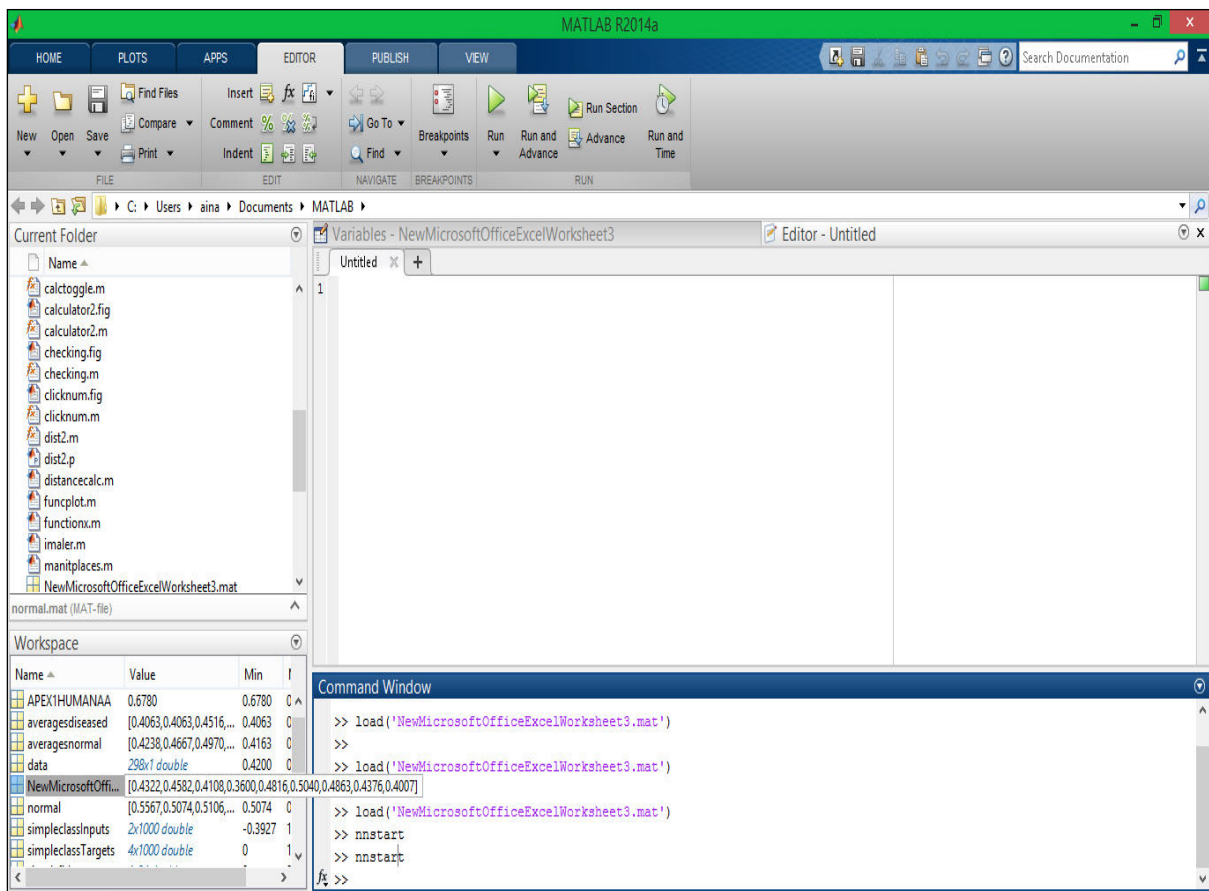
Furthermore, MATLAB is a modern programming language environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research. MATLAB has many advantages compared to conventional computer languages (e.g., C, FORTRAN) for solving technical problems.

MATLAB is an interactive system whose basic data element is an array and does not require dimensioning. Its software package has been available commercially since 1984 and is now regarded as a standardized tool at most universities and industries worldwide. It has built-in applications that enable a very wide variety of computations which are very powerful. It also has user-friendly graphics commands that make the visualization of results immediately possible.

Many specific routines applications are collected in packages referred to as ‘toolbox’. The different types of toolboxes are for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering. MATLAB is a tool that integrates numerical computation and visualization. The basic data element is a matrix or array, so if one needs a program that manipulates array-based data it is generally fast to write and run in MATLAB.

### Starting MATLAB

This is the window in which one interacts with MATLAB. The main window on the right below is called the *Command Window*. One can see the command prompt in this window, which looks like `>>`. If this prompt is visible MATLAB is ready to enter a command. Also on the right is a *Editor Window or Script*. In the top left corner *Current Folder* window can be viewed. At the below left is the *Workspace window*. The *Workspace window* will shows all variables that are used in the current MATLAB session.



**Figure 4.1: The graphical interface to the MATLAB workspace.**

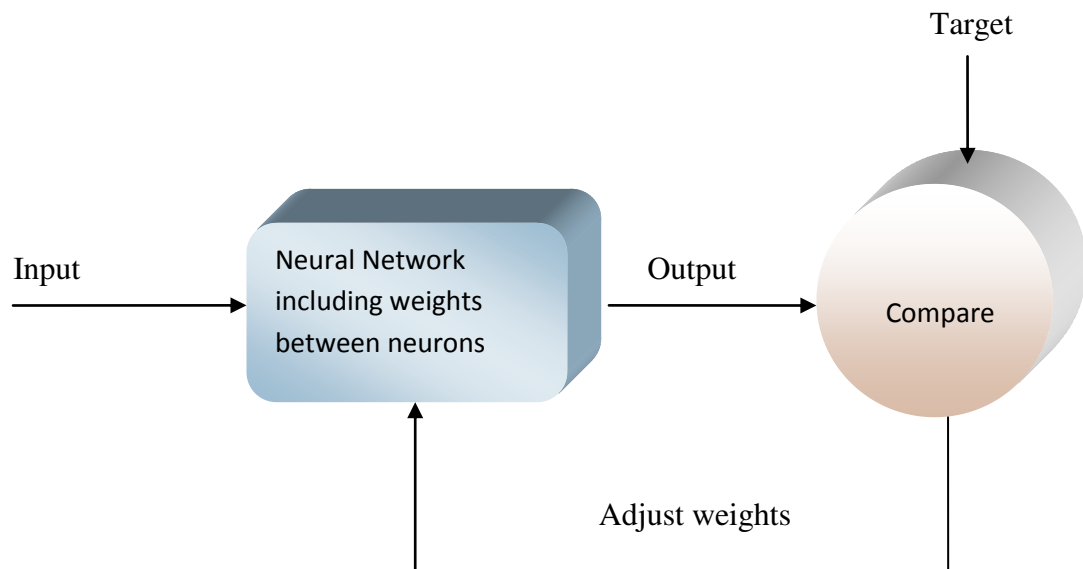
## Features of MATLAB

Following are the basic features of MATLAB:

- 1) Provides an interactive environment for iterative exploration, design and problem solving.
- 2) Has built-in graphics for visualizing data and tools for creating custom plots.
- 3) Provides tools for building applications with custom graphical interfaces.
- 4) Its programming interface provides tools for improving code quality and maintainability and maximizing good performance.
- 5) Provides vast platform for mathematical functions for linear algebra, statistics, fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- 6) Helps in joining MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

### 4.3.2 Neural Network Toolbox™

Neural Networks consists of simple elements like biological neurons that work parallel to each other. One can train a NN to perform a function by adjusting the values of the weights between elements and these weights determine the network function.



**Figure 4.2: Basic Neural Network**

Usually, NN weights are adjusted, or it is trained, so that certain input leads to desired target output. Until the network output matches the target, the network is adjusted, after comparing with the output and the target, until the network output matches the target. Hence, large numbers of input and target pairs are needed to train a network.

NN are helpful solve problems that are difficult for or human beings conventional computers. This toolbox uses NN paradigms that are used for practical applications in engineering, financial, and other fields.

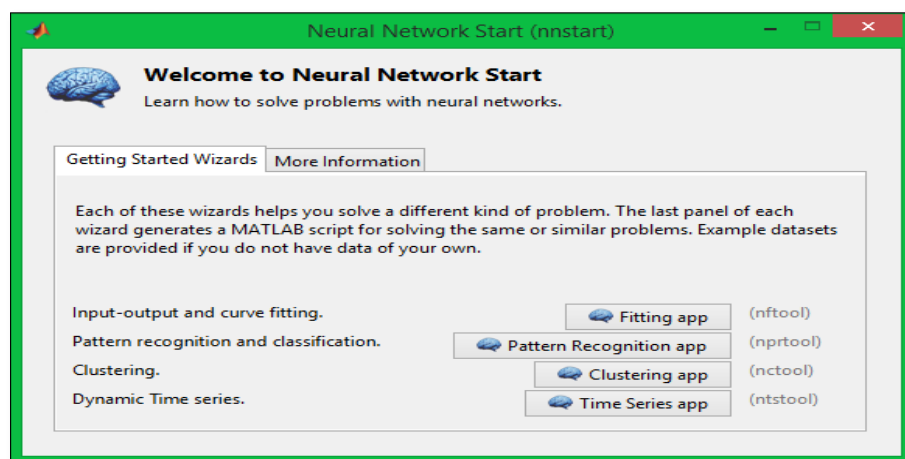
Neural Network Toolbox™ consists of apps and in-built functions. Using it complex nonlinear systems can be modelled which are not easily modelled using a closed-form equation. Using this toolbox one can do designing, training, visualization, and also simulate neural networks. In it supervised learning is supported via feed-forward, radial basis, and dynamic networks and unsupervised learning is supported via self-organizing maps and competitive layers. Another important use of Neural Network Toolbox is for applications

such as data fitting, pattern recognition, clustering, dynamic system modelling and control and also time-series prediction.

For speeding up the training part and handling of the large data sets, computations and calculations is distributed across multi-core processors, GPUs and computer clusters. Parallel Computing Toolbox™ is used for this. It also provides pre-processing and post-processing for improving the training efficiency and assessing network performance. It also gives modular network representation for managing and visualization of networks of arbitrary size Simulink® so that NN can be evaluated.

It has four kinds of wizards and apps-

- 1) Input/output curve fitting
- 2) Pattern recognition and classification
- 3) Clustering
- 4) Dynamic time series



**Figure 4.3: Neural Network Toolbox**

### Neural Fitting (nftool)

In a fitting problem one wants NN to work on the datasets of numeric inputs and set a of numeric targets. Examples include estimation of house pricing by input variables like tax rate, teacher/pupil ratio in local schools and crime rates which is called house datasets.

The NN fitting app helps to select data, create and train the network and evaluate its performance using mean square error and regression analysis.

A two layer feed-forward network is used which has sigmoid input neurons and linear output neurons to fit multi-dimensional mapping problems.

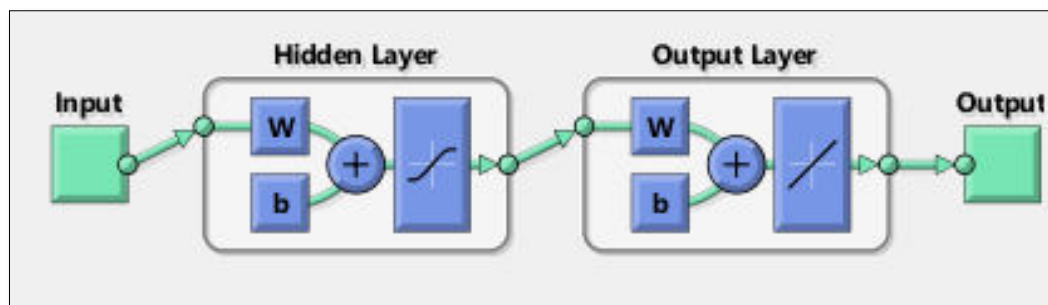


Figure 4.4: Two layer feed-forward network used by nftool.

#### Basic terminology:

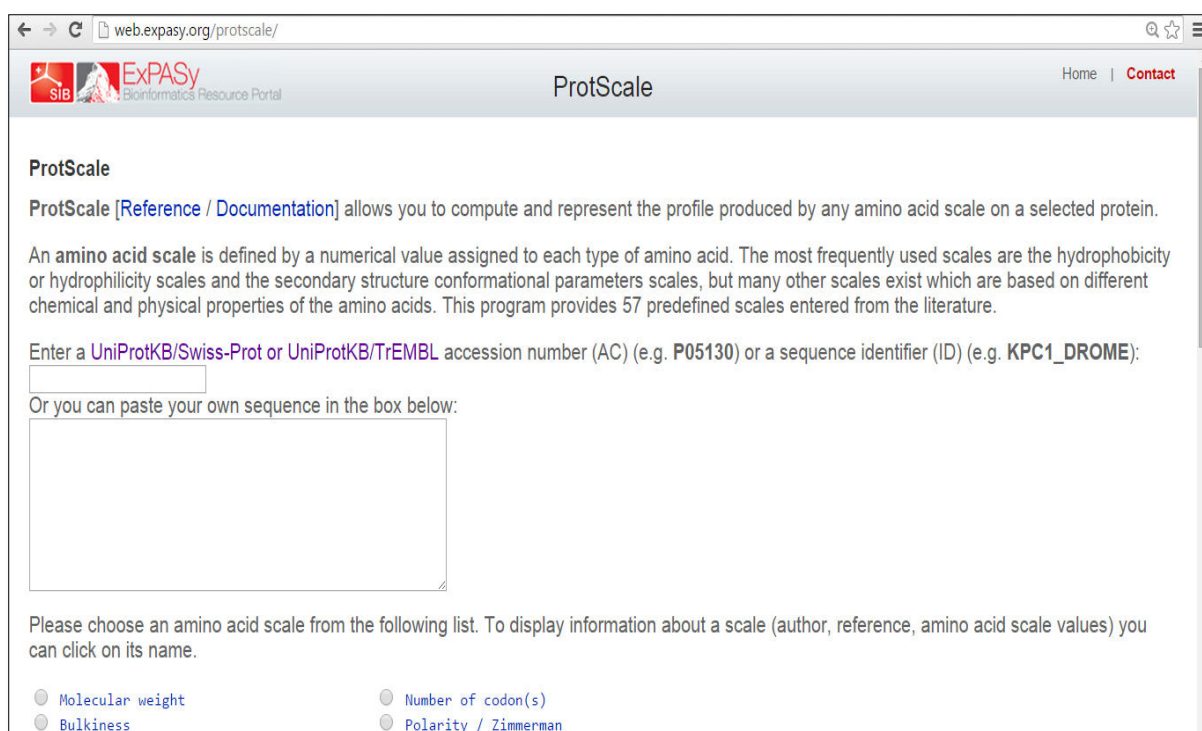
- 1) **Levenberg-Marquardt Algorithm-** Training was done using this algorithm. It takes more memory but less time. Training automatically stops when generalization stops improving, as indicated by mean square error of validation samples.
- 2) **Regression-** R values measure correlation between outputs and targets. An R value 1 means close relationships and 0 means no relationship.
- 3) **Mean square error-** Average squared difference between outputs and targets. Lower values are better. Zero means no error.

#### 4.3.3 ExPASy-ProtScale

It is a tool that helps to compute and represent (in the form of a two-dimensional plot) the profile (also called scales) produced by any amino acid scale of a selected protein.

An amino acid scale can be defined as a numerical value assigned to each type of amino acid. The most frequently used scales are hydrophobicity scales. Additionally, other scales also exist based on different chemical and physical properties of the amino acids.

There are 57 predefined scales entered from the literature in this tool. To generate data the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is assigned to the midpoint of the window.



The screenshot shows the web interface of the ProtScale tool at web.expasy.org/protscale/. The page has a header with the ExPASy logo and navigation links for Home and Contact. The main content area is titled "ProtScale" and contains a description of the tool, input fields for UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession numbers or sequence identifiers, and a list of amino acid scales to choose from. The scales listed are Molecular weight, Bulkiness, Number of codon(s), and Polarity / Zimmerman.

ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This program provides 57 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. P05130) or a sequence identifier (ID) (e.g. KPC1\_DROME):

Or you can paste your own sequence in the box below:

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

- ☐ Molecular weight
- ☐ Bulkiness
- ☐ Number of codon(s)
- ☐ Polarity / Zimmerman

**Figure 4.5: ExPASy-ProtScale Tool**



## 4.4 Methodology

The steps of the proposed framework are discussed in the subsections below:

### 4.4.1 Creation of the input and target datasets:

Two kinds of input datasets were created, one for the ordered proteins and other for the disordered proteins.

#### a) Input or ordered and disordered proteins dataset-

**Step1-** For this, 100 ordered proteins were chosen from PDBSelect\_25 and 100 disordered from Disprot Database\_2013 list of proteins.

**Step2-** The sequence of these proteins was then retrieved from the NCBI in FASTA format.

**Step3-** These sequences were then pasted one by one on the ExPASy-protScale tool and the parameters selected were-

1. Hydrophobicity scale (Kyte-Doolittle) [60]
2. Window size- 21 [60]
3. Normalization- Yes

The screenshot shows the ExPASy-protScale tool interface. It features a list of 40 radio buttons for selecting a hydrophobicity scale, arranged in two columns. The first column includes scales like Kyte & Doolittle, Abraham & Leo, Bull & Breese, Guy, Miyazawa et al., Roseman, Wolfenden et al., HPLC / Wilson & al, HPLC pH3.4 / Cowan, Rf mobility, TFA retention, retention pH 2.1, buried residues, Chothia, hetero end/side, flexibility, beta-sheet / Chou & Fasman, alpha-helix / Deleage & Roux, beta-turn / Deleage & Roux, alpha-helix / Levitt, beta-turn / Levitt, antiparallel beta-strand, A.A. composition, and Relative mutability. The second column includes scales like Manavalan et al., Black, Fauchere et al., Janin, Rao & Argos, Tanford, Welling & al, HPLC / Parker & al, HPLC pH7.5 / Cowan, HFBA retention, Transmembrane tendency, retention pH 7.4, accessible residues, Rose & al, average area buried, alpha-helix / Chou & Fasman, beta-turn / Chou & Fasman, beta-sheet / Deleage & Roux, Coil / Deleage & Roux, beta-sheet / Levitt, Total beta-strand, Parallel beta-strand, and A.A. comp. in Swiss-Prot. Below the scales, there is a 'Window size' dropdown set to 21, a text input for 'Relative weight of the window edges compared to the window center (in %)' set to 100, a 'Weight variation model' section with 'linear' selected over 'exponential', and a 'Do you want to normalize the scale from 0 to 1?' section with 'yes' selected over 'no'. At the bottom, there is a link for more information and 'Submit' and 'Reset' buttons.

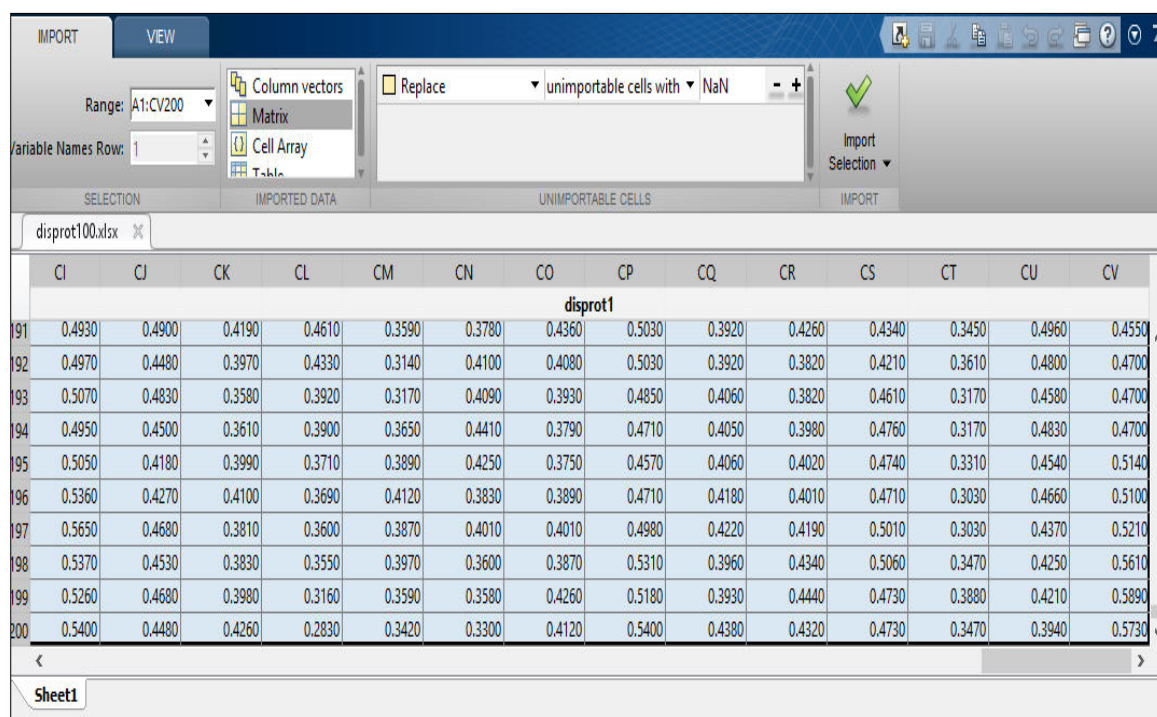
**Figure 4.6: Parameters to be set before calculating hydrophobicity scales.**

**Step4-** Then the sequences were submitted and the numerical format of the result of the query selected.

**Step5-** This was saved as a text file and converted to an excel file. Likewise these steps were repeated to create the excel sheet of hydrophobicity scores of 100 ordered proteins and another sheet of 100 disordered proteins.

**Step6-** The hydrophobicity scores of first 200 amino acids in a sequence were taken to make matrix of 200X100.

**Step7-** Both these excel files were then imported in matrix form in the MATLAB2014a software.



	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV
	<b>disprot1</b>													
191	0.4930	0.4900	0.4190	0.4610	0.3590	0.3780	0.4360	0.5030	0.3920	0.4260	0.4340	0.3450	0.4960	0.4550
192	0.4970	0.4480	0.3970	0.4330	0.3140	0.4100	0.4080	0.5030	0.3920	0.3820	0.4210	0.3610	0.4800	0.4700
193	0.5070	0.4830	0.3580	0.3920	0.3170	0.4090	0.3930	0.4850	0.4060	0.3820	0.4610	0.3170	0.4580	0.4700
194	0.4950	0.4500	0.3610	0.3900	0.3650	0.4410	0.3790	0.4710	0.4050	0.3980	0.4760	0.3170	0.4830	0.4700
195	0.5050	0.4180	0.3990	0.3710	0.3890	0.4250	0.3750	0.4570	0.4060	0.4020	0.4740	0.3310	0.4540	0.5140
196	0.5360	0.4270	0.4100	0.3690	0.4120	0.3830	0.3890	0.4710	0.4180	0.4010	0.4710	0.3030	0.4660	0.5100
197	0.5650	0.4680	0.3810	0.3600	0.3870	0.4010	0.4010	0.4980	0.4220	0.4190	0.5010	0.3030	0.4370	0.5210
198	0.5370	0.4530	0.3830	0.3550	0.3970	0.3600	0.3870	0.5310	0.3960	0.4340	0.5060	0.3470	0.4250	0.5610
199	0.5260	0.4680	0.3980	0.3160	0.3590	0.3580	0.4260	0.5180	0.3930	0.4440	0.4730	0.3880	0.4210	0.5890
200	0.5400	0.4480	0.4260	0.2830	0.3420	0.3300	0.4120	0.5400	0.4380	0.4320	0.4730	0.3470	0.3940	0.5730

**Figure 4.7: Snapshot of excel file imported in the form of matrix to create .mat file.**

**Step7-** .mat files of both the excel sheets were created by using save function in command window.

#### **b) Target or disease proteins dataset-**

**Step1-** Proteins which were responsible for causing 9 CAG repeat diseases were selected.

**Step2-** The sequences of these were retrieved from the NCBI in the FASTA format.

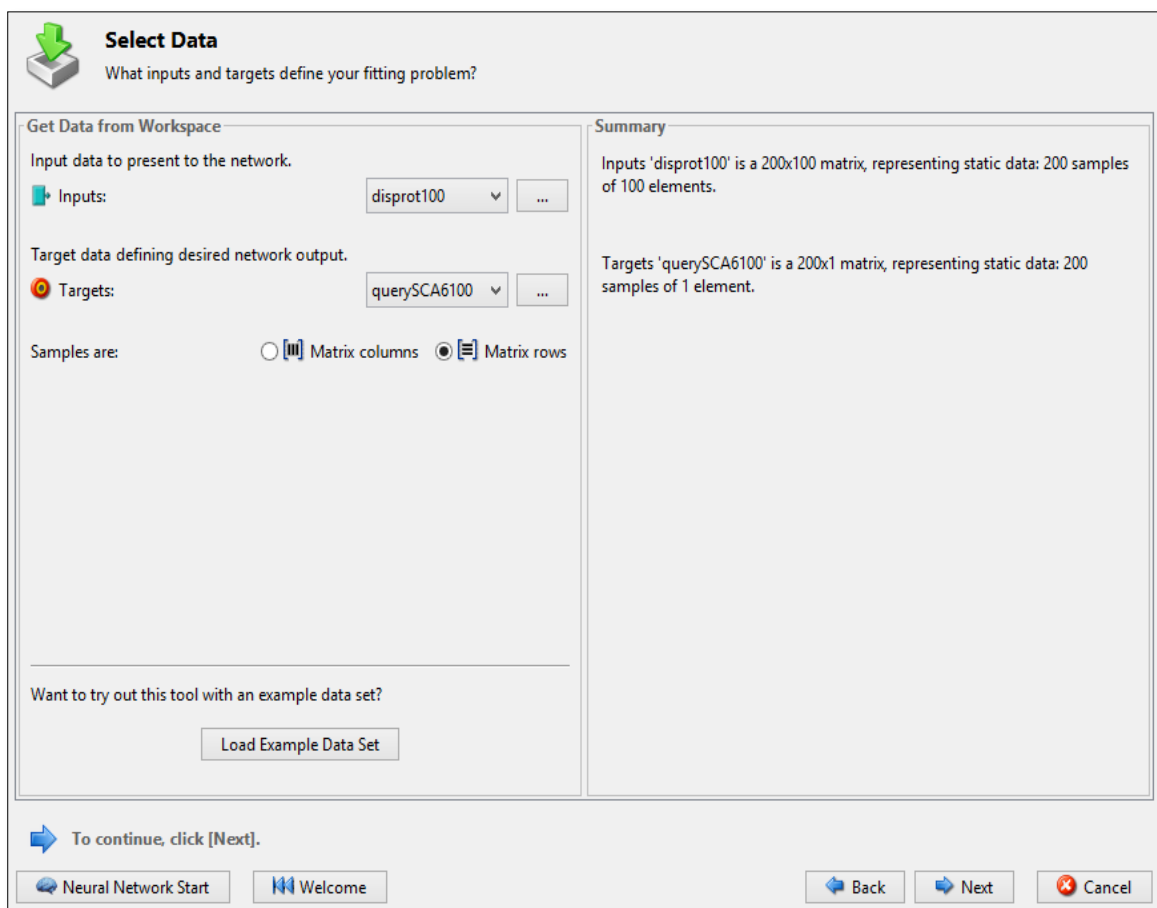
**Step3-** Rest all the steps were same as of input datasets to create .mat file of target datasets of 200X1.

#### 4.4.2 Training the datasets using nftool

**Step1-** In the command window of MATLAB2014a, nnstart command was given to open Neural Network Toolbox

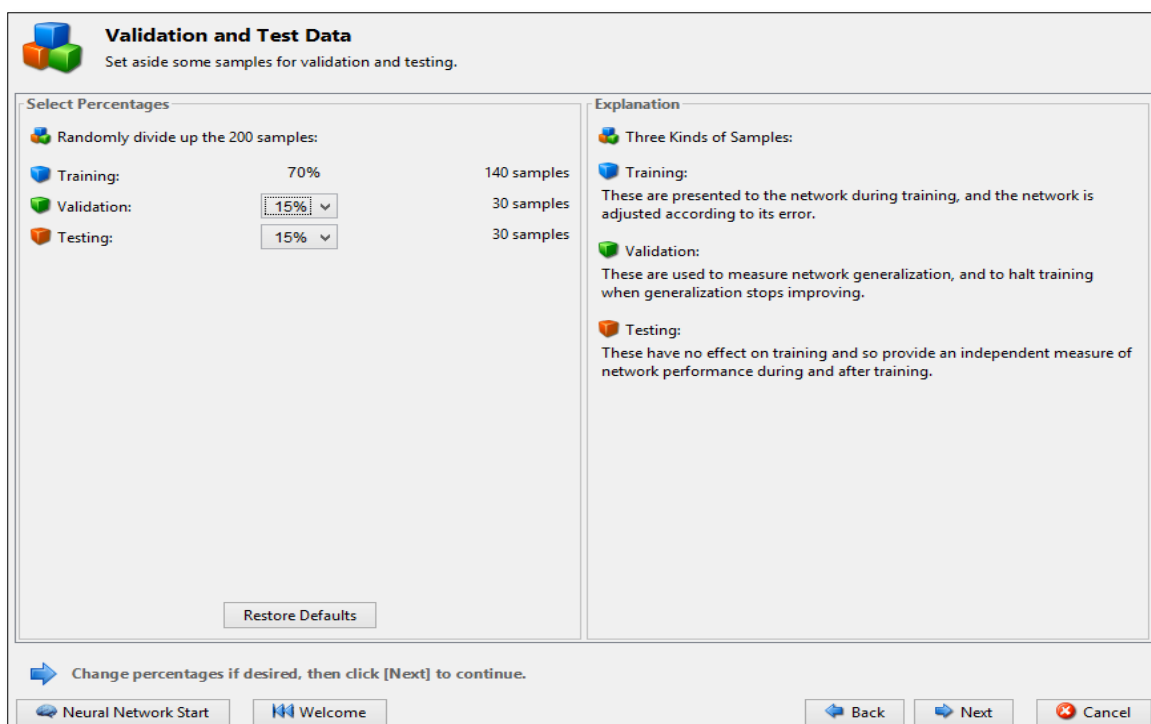
**Step2-** nftool was selected.

**Step3-** For input, .mat file of ordered/unordered proteins was selected and for target, .mat file of disease protein was selected.



**Figure 4.8:** Snapshot of inputs and targets being given in the form of Matrix rows.

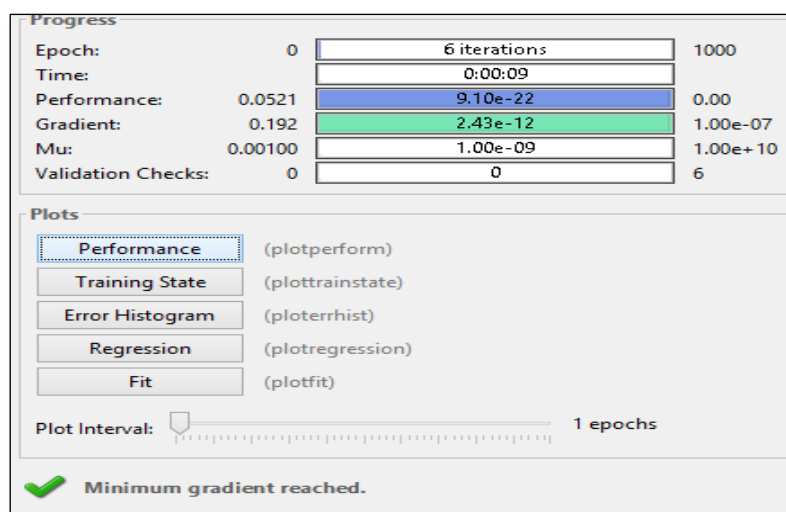
**Step4-** Validation and test data were selected as 15% both respectively, while training data being 70%.



**Figure 4.9: Division of hydrophobicity scores of for training, testing and validation.**

**Step5-** The datasets were then trained by selecting Levenberg-Marquardt Algorithm.

**Step6-** After that regression plot was obtained by clicking on plotregression.



**Figure 4.10: Regression plot was obtained using plotregression.**

## 4.5 Conclusion

Hence, it can be observed that the hydropathy scores of the nine proteins responsible for pathogenesis of the CAG diseases were obtained using ProtScale tool. Also similar scores were obtained for 100 ordered and 100 disordered proteins in the similar manner. These scores were used to create balancing and unbalancing datasets for training of the neural network using MATLAB 2014a Neural Network Toolbox<sup>TM</sup>. Finally, the regression plots were obtained using the same toolbox.

# **Chapter 5**

## **Results and Discussions**

## 5.1 Overview

In this study, the results were obtained through the analysis of regression plots. Regression lines in the plots are used to visually depict the relationship between the independent (x) and dependent (y) variables in the graph. A straight line depicts a linear trend in the data.

In addition to visually depict the trend in the data with a regression line, equation of the regression line is also calculated. This equation can be seen on the graph. How well this equation describes the data (the 'fit'), is expressed as a correlation coefficient, R. The closer R is to 1.00, the better the fit.

The prediction of disorder was done in proteins responsible for 9 CAG diseases on the basis of hydrophobicity scores. For that the scores of 9 proteins were individually trained against the scores of 100 ordered and 100 disordered proteins. The plots obtained are discussed in the following section.

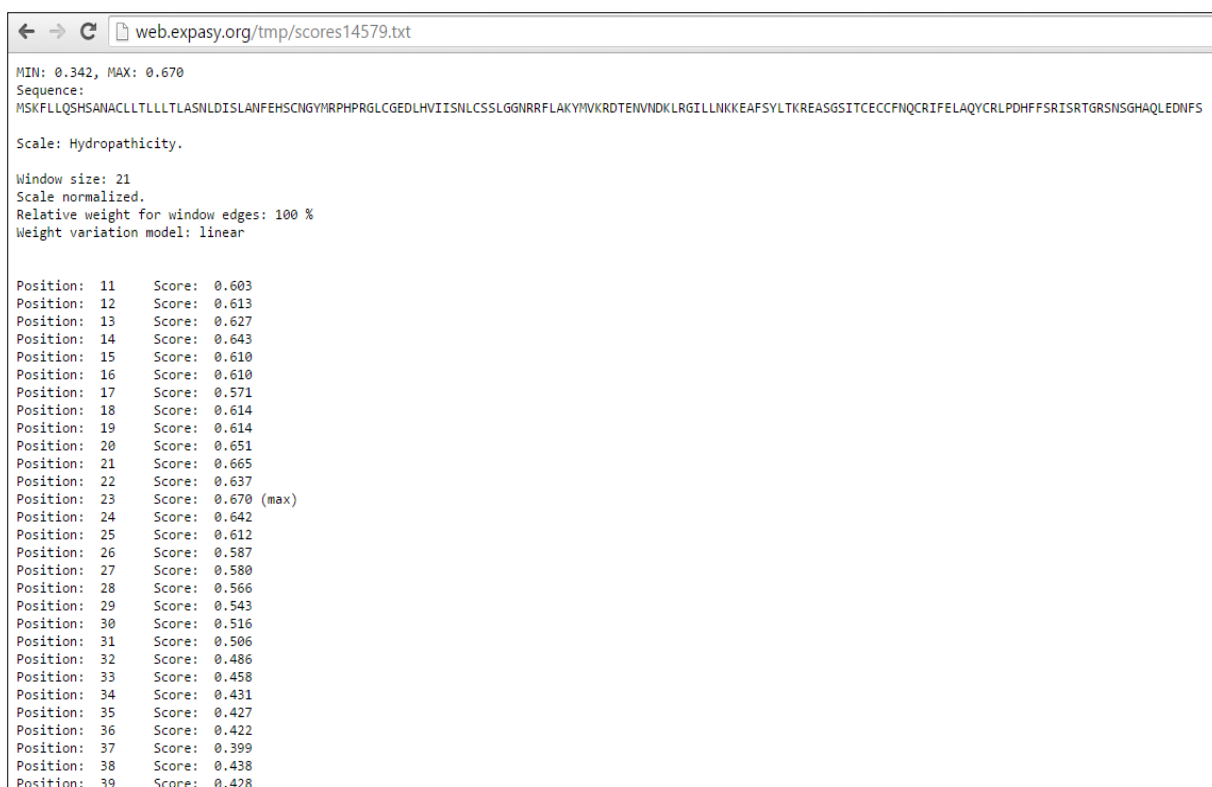
## 5.2 Results

### 5.2.1 Datasets

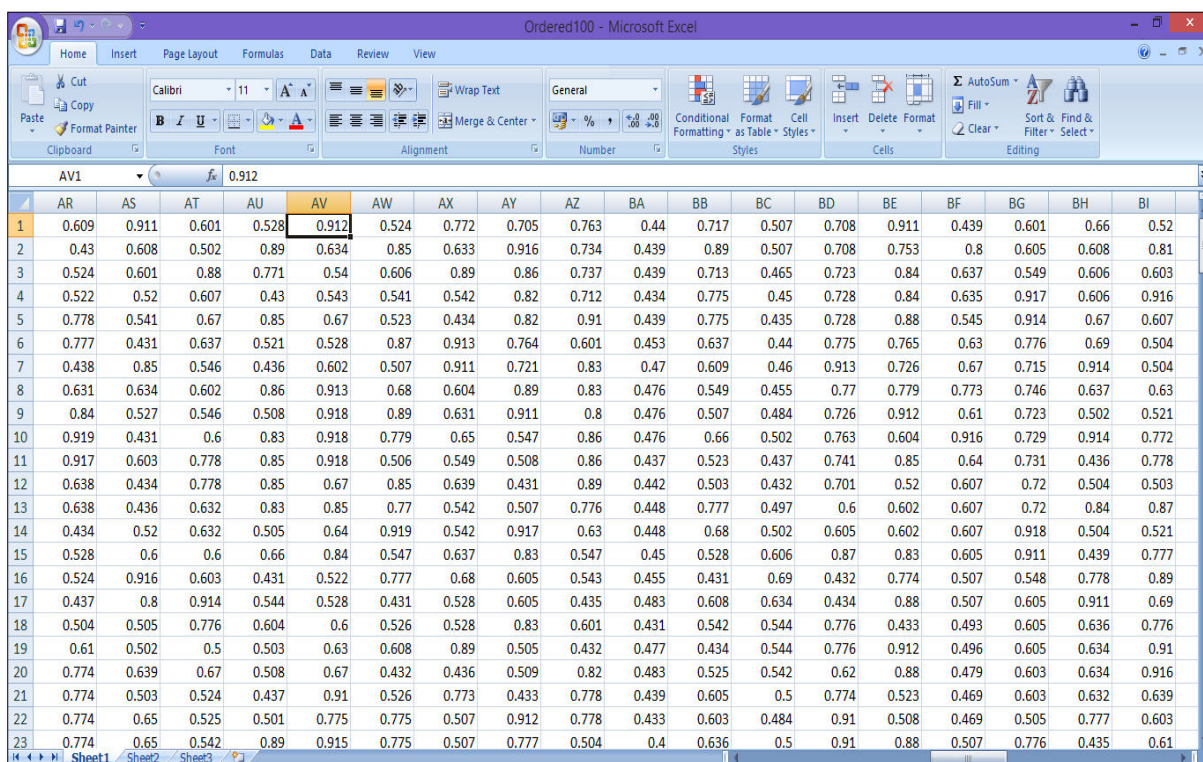
Two datasets of 200X100 were created separately for ordered and disordered proteins. And 9 datasets of 200X1 were created for the 9 proteins responsible for CAG diseases. The result of hydrophobicity scores always starts from 11<sup>th</sup> amino acid due to windowing which was taken to be 21. Windowing is selected on the basis that the length of the protein sequence should be greater than double the size of window.

The values were calculated through Kyte-Doolittle method, in such a manner that the pre-standardized hydrophobicity scores of first 21 amino acids were summed up and their average was assigned to the amino acid at the mid position of the window as shown in the figure below. These scores were obtained from ExPASy Prot-Scale Tool.

An exemplary protein hydrophobicity scores is shown below in the figure. Likewise, scores were obtained for 100 ordered and 100 disordered proteins. Also scores in similar manner were obtained for the 9 proteins responsible for CAG diseases.



**Figure 5.1: Snapshot of the result of the hydropobicity scores of a single protein.**



**Figure 5.2: Snapshot of dataset of 100 ordered proteins.**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0.36	0.62	0.471	0.447	0.387	0.469	0.521	0.387	0.469	0.593	0.575	0.442	0.453	0.362	0.567	0.524	0.605	0.508	0.424	0.535	0.4
2	0.37	0.631	0.459	0.433	0.358	0.457	0.525	0.358	0.457	0.579	0.587	0.424	0.467	0.334	0.552	0.502	0.591	0.477	0.434	0.504	0.4
3	0.404	0.664	0.476	0.461	0.353	0.457	0.497	0.353	0.457	0.579	0.589	0.409	0.467	0.344	0.577	0.474	0.596	0.467	0.472	0.474	0
4	0.428	0.663	0.505	0.459	0.355	0.459	0.525	0.355	0.459	0.568	0.605	0.407	0.5	0.331	0.545	0.474	0.609	0.461	0.456	0.499	0.4
5	0.444	0.623	0.547	0.433	0.357	0.418	0.523	0.357	0.418	0.569	0.598	0.424	0.459	0.374	0.54	0.512	0.623	0.473	0.456	0.538	0.4
6	0.468	0.592	0.549	0.448	0.398	0.434	0.501	0.398	0.434	0.597	0.598	0.439	0.445	0.346	0.54	0.502	0.581	0.484	0.476	0.499	0.4
7	0.508	0.568	0.595	0.461	0.412	0.468	0.517	0.412	0.468	0.587	0.593	0.441	0.429	0.346	0.551	0.459	0.548	0.461	0.476	0.526	0.4
8	0.506	0.525	0.6	0.504	0.374	0.448	0.556	0.374	0.448	0.59	0.563	0.439	0.429	0.384	0.565	0.416	0.557	0.431	0.481	0.568	0.4
9	0.492	0.525	0.612	0.489	0.392	0.494	0.517	0.392	0.494	0.59	0.531	0.439	0.412	0.42	0.532	0.458	0.559	0.401	0.481	0.552	0.4
10	0.533	0.542	0.586	0.446	0.394	0.493	0.493	0.394	0.493	0.59	0.522	0.452	0.406	0.439	0.491	0.472	0.52	0.375	0.464	0.566	0.4
11	0.531	0.508	0.592	0.47	0.375	0.454	0.505	0.375	0.454	0.546	0.498	0.438	0.389	0.469	0.48	0.501	0.488	0.414	0.466	0.519	0.4
12	0.546	0.534	0.594	0.446	0.395	0.477	0.478	0.395	0.477	0.546	0.458	0.454	0.371	0.485	0.48	0.462	0.45	0.452	0.462	0.501	0.4
13	0.585	0.532	0.59	0.421	0.434	0.448	0.447	0.434	0.448	0.517	0.417	0.497	0.371	0.485	0.462	0.463	0.437	0.439	0.46	0.515	0.4
14	0.596	0.543	0.592	0.423	0.429	0.462	0.46	0.429	0.462	0.503	0.421	0.469	0.359	0.5	0.469	0.463	0.461	0.444	0.436	0.487	0.4
15	0.582	0.516	0.581	0.461	0.445	0.45	0.482	0.445	0.45	0.501	0.421	0.471	0.401	0.502	0.452	0.468	0.507	0.444	0.45	0.474	0.4
16	0.582	0.538	0.593	0.456	0.451	0.45	0.497	0.451	0.45	0.525	0.424	0.485	0.396	0.536	0.439	0.435	0.479	0.434	0.464	0.49	0.4
17	0.628	0.494	0.593	0.498	0.465	0.449	0.52	0.465	0.449	0.511	0.41	0.475	0.392	0.52	0.423	0.413	0.44	0.432	0.459	0.492	0.4
18	0.656	0.493	0.579	0.46	0.466	0.401	0.503	0.466	0.401	0.525	0.449	0.475	0.364	0.495	0.399	0.413	0.427	0.449	0.443	0.49	0.4
19	0.661	0.493	0.565	0.474	0.466	0.405	0.501	0.466	0.405	0.538	0.449	0.485	0.392	0.515	0.425	0.406	0.415	0.449	0.443	0.5	0.4
20	0.688	0.462	0.524	0.504	0.425	0.4	0.468	0.425	0.4	0.505	0.419	0.454	0.406	0.491	0.437	0.406	0.446	0.465	0.399	0.495	0.4
21	0.668	0.437	0.489	0.488	0.438	0.415	0.506	0.438	0.415	0.501	0.419	0.432	0.407	0.483	0.437	0.44	0.456	0.465	0.433	0.49	0.4
22	0.656	0.442	0.465	0.488	0.452	0.431	0.49	0.452	0.431	0.501	0.404	0.422	0.389	0.483	0.435	0.466	0.497	0.465	0.416	0.459	0.4
23	0.617	0.42	0.463	0.474	0.48	0.403	0.476	0.48	0.403	0.525	0.361	0.453	0.347	0.497	0.435	0.499	0.498	0.477	0.376	0.456	0.4
24	0.602	0.386	0.463	0.474	0.499	0.415	0.476	0.499	0.415	0.523	0.376	0.451	0.337	0.503	0.411	0.499	0.526	0.488	0.376	0.485	0.4
25	0.563	0.411	0.476	0.463	0.497	0.396	0.472	0.497	0.396	0.51	0.4	0.492	0.303	0.473	0.425	0.51	0.486	0.493	0.378	0.481	0.4

Figure 5.3: Snapshot of dataset of 100 disordered proteins.

## 5.2.2 Regression Plots

The regression plots were obtained by training two layer feed-forward neural network using nftool of Neural Network Toolbox<sup>TM</sup> of MATLAB 2014a. For each CAG disease two plots were obtained, one by training the input by 100 ordered and other by 100 disordered proteins against the target diseased protein. Hence eighteen different plots were obtained which are illustrated and discussed below.

### 1) Atrophin-1 - DRPLA

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Atrophin-1 hydrophobicity scores and the R value obtained is close to 1, i.e. resultant  $R=0.97607$  and the validation R value was  $R=0.94603$ . While from the plot against ordered proteins it can be seen that the data does not fits the fits line properly and the R value was resultant  $R=0.32315$  and validation R value was  $R=0.23049$ .

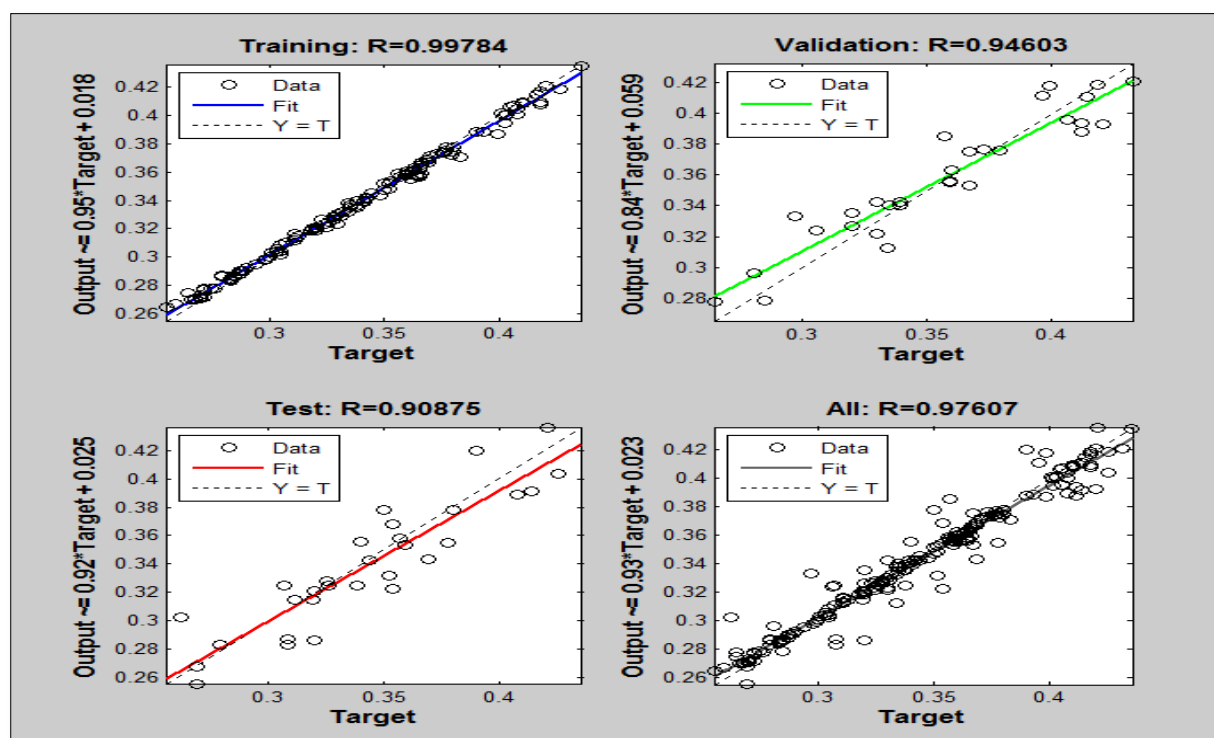


Figure 5.4(a): Snapshot of DRPLA protein Atrophin-1 against disordered proteins.

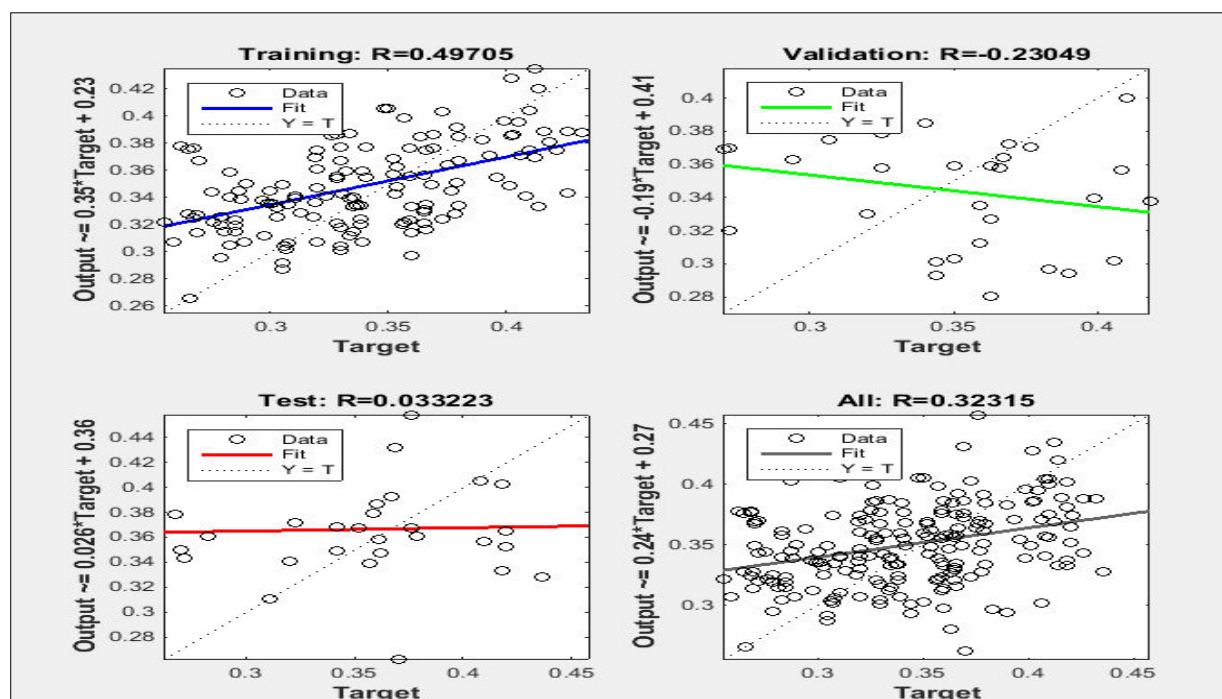


Figure 5.4(b): Snapshot of DRPLA protein Atrophin-1 against ordered proteins.

## 2) Huntingtin – Huntington’s disease

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Huntingtin’s hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.95287$  and the validation R value was  $R=0.90894$ . While from the plot against ordered proteins it can be seen that the data does not fits the fit line properly and the resultant R value was  $R=0.74341$  and validation R value was  $R=0.23158$ .

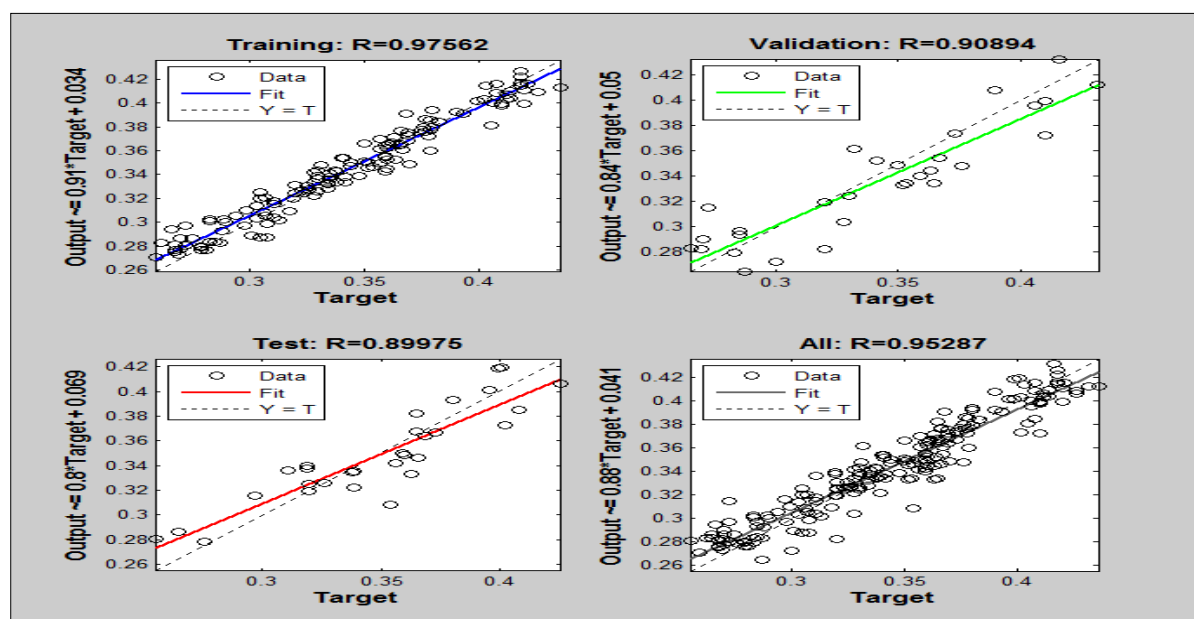


Figure 5.5(a): Snapshot of HD protein Huntingtin against disordered proteins.

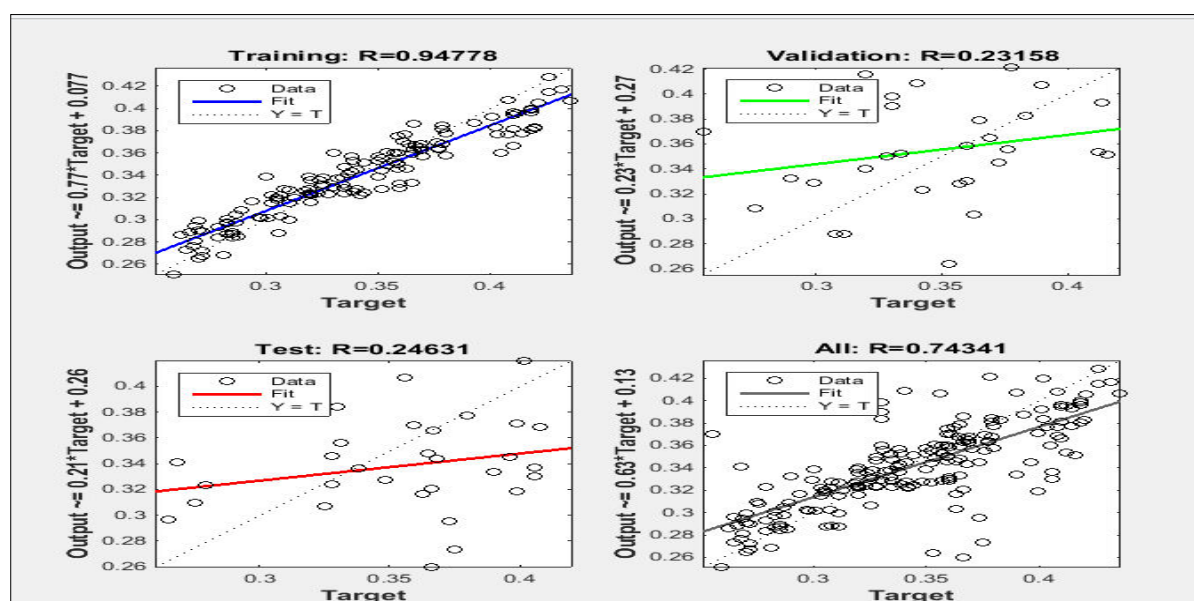


Figure 5.5(b): Snapshot of HD protein Huntingtin against ordered proteins.

### 3) Androgen Receptor(AR)- Kennedy's disease

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Androgen Receptor's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.98677$  and the validation R value was  $R=0.95716$ . While from the plot against ordered proteins it can be seen that the data fits at the values higher than 0.4 and the resultant R value was  $R=0.81914$  and validation R value was  $R=0.4844$ .

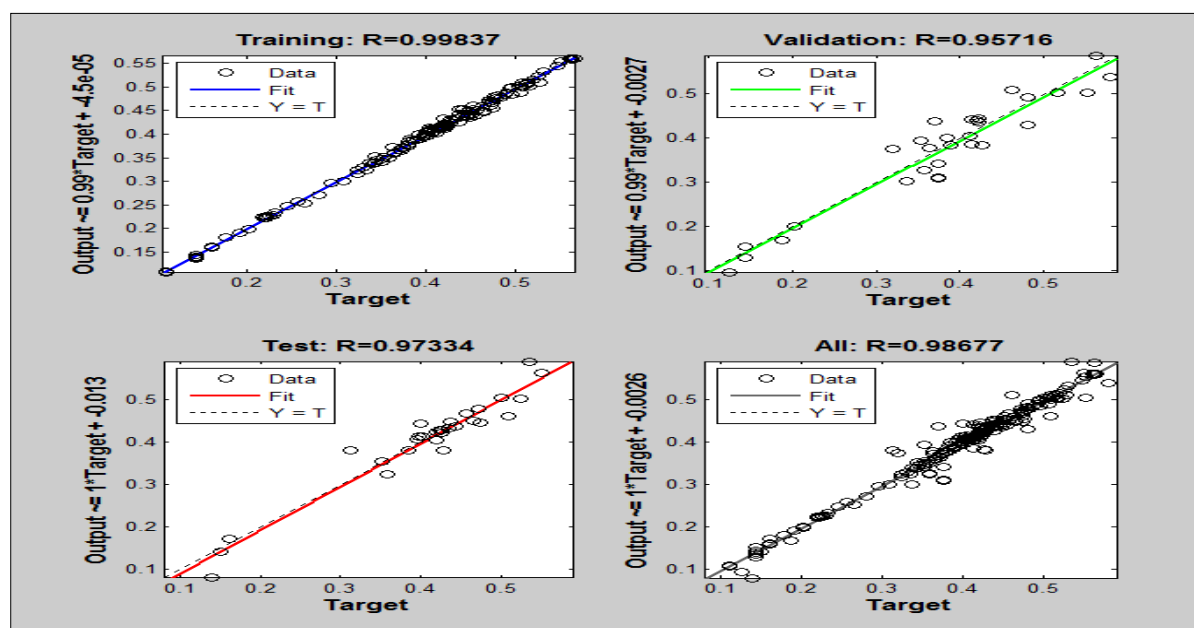


Figure 5.6(a): Snapshot of KD protein Androgen Receptor against disordered proteins.

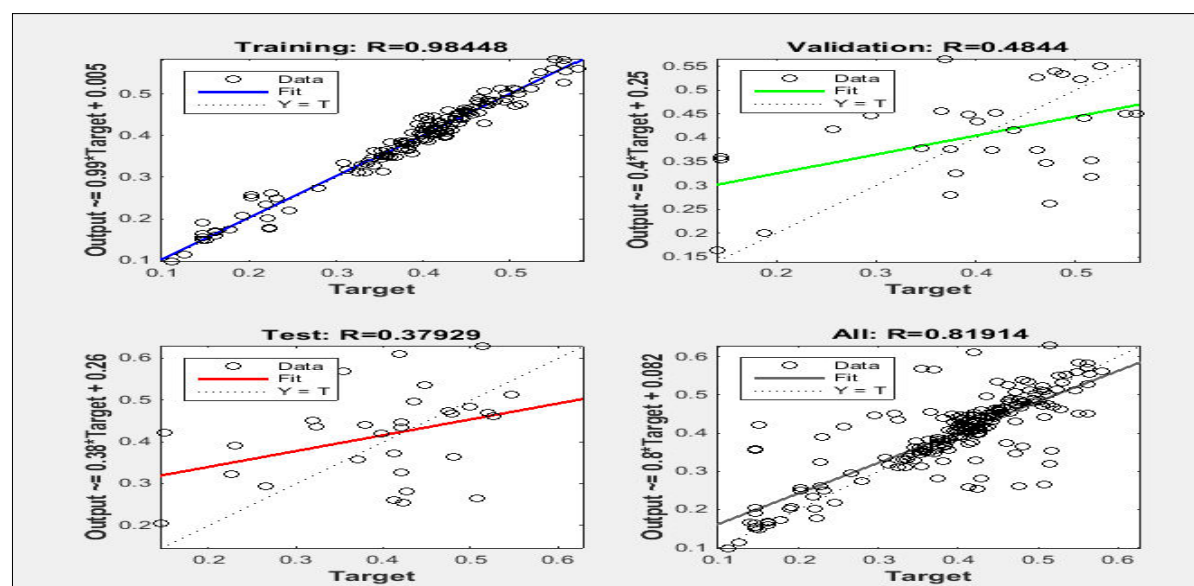


Figure 5.6(b): Snapshot of KD protein Androgen Receptor against ordered proteins.



#### 4) Ataxin-1– SCA1

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Ataxin-1's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.97977$  and the validation R value was  $R=0.91154$ . While from the plot against ordered proteins it can be seen that the data fits at the values higher than 0.5 and the resultant R value was  $R=0.58725$  and validation R value was  $R= -0.0636$ .

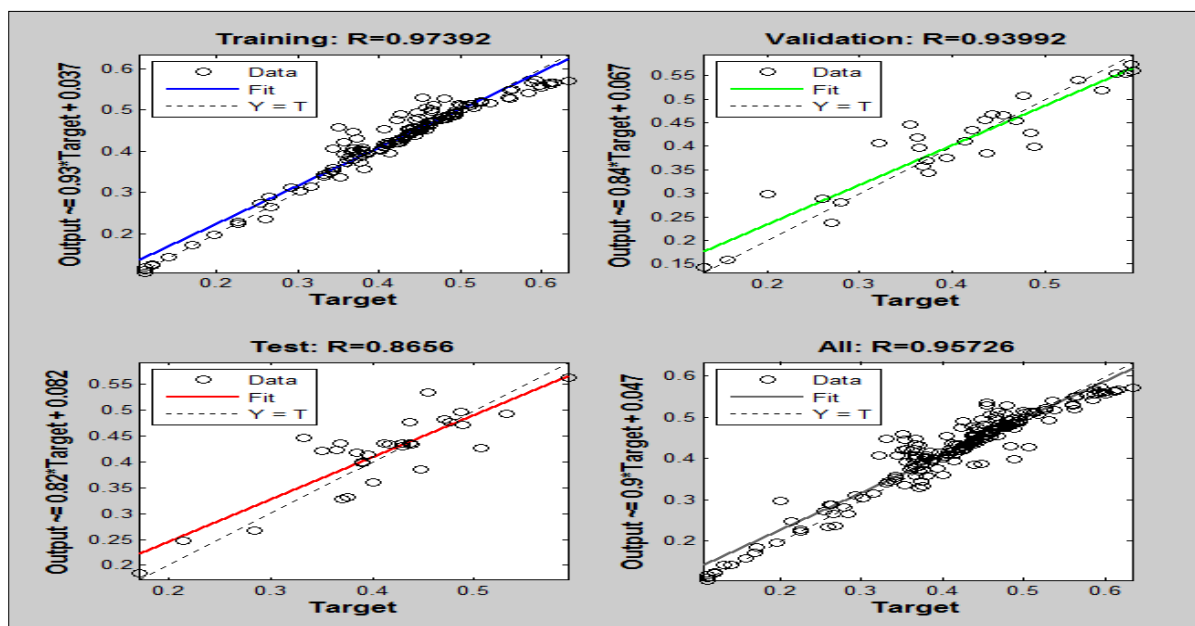
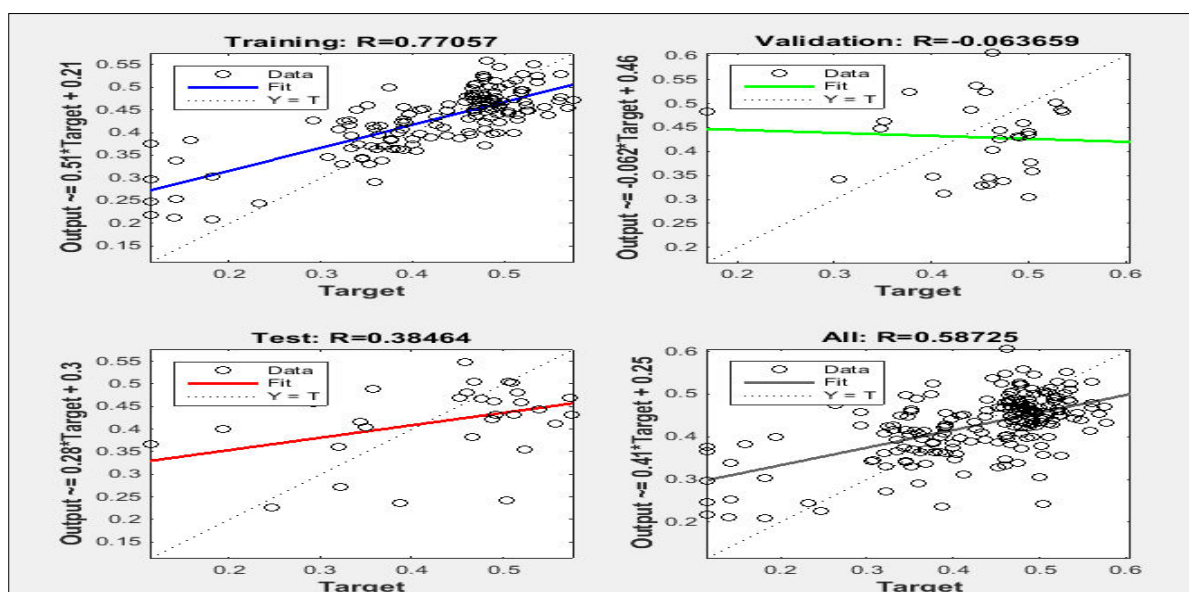


Figure 5.7(a): Snapshot of SCA1 protein Ataxin-1 against disordered proteins.



. Figure 5.7(b): Snapshot of SCA1 protein Ataxin-1 against ordered proteins.

### 5) Ataxin-2– SCA2

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Ataxin-2's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.95726$  and the validation R value was  $R=0.93992$ . While from the plot against ordered proteins it can be seen that the data fits at the values higher than 0.4 and the resultant R value was  $R=0.81716$  and validation R value was  $R=0.68285$ .

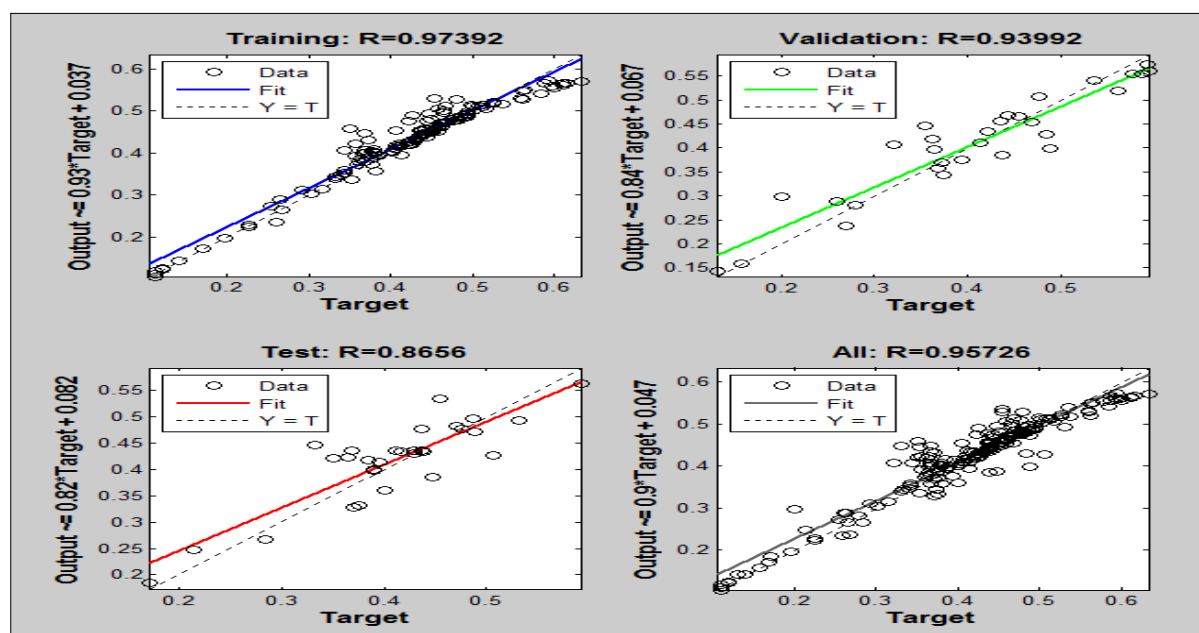


Figure 5.8(a): Snapshot of SCA2 protein Ataxin-2 against disordered proteins.

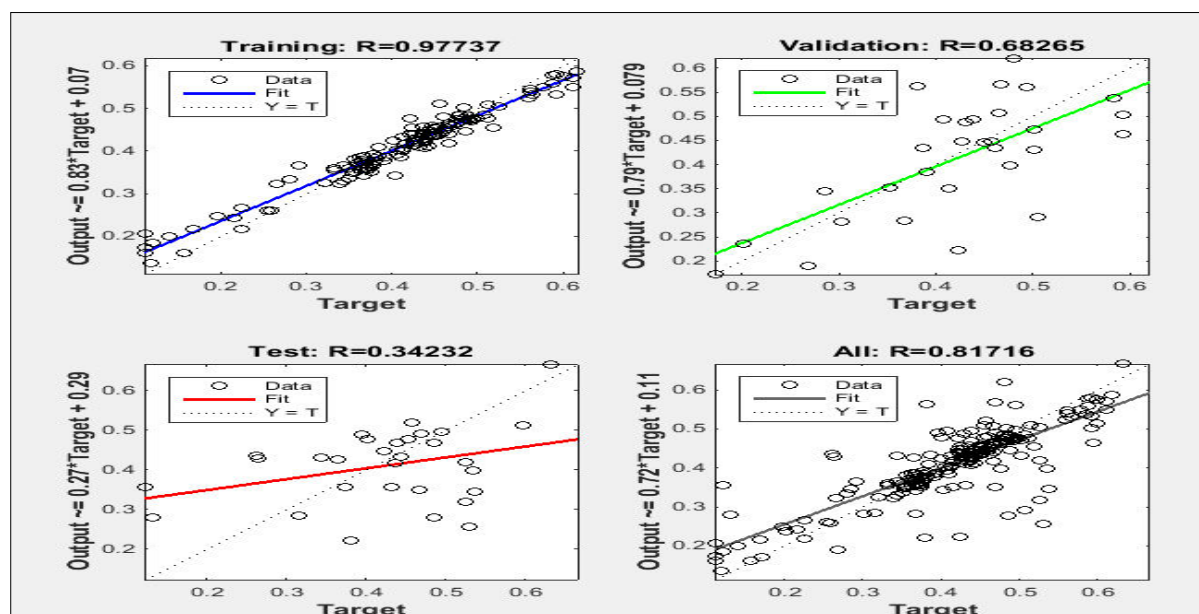


Figure 5.8(b): Snapshot of SCA2 protein Ataxin-2 against ordered proteins.

## 6) Ataxin-3– SCA3

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Ataxin-3's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.9651$  and the validation R value was  $R=0.93297$ . While from the plot against ordered proteins it can be seen that the data fits at the values between 0.4 and 0.5 and the resultant R value was  $R=0.80952$  and validation R value was  $R=0.64659$ .

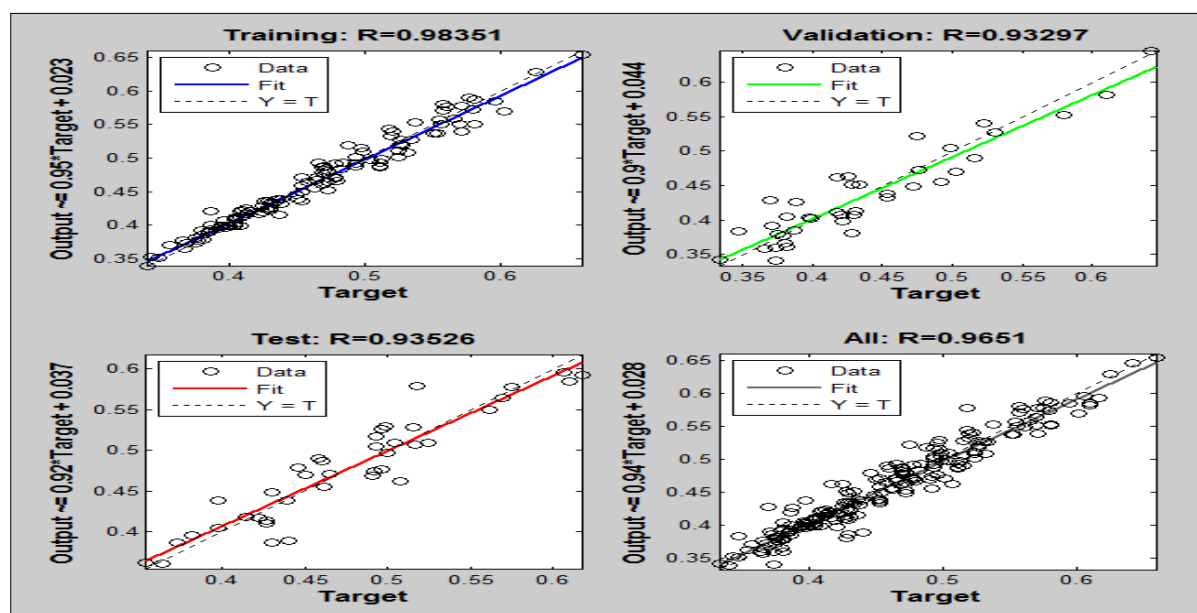


Figure 5.9(a): Snapshot of SCA3 protein Ataxin-3 against disordered proteins.

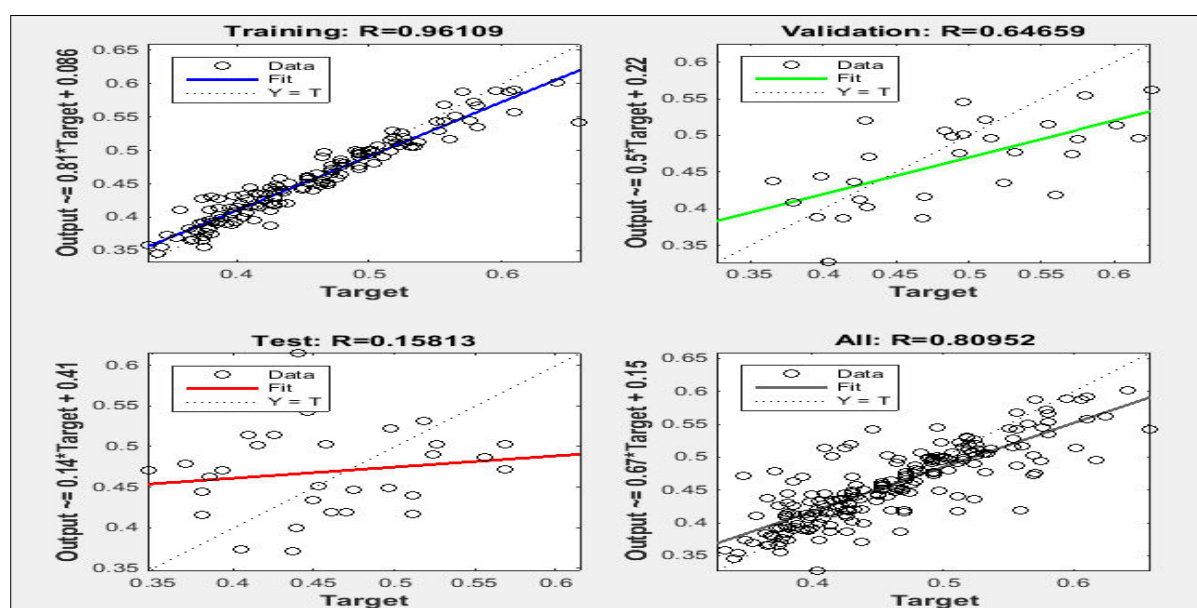


Figure 5.9(b): Snapshot of SCA3 protein Ataxin-3 against ordered proteins.

### 7) $\alpha 1A$ – SCA6

From the plot against disordered proteins it was concluded that the fit line fits maximum of the  $\alpha 1A$ 's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.94742$  and the validation R value was  $R=0.89579$ . While from the plot against ordered proteins it can be seen that the resultant R value was  $R=0.66792$  and validation R value was  $R=0.13754$ .

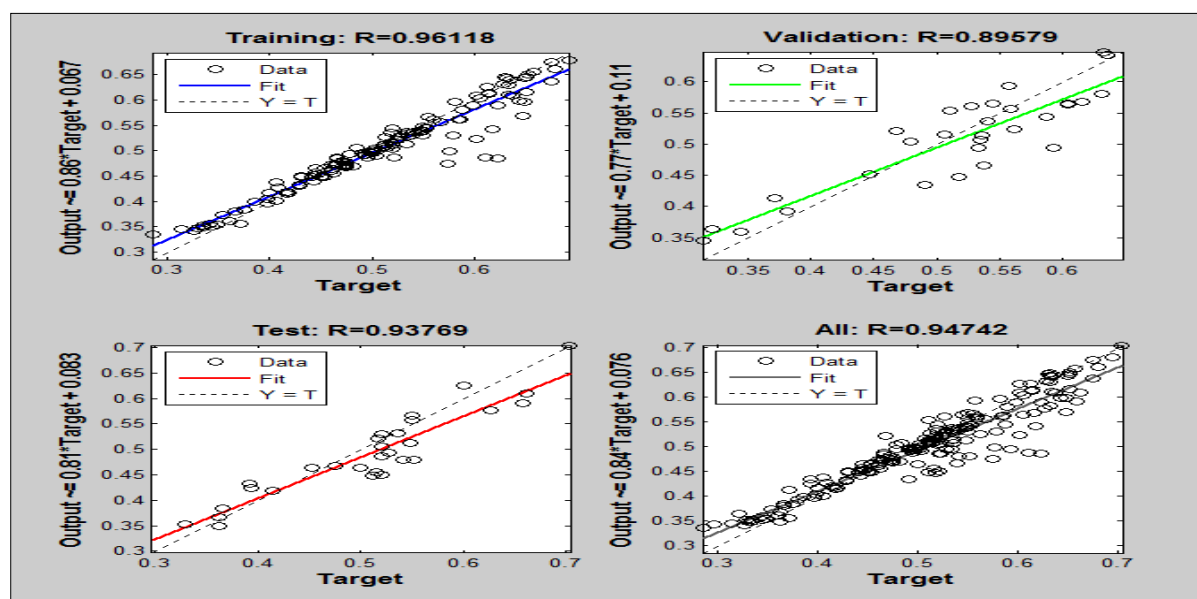


Figure 5.10(a): Snapshot of SCA6 protein  $\alpha 1A$  against disordered proteins.

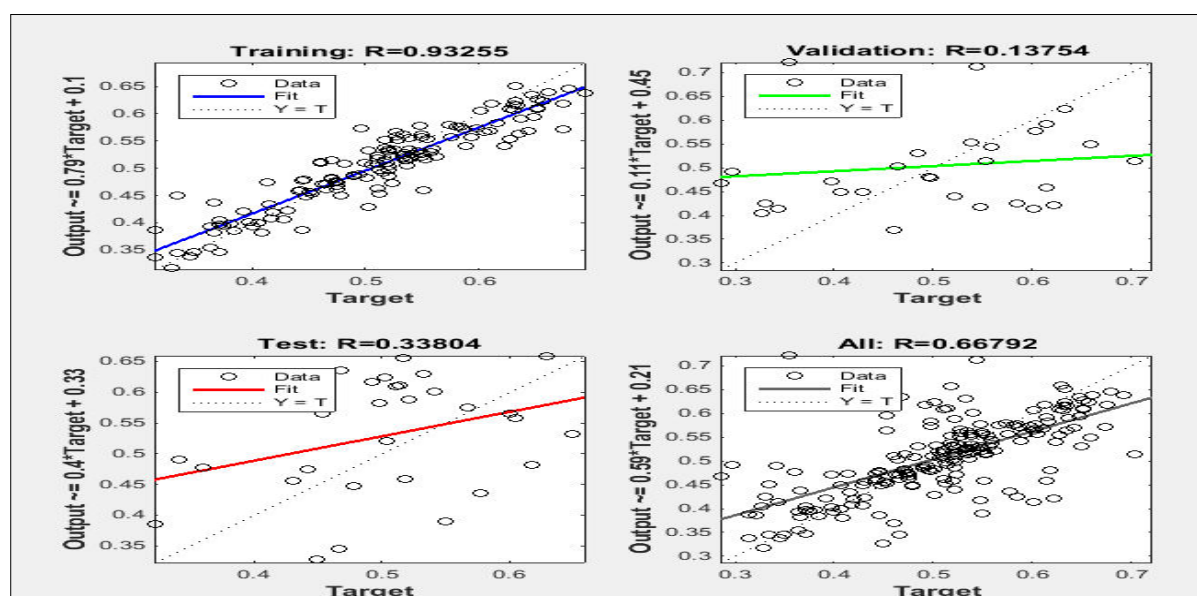


Figure 5.10(b): Snapshot of SCA6 protein  $\alpha 1A$  against ordered proteins.



### 8) Ataxin-7 – SCA7

From the plot against disordered proteins it was concluded that the fit line fits maximum of the Ataxin-7's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.97699$  and the validation R value was  $R=0.9540$ . While from the plot against ordered proteins it can be seen that the data fits between 0.4 and 0.5 and the resultant R value was  $R=0.82961$  and but the validation R value was  $R=0.55342$ .

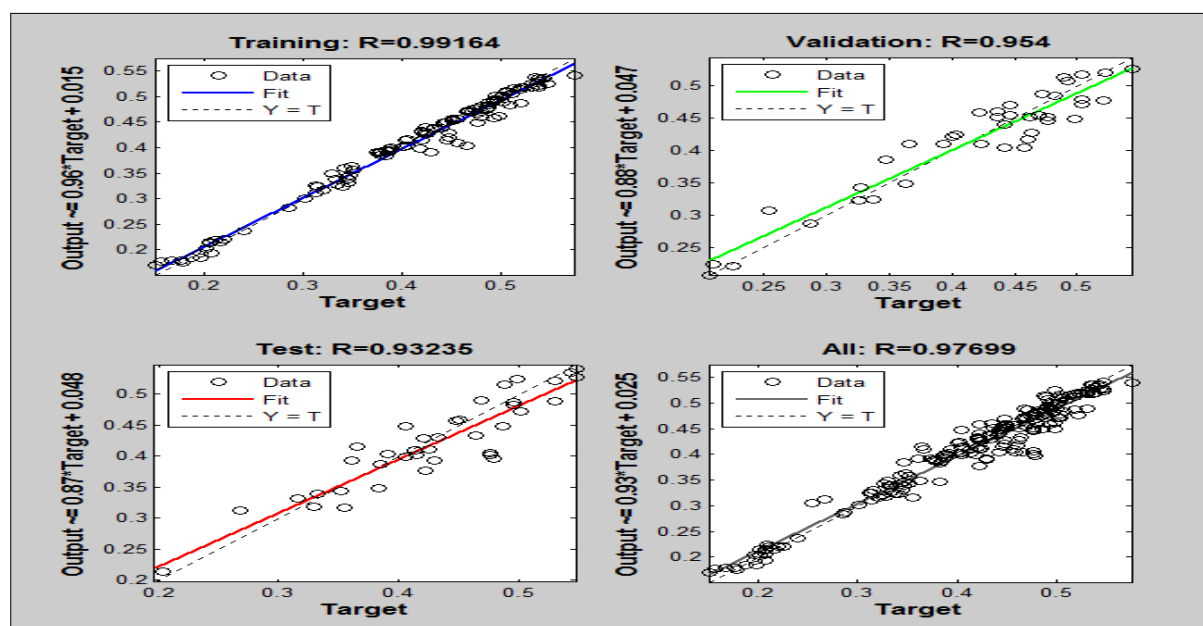


Figure 5.11(a): Snapshot of SCA7 protein Ataxin-7 against disordered proteins.

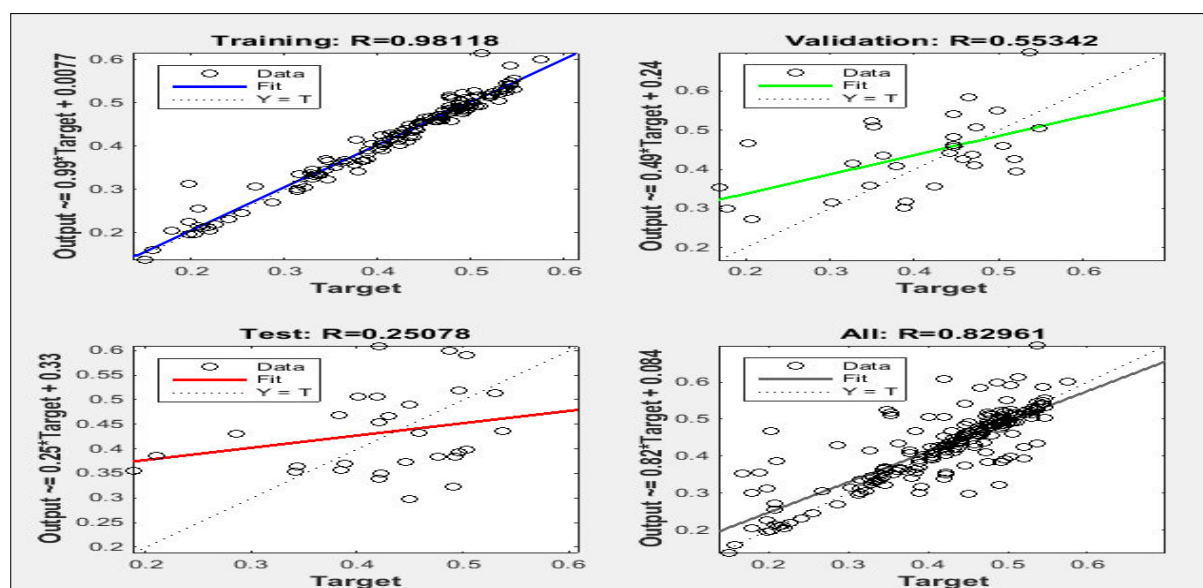


Figure 5.11(b): Snapshot of SCA7 protein Ataxin-7 against ordered proteins.

### 9) TBP – SCA17

From the plot against disordered proteins it was concluded that the fit line fits maximum of the TBP's hydrophobicity scores and the resultant R value obtained is close to 1, i.e.  $R=0.99112$  and the validation R value was  $R=0.96832$ . While from the plot against ordered proteins it can be seen that the resultant R value was  $R=0.66242$  and but the validation R value was  $R=0.42181$ .

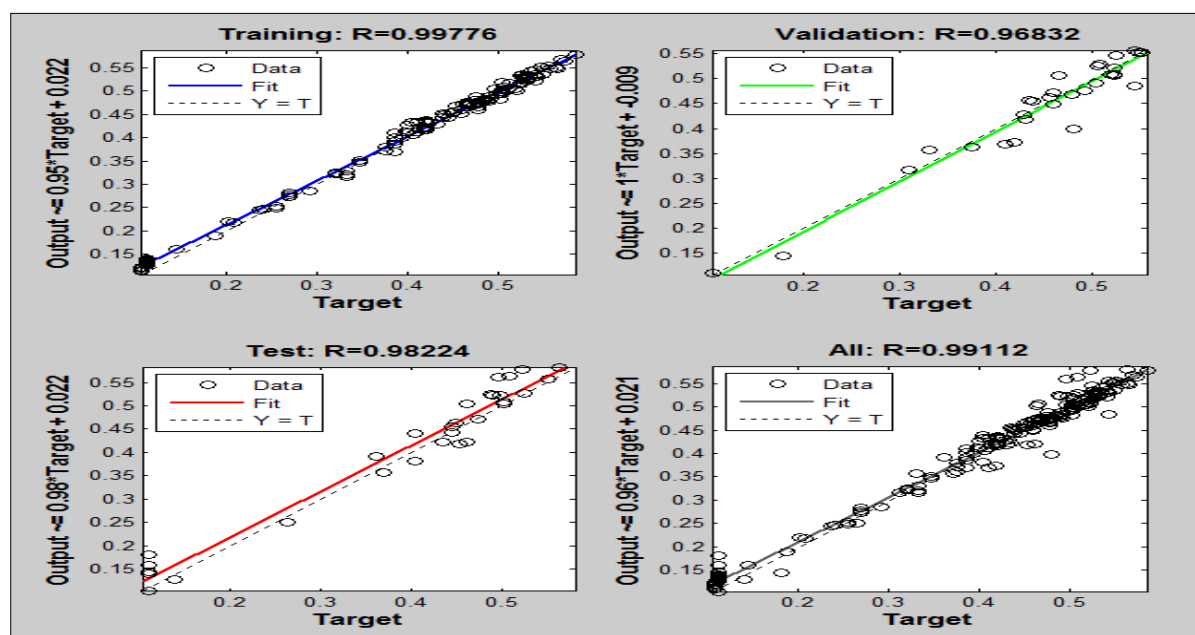


Figure 5.12(a): Snapshot of SCA17 protein TBP against disordered proteins.

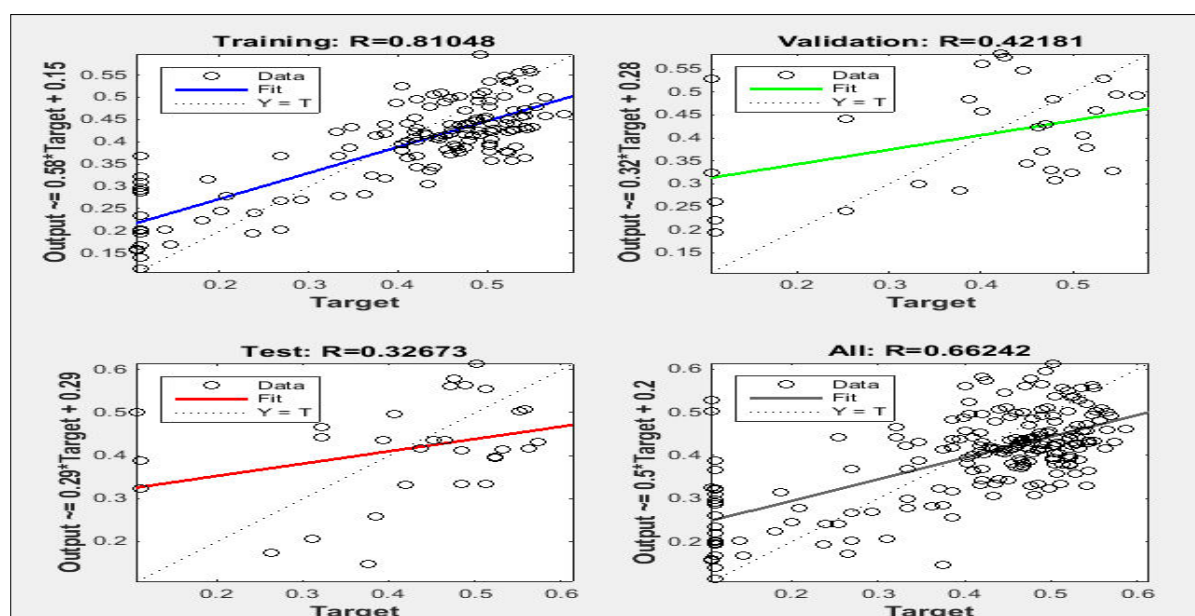


Figure 5.12(b): Snapshot of SCA17 protein TBP against ordered proteins.

### 5.3 Discussion

The hydrophobicity in a protein is important for its folding and so for order promotion and stability the protein should have high hydrophobic scores for its sequences, i.e. the number of hydrophobic residues should be greater. Likewise for disorder promotion the protein should possess lower hydrophobic scores for its amino acid sequence.

From the resultant plots and the R values above it can be easily seen that validation R values are closer to 1 when the input dataset was of disordered proteins, while, the validation R values were significantly less when input dataset was of ordered proteins. Hence, it can be predicted that all the 9 proteins responsible for the CAG repeat diseases have disordered regions in their sequences.

Previously, 7 out of 9 proteins i.e. atrophin-1, AR, ataxin-2, ataxin-3,  $\alpha$ 1A, ataxin-7 and TBP, responsible for the pathogenesis of CAG repeat diseases have been predicted to be either partially or completely disordered [16].

**Table 5.1: R values of resultant and validation regression plots of nine pathogenic proteins against 100 ordered and 100 disordered proteins datasets.**

Pathogenic Protein	Resultant R value		Validation R value	
	Disordered	Ordered	Disordered	Ordered
<b>Atrophin-1</b>	0.97607	0.32315	0.94603	0.23049
<b>Huntingtin</b>	0.95287	0.74341	0.90894	0.23158
<b>Androgen Receptor</b>	0.98677	0.81914	0.95716	0.48440
<b>Ataxin-1</b>	0.97977	0.58725	0.91154	-0.06360
<b>Ataxin-2</b>	0.95726	0.81716	0.93992	0.68285
<b>Ataxin-3</b>	0.95510	0.80952	0.93297	0.64659
<b><math>\alpha</math>1A</b>	0.94742	0.66792	0.89579	0.13754
<b>Ataxin-7</b>	0.97699	0.82961	0.95400	0.55342
<b>TBP</b>	0.99112	0.66242	0.96382	0.42181

# **Chapter 6**

## **Conclusion and Future Scope**

## 6.1 Conclusion

In this thesis, the lack of structure, i.e. disorder in the protein structure in the protein structure was studied. The peculiarities in the amino acid sequences and their properties were studied which are required for the prediction of such disordered regions in a protein. Also some of the commonly used machine languages to study disorder in the protein sequence have also been reviewed.

Introduction to the problem of pathogenesis of polyglutamine disease due to responsible proteins was done in this study and the disorderiness in their sequences was predicted. For that the hydrophobicity scores of ordered and disordered proteins were collected and trained to the learning neural network and were tested against the target proteins responsible for polyglutamine diseases separately.

From the regression plot obtained it was analysed that all the nine proteins responsible for the pathogenesis of polyglutamine diseases had disordered regions in them. This prediction was made on the basis of the resultant and validation R values of the plots obtained which were found to be equal to 1.

## 6.2 Related Work

Apart from polyglutamine diseases, disorderiness have been found in the proteins responsible for other diseases also. Hence, due to their structural range, they serve as the potential drug targets.

Also, many others machine learning techniques are coming in picture for predicting and studying such regions which have greater accuracies and efficiencies.

## 6.3 Future Scope

In the introduction it was proposed to be predicted that the proteins related to the CAG diseases have regions of disorder through neural network training and regression plot analysis. This can also be achieved by various other machine learning algorithms, namely the parameters based clustering.

Also not only the proteins of CAG repeat diseases, proteins responsible for other diseases can also be studied extensively for disorder prediction. Apart from disease proteins, other

functionally significant proteins can also be nailed for disorder prediction. In past research done numerous IDPs have been found to have close relationships with human diseases such as Alzheimer disease, Parkinson disease, tumor, diabetes, and so on. It has also been found that most of the disease connected IDPs play primary roles in the disease-linked protein-protein interactions. Hence, the disease-linked IDPs may prove to be the potential targets for drugs modulating protein-protein interactions. Because of this reason novel strategies for drug discovery based on IDPs are in domination.

In order to promote novel strategies for drug discovery, it is essential that more and more features of IDPs are revealed by computing methods like their parameter based neural network training, using which various parameters responsible for disorder can be studied.

# References

- [1] Tompa, Peter. "Intrinsically unstructured proteins." *Trends in biochemical sciences* 27, no. 10 (2002): 527-533..
- [2] Dedmon, Matthew M., Chetan N. Patel, Gregory B. Young, and Gary J. Pielak. "FlgM gains structure in living cells." *Proceedings of the National Academy of Sciences* 99, no. 20 (2002): 12681-12684.
- [3] Radivojac, Predrag, Lilia M. Iakoucheva, Christopher J. Oldfield, Zoran Obradovic, Vladimir N. Uversky, and A. Keith Dunker. "Intrinsic disorder and functional proteomics." *Biophysical journal* 92, no. 5 (2007): 1439-1456.
- [4] Daughdrill, Gary W., Stepan Kashtanov, Amber Stancik, Shannon E. Hill, Gregory Helms, Martin Muschol, Véronique Receveur-Bréchet, and F. Marty Ytreberg. "Understanding the structural ensembles of a highly extended disordered protein." *Molecular BioSystems* 8, no. 1 (2012): 308-319.
- [5] Babu, M. Madan, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. "Intrinsically disordered proteins: regulation and disease." *Current opinion in structural biology* 21, no. 3 (2011): 432-440.
- [6] Gunasekaran, Kannan, Chung-Jung Tsai, Sandeep Kumar, David Zanuy, and Ruth Nussinov. "Extended disordered proteins: targeting function with less scaffold." *Trends in biochemical sciences* 28, no. 2 (2003): 81-85.
- [7] Brown, Celeste J., Sachiko Takayama, Andrew M. Campen, Pam Vise, Thomas W. Marshall, Christopher J. Oldfield, Christopher J. Williams, and A. Keith Dunker. "Evolutionary rate heterogeneity in proteins with long disordered regions." *Journal of molecular evolution* 55, no. 1 (2002): 104-110.
- [8] Radivojac P, Obradovic Z, Brown CJ, Dunker AK. "Improving sequence alignments for intrinsically disordered proteins." *Proc. 7th Pacific Symposium on Biocomputing* (2002): 589-600.



- [9] Sickmeier, Megan, Justin A. Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S. Cortese, Agnes Tantos, Beata Szabo et al. "DisProt: the database of disordered proteins." *Nucleic acids research* 35, no. suppl 1 (2007): D786-D793.
- [10] Brown, Terence A. *Genomes*. Garland science, 2006.
- [11] Li, Ling-Bo, Zhenming Yu, Xiuyin Teng, and Nancy M. Bonini. "RNA toxicity is a component of ataxin-3 degeneration in *Drosophila*." *Nature* 453, no. 7198 (2008): 1107-1111.
- [12] Miller, Jill W., Carl R. Urbinati, Patana Teng-umnuay, Myrna G. Stenberg, Barry J. Byrne, Charles A. Thornton, and Maurice S. Swanson. "Recruitment of human muscleblind proteins to (CUG) n expansions associated with myotonic dystrophy." *The EMBO journal* 19, no. 17 (2000): 4439-4448.
- [13] Ho, Thai H., Rajesh S. Savkur, Michael G. Poulos, Michael A. Mancini, Maurice S. Swanson, and Thomas A. Cooper. "Colocalization of muscleblind with RNA foci is separable from mis-regulation of alternative splicing in myotonic dystrophy." *Journal of cell science* 118, no. 13 (2005): 2923-2933.
- [14] Kino, Yoshihiro, Daisuke Mori, Yoko Oma, Yuya Takeshita, Noboru Sasagawa, and Shoichi Ishiura. "Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats." *Human molecular genetics* 13, no. 5 (2004): 495-507.
- [15] Vacic, Vladimir, Vladimir N. Uversky, A. Keith Dunker, and Stefano Lonardi. "Composition Profiler: a tool for discovery and visualization of amino acid composition differences." *BMC bioinformatics* 8, no. 1 (2007): 211.
- [16] Uversky, Vladimir N. "Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the D2 concept." *Expert review of proteomics* 7, no. 4 (2010): 543-564.

- [17] Mathura, V., W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown, and A. K. Dunker. "The protein non-folding problem: amino acid determinants of intrinsic order and disorder." In Pacific Symposium on Biocomputing, vol. 6, pp. 89-100. 2001.
- [18] Lavery, Derek N., and Iain J. McEwan. "Structural Characterization of the Native NH2-Terminal Transactivation Domain of the Human Androgen Receptor: A Collapsed Disordered Conformation Underlies Structural Plasticity and Protein-Induced Folding†." *Biochemistry* 47, no. 11 (2008): 3360-3369.
- [19] Yazawa, Ikuru, Nobuyuki Nukina, Hideji Hashida, Jun Goto, Masao Yamada, and Ichiro Kanazawa. "Abnormal gene product identified in hereditary dentatorubral–pallidoluysian atrophy (DRPLA) brain." *Nature genetics* 10, no. 1 (1995): 99-103.
- [20] Albrecht, Mario, Michael Golatta, Ullrich Wüllner, and Thomas Lengauer. "Structural and functional analysis of ataxin-2 and ataxin-3." *European Journal of Biochemistry* 271, no. 15 (2004): 3155-3170.
- [21] Masino, Laura, Valeria Musi, Rajesh P. Menon, Paola Fusi, Geoff Kelly, Thomas A. Frenkiel, Yvon Trottier, and Annalisa Pastore. "Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail." *FEBS letters* 549, no. 1 (2003): 21-25.
- [22] Palhan, Vikas B., Shiming Chen, Guang-Hua Peng, Agneta Tjernberg, Armin M. Gamper, Yuxin Fan, Brian T. Chait, Albert R. La Spada, and Robert G. Roeder. "Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 24 (2005): 8472-8477.
- [23] Friedman, Meyer J., Chuan-En Wang, Xiao-Jiang Li, and Shihua Li. "Polyglutamine expansion reduces the association of TATA-binding protein with DNA and induces DNA binding-independent neurotoxicity." *Journal of Biological Chemistry* 283, no. 13 (2008): 8283-8290.

- [24] Dunker, A. Keith, J. David Lawson, Celeste J. Brown, Ryan M. Williams, Pedro Romero, Jeong S. Oh, Christopher J. Oldfield et al. "Intrinsically disordered protein." *Journal of Molecular Graphics and Modelling* 19, no. 1 (2001): 26-59.
- [25] Romero, Pedro, Zoran Obradovic, Xiaohong Li, Ethan C. Garner, Celeste J. Brown, and A. Keith Dunker. "Sequence complexity of disordered protein." *Proteins: Structure, Function, and Bioinformatics* 42, no. 1 (2001): 38-48.
- [26] Li, Xiaohong, Celeste J. Brown, Zoran Obradovic, Ethan C. Garner, and A. Keith Dunker. "Comparing predictors of disordered protein." *Genome Informatics* 11 (2000): 172-184.
- [27] Iakoucheva, Lilia M., Celeste J. Brown, J. David Lawson, Zoran Obradović, and A. Keith Dunker. "Intrinsic disorder in cell-signaling and cancer-associated proteins." *Journal of molecular biology* 323, no. 3 (2002): 573-584.
- [28] Uversky, Vladimir N., Christopher J. Oldfield, and A. Keith Dunker. "Intrinsically disordered proteins in human diseases: introducing the D2 concept." *Annu. Rev. Biophys.* 37 (2008): 215-246.
- [29] Uversky, Vladimir N., Ann Roman, Christopher J. Oldfield, and A. Keith Dunker. "Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs." *Journal of proteome research* 5, no. 8 (2006): 1829-1842.
- [30] Cheng, Yugong, Tanguy LeGall, Christopher J. Oldfield, A. Keith Dunker, and Vladimir N. Uversky. "Abundance of intrinsic disorder in protein associated with cardiovascular disease." *Biochemistry* 45, no. 35 (2006): 10448-10460.
- [31] Midic, Uros, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. "Protein disorder in the human diseasome: unfoldomics of human genetic diseases." *Bmc Genomics* 10, no. Suppl 1 (2009): S12.

- [32] Vucetic, Slobodan, Hongbo Xie, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions." *Journal of proteome research* 6, no. 5 (2007): 1899-1916.
- [33] Xie, Hongbo, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins." *Journal of proteome research* 6, no. 5 (2007): 1917-1932.
- [34] Xie, Hongbo, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Vladimir N. Uversky, and Zoran Obradovic. "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions." *Journal of proteome research* 6, no. 5 (2007): 1882-1898.
- [35] Uversky, Vladimir N., Christopher J. Oldfield, Uros Midic, Hongbo Xie, Bin Xue, Slobodan Vucetic, Lilia M. Iakoucheva, Zoran Obradovic, and A. Keith Dunker. "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases." *BMC genomics* 10, no. Suppl 1 (2009): S7.
- [36] Uversky, Vladimir N. "Intrinsic disorder in proteins associated with neurodegenerative diseases." In *Protein folding and misfolding: neurodegenerative diseases*, pp. 21-75. Springer Netherlands, 2009.
- [37] Ishida, Takashi, and Kengo Kinoshita. "Prediction of disordered regions in proteins based on the meta approach." *Bioinformatics* 24, no. 11 (2008): 1344-1348.
- [38] He, Bo, Kejun Wang, Yunlong Liu, Bin Xue, Vladimir N. Uversky, and A. Keith Dunker. "Predicting intrinsic disorder in proteins: an overview." *Cell research* 19, no. 8 (2009): 929-949.

- [39] Ferron, François, Sonia Longhi, Bruno Canard, and David Karlin. "A practical overview of protein disorder prediction methods." *Proteins: Structure, Function, and Bioinformatics* 65, no. 1 (2006): 1-14.
- [40] Dosztányi, Zsuzsanna, Márk Sándor, Peter Tompa, and István Simon. "Prediction of protein disorder at the domain level." *Current Protein and Peptide Science* 8, no. 2 (2007): 161-171.
- [41] Schlessinger, Avner, Marco Punta, Guy Yachdav, Laszlo Kajan, and Burkhard Rost. "Improved disorder prediction by combination of orthogonal approaches." *PLoS One* 4, no. 2 (2009): e4433.
- [42] Xue, Bin, Christopher J. Oldfield, A. Keith Dunker, and Vladimir N. Uversky. "CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions." *FEBS letters* 583, no. 9 (2009): 1469-1474.
- [43] Romero, Pedro, Z. O. R. A. N. Obradovic, Charles R. Kissinger, J. Ernest Villafranca, E. T. H. A. N. Garner, S. T. E. P. H. E. N. Guilliot, and A. KEITH Dunker. "Thousands of proteins likely to have long disordered regions." In *Pac Symp Biocomput*, vol. 3, pp. 437-448. 1998.
- [44] Li, Xiaohong, Pedro Romero, Meeta Rani, A. Keith Dunker, and Zoran Obradovic. "Predicting protein disorder for N-, C-and internal regions." *Genome Informatics* 10 (1999): 30-40.
- [45] Peng, Kang, Predrag Radivojac, Slobodan Vucetic, A. Keith Dunker, and Zoran Obradovic. "Length-dependent prediction of protein intrinsic disorder." *BMC bioinformatics* 7, no. 1 (2006): 208.
- [46] Ward, Jonathan J., Liam J. McGuffin, Kevin Bryson, Bernard F. Buxton, and David T. Jones. "The DISOPRED server for the prediction of protein disorder." *Bioinformatics* 20, no. 13 (2004): 2138-2139.

- [47] Jin, Yumi, and Roland L. Dunbrack. "Assessment of disorder predictions in CASP6." *Proteins: Structure, Function, and Bioinformatics* 61, no. S7 (2005): 167-175.
- [48] Bordoli, Lorenza, Florian Kiefer, and Torsten Schwede. "Assessment of disorder predictions in CASP7." *Proteins: Structure, Function, and Bioinformatics* 69, no. S8 (2007): 129-136.
- [49] Tompa, Peter. "Intrinsically disordered proteins: a 10-year recap." *Trends in biochemical sciences* 37, no. 12 (2012): 509-516.
- [50] Xue, Bin, Roland L. Dunbrack, Robert W. Williams, A. Keith Dunker, and Vladimir N. Uversky. "PONDR-FIT: a meta-predictor of intrinsically disordered amino acids." *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1804, no. 4 (2010): 996-1010.
- [51] Müller-Späth, Sonja, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rügger, Luc Reymond, Daniel Nettels, and Benjamin Schuler. "Charge interactions can dominate the dimensions of intrinsically disordered proteins." *Proceedings of the National Academy of Sciences* 107, no. 33 (2010): 14609-14614.
- [52] Eliezer, David. "Biophysical characterization of intrinsically disordered proteins." *Current opinion in structural biology* 19, no. 1 (2009): 23-30.
- [53] Linding, Rune, Joost Schymkowitz, Frederic Rousseau, Francesca Diella, and Luis Serrano. "A comparative study of the relationship between protein structure and  $\beta$ -aggregation in globular and intrinsically disordered proteins." *Journal of molecular biology* 342, no. 1 (2004): 345-353.
- [54] Yang, Zheng Rong, Rebecca Thomson, Philip McNeil, and Robert M. Esnouf. "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins." *Bioinformatics* 21, no. 16 (2005): 3369-3376.

- [55] Xie, Qian, Zoran Obradovic, Gregory E. Arnold, Ethan Garner, Pedro Romero, and A. Keith Dunker. "The sequence attribute method for determining relationships between sequence and protein disorder." *genome Informatics* 9 (1998): 193-200.
- [56] Hobohm, Uwe, Michael Scharf, Reinhard Schneider, and Chris Sander. "Selection of representative protein data sets." *Protein Science* 1, no. 3 (1992): 409-417.
- [57] Hobohm, Uwe, and Chris Sander. "Enlarged representative set of protein structures." *Protein Science* 3, no. 3 (1994): 522-524.
- [58] Griep, Sven, and Uwe Hobohm. "PDBselect 1992–2009 and PDBfilter-select." *Nucleic acids research* (2009): gkp786.
- [59] Guide, MATLAB User's. "The mathworks." Inc., Natick, MA 5 (1998): 333.
- [60] Kyte, Jack, and Russell F. Doolittle. "A simple method for displaying the hydropathic character of a protein." *Journal of molecular biology* 157, no. 1 (1982): 105-132.

# Plagiarism Detector - Originality Report

Plagiarism Detector copy registered to:

Akhtar

Rasool\_License2

Software core version: 850

Originality

report details:

**Generation Time**

7/8/2015 1:35:38

PM and Date:

Document

Name: ReportAina1.docx

Document

Location: C:\Users\hcl\Desktop\ReportAina1.docx

Document Words **9125**

Count:

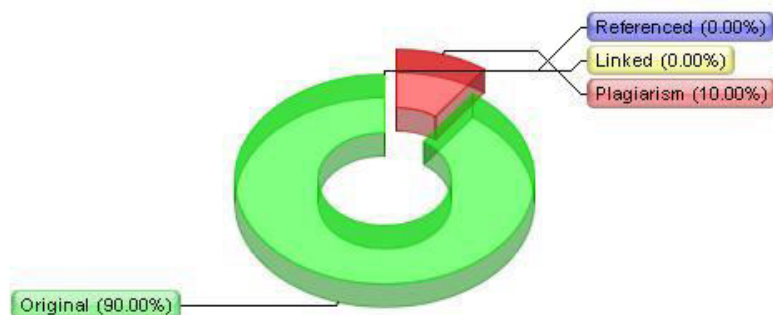
Check time

00:06:18

[hs:ms:ss]:

Important Hint: to understand what exactly is meant by any report value - you can click [?](#). It will navigate you to the most detailed explanation at our web site.

[?](#) Plagiarism Detection Chart:



Referenced 0% / Linked 0%  
Original - 90% / 10% - Plagiarism

[?](#) Processed Resources List: [click below to open] Processed Ok: 98 Failed: 46

[\[Toggle other sources:\]](#)