Genetics of Cardiovascular Disorders: An *in-silico* Analysis using Protein Structural and Interaction Information

Prerna Jain

Delhi Technological University, Delhi, India

1. ABSTRACT

Proteins are the structural and functional workhouse in the cell that takes part in virtually every event within and between cells. Proteins in association with other molecules determine the ultimate behavior of biological system. Recently, network-centered approaches have been increasingly used to comprehend the fundamentals of biology. There are different databases documenting the interactions of proteins as PPI networks but they do not reveal the molecular mechanism behind the binding process occurring between molecules. This problem can only be solved by including the structural details of the complexes which includes the 3 dimensional structures of proteins, interface as well as topological properties. Interface is the region where two protein chains interact leading to formation of protein complex. The main concern of the present study is to present interface analysis of cardiovascular-disorder (CVD) related proteins to shed lights on details of interactions and to emphasize the importance of using structures in network studies. We have used interface properties as parameters to classify the CVD associated proteins and non CVD proteins. Machine learning algorithm was used to generate a classifier based on the training set which was used to predict potential CVD related proteins from a set of polymorphic proteins which are not known to be involved in any disease. The predicted CVD related proteins may not be the causing factor of particular disease but can be involved in pathways and reactions yet unknown to us thus permitting a more rational analysis of disease mechanism. Study of their interactions with other proteins can significantly improve our understanding in the molecular mechanism of diseases. The wider scope of this study is the characterization of all the hereditary disorders based on their structural properties to gain better understanding of the molecular machinery within the cells of living organism.

2. INTRODUCTION

The association between genes and diseases has been well studied in the past and been documented in various databases such as Online Mendelian Inheritance in Man, Genetic Association Database, Disease ontology and many. These databases not only provide genes associated with the disorders but also, insight about the common genes or essential genes which are involved in pathways and process of multiple disorders. The human disease gene network has already established link between the genetic disorders with the genes. An important conclusion inferred from this network is that genes associated with similar disorders show higher likelihood of physical interaction between their products, thus forming a hub of essential genes and their products (Goh *et al.,* 2007).

Documentation of genes with the associated disease is not enough to understand the biological details of pathogenesis and disease progression. It is important to identify various molecules and mechanism triggering, participating and controlling the biological process. This understanding of molecular mechanism is a complex process and not much is known about the mechanism (Gonzalez *et al.,* 2012). Now with the tremendous increase in human interaction data, it is important to understand how any biological function is exerted over the body and this can only be made possible by the inclusion of structural details in the networks.

Undoubtedly, sequence based annotation is important in unraveling the encoded information but the finer details of the molecular mechanism within the cell is possible only with the structural information (Marini *et al.,* 2010). Protein structures provide a higher resolution of information and a more sensitive approach for detecting similarities among proteins by including details about structures of the interacting proteins in the network, protein hubs and protein interfaces. Structural profile of the proteins provides the opportunities in understanding the cellular functioning in terms of structural scaffolds which facilitate the underlying molecular recognition events. Moreover, a protein’s structure is better conserved than its amino acid structure (Choura *et al.,* 2011).

Protein-protein interactions (PPI) network is a way of representing how two or more proteins interact with each other in a cell and biological processes are essentially interactions between multiple proteins (Zhang *et al.*, 2011) with PPI networks controlling the flow of information both within and between the cells. Protein interactions are mediated by specific recognition of distinct binding regions on the surface of interacting protein. Such recognition should be of sufficient affinity to effectively bind fragments of proteins, which is decided by specificities such as interfaces, thus interface properties and the topology of the proteins involved is very important in deciphering their function in the network.

Various studies have been carried to integrate protein-protein interaction network with the structural properties. Major work in this area has been done in the field of cancer biology.

Kar *et al.,* 2009; mapped the cancer genes onto human PPI network and studied the network with respect to structural features such as interface properties and they were successful in classifying cancer and non-cancer proteins. Zhang *et al.,* 2011; proposed a novel integrated approach named CAERUS, for identification of gene signatures to predict cancer outcomes based on domain interaction network in human proteome. Taylor *et al.,* 2009; proposed a new methodology to predict breast cancer outcome based on the correlation of gene expression profile between hub proteins and their interacting partners in PPI network. Chuang *et al.,* 2007; developed a method to find sub network – based signatures by incorporating PPI networks and gene expression profiles to classify metastatic and non-metastatic tumors.

Studies have shown that a molecular mechanism leading to diseases is dependent on interlinked proteins and networks, but we are far from unraveling the difference in molecular interactions in healthy and diseased organism. Mechanism of genetic malfunctioning, that leads to one or several diseases can be understood when the molecular level of the protein interactions are known. Understanding the interactions and hence binding between proteins is critical for the rational design of new therapeutic agents targeted to disrupt the interactions that cause disorders (Gonzalez *et al.,* 2012).

PPIs not only identify disease-associated interacting proteins but also the potentially interesting disease-associated gene candidates (genes coding for interacting proteins are putative disease causing genes). This theory has been used in the present work where all the protein associated with cardiovascular disease have been retrieved and a classifier using machine learning algorithms is used to identify potential proteins which are associated with cardiovascular diseases. This classifier is based on the interfacial properties of the interacting complexes, to differentiate cardiovascular disease related proteins from non-cardiovascular disease associated proteins. The wider scope of this study is the characterization of all the hereditary disorders based on their structural properties to gain better understanding in the molecular mechanism (such as effect of mutation on protein structure, which residue is being affected, affect of all the residues including charged, polar, uncharged, how hydrogen bonding between complexes are important and so on) behind these diseases.

3. REVIEW OF LITERATURE

The central dogma of molecular biology states that DNA is transcribed into RNA, and RNA is translated into protein. Protein, thus formed is in the form of polypeptide chain which folds into a three dimensional structure that requires minimum energy (Figure 1). The amino acid residue composition of protein uniquely determines its 3D structure. Proteins are the functional and structural workhouses in the cells of living organisms as they are involved in transport, storage, catalysis, signaling and many more functions inside the body. The function of protein can be determined by analysis of either amino acid sequence or the 3D structure. Previous studies have shown that sequence based analysis is less accurate and less sensitive than structure based because, the later is better conserved during the evolution, for example there are large number of proteins whose structure (hence function) are same but sequences are quite different. Thus the function of protein is more closely associated to its three dimensional structure that to amino acid compositions (Aung, Z. 2006).

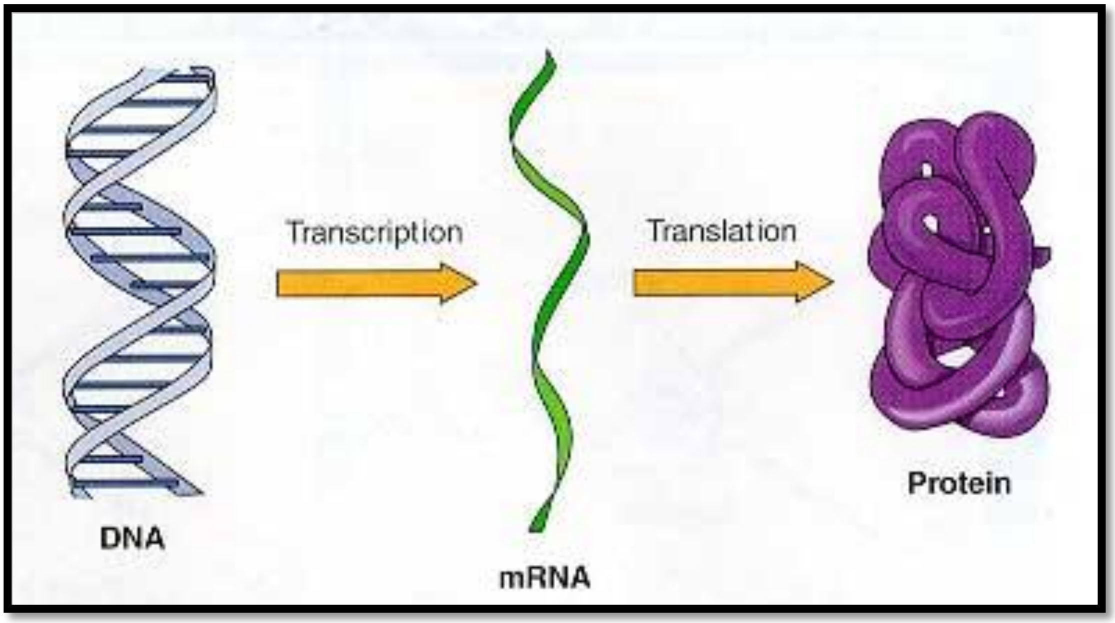


Figure 1: Central dogma of gene expression

Mutations in genes can change the sequence and structure of the encoded protein, affecting its binding sites and ultimately impairing the original function of protein and can possibly lead to disease (Steward *et al.*, 2003). Most of the mutation in diseased genes map to non coding regions but the mutations that occur in the coding region of genes, severely affect protein function by affecting its structure due to deletions, insertion, inversions and other abnormalities. Further, mutations at highly conserved residues are more likely to cause disruption in the protein structure as these conserved residues are buried deep inside the protein structures. However majority of the mutations occur at solvent accessible sites where they generally affect solubility and interaction patterns of protein with other proteins, ligands and nucleic acids (Venselaar *et al.,* 2010).

For example in Porphyria cutanea tarda (PCT), caused by an accumulation of uroporphyrins in the liver and plasma is controlled by gene the uroporphyrinogen decarboxylase (URO-D). URO-D is a α/β barrel – a series of alternating α-helical and β-strand. Mutation of Leu195Phe substitutes a hydrophobic leucine residue buried deep within the protein, for a larger aromatic phenylalanine residue, causing the surrounding side-chains to rearrange which results in only 30% of the mean activity of the normal allele, leading to an accumulation of substrate and, hence, the disease. Similarly mutations leads to variety of changes in protein structure such as disruption of protein-protein interaction, disruption of hydrogen bonding network, interference with DNA binding, breaking of disulfide bonds, mutation in catalytic site and many more leading to severe disease (Steward *et al.,* 2003). By analyzing the 3D structure of the protein, more detailed information on the role of mutated residue can be inferred (Figure 2); moreover structural analysis can often explain why different phenotypes originate from mutations in the same gene. Thus, the major aim is to study the interaction sites which can be hampered due to these mutations thus affecting the normal function of proteins.

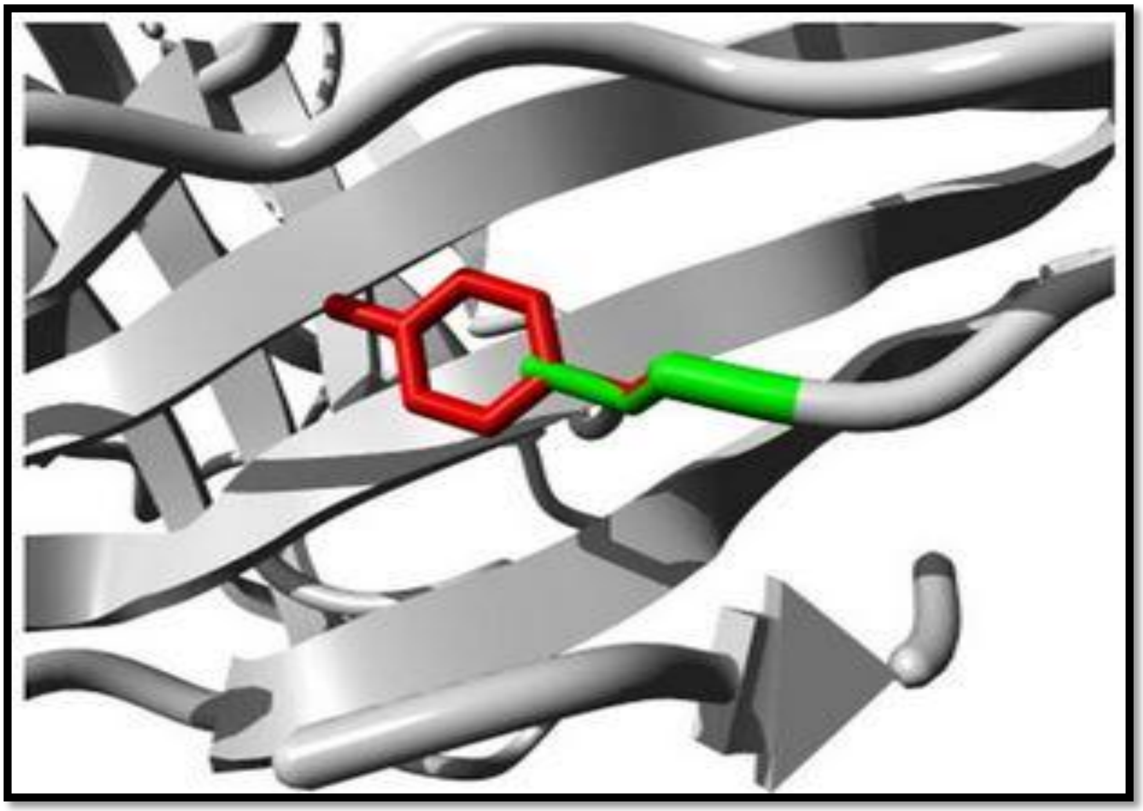


Figure 2: Image of mutated protein where protein is colored gray and the side chains of both wild type and new mutant residues are shown and colored green and red respectively.

It is well known that proteins are the main agents of biological function based on their interactions with other proteins and to other biomolecules such as DNA and RNA, mediating metabolic and signaling pathways and cellular processes. Protein interaction do not necessarily mean the physical association of proteins but can be metabolic interactions where proteins are involved in the same pathway or genetic interactions displaying co-expressed and co-regulated proteins (Gonzalex *et al.*, 2012). Any protein rarely acts alone but as mentioned above interacts with other protein to carry out a specific function. This pair of interacting protein forms a protein complex. The protein fragments within the complex that actually comes together and interact forms a special region called protein interface (Aung, Z. 2006). Thus, protein interaction networks are useful resources to gain knowledge about the evolution of proteins and about the different systems in which they are involved.

With the tremendous increase in human interaction data, it is important to understand how any biological function is exerted over the body and this can only be made possible by the inclusion of structural details in the networks, such details include three dimensional structure of network and protein interfaces. Protein interactions are mediated by specific recognition of distinct binding regions on the surface of interacting protein. Such recognition should be of sufficient affinity to effectively bind fragments of proteins, which is decided by specificities such as interfaces. Moreover, in protein-protein network, most of the proteins have few interactions whereas other proteins can have multiple interactions, these protein are central to the stability and normal functioning of the proteins in the network for example p53, p21, p27, BRCA1, ubiquitin, calmodulin are extensively involved in diseases such as different forms of cancer, and the product of these genes forms a hub. Deletions of these proteins, popularly known as hubs are lethal to the organism, hence it can be inferred that these proteins are encoded by essential genes and are important targets for molecular and structural studies (Choura *et al.*, 2011).

Protein-protein interactions (PPI) network is a way of representing how two or more proteins interact with each other and this network is the key to understand all the biological processes occurring within and between the cells (Figure 3). Protein-protein interaction network provides valuable information on the biological process and the cellular function (Chen *et al.*, 2009). Biological processes are essentially interactions between multiple proteins (Zhang *et al.*, 2011) with PPI networks controlling the flow of information both within and between biological processes. Network representation where proteins are nodes and interactions are edges is useful indicator of biological process and protein function (Choura *et al.*, 2011). Alteration in this network can potentially lead to various disorders because disorders are often caused by alteration in the binding sites or allosteric changes in the protein. For example, mutation in the zinc finger domain present in the oncoprotein MDM2 can disrupt the interaction of MDM2 with L5 AND L1 (ribosomal proteins) that mediates p53 degradation and leads to cancer (Zhang *et al.*, 2011).

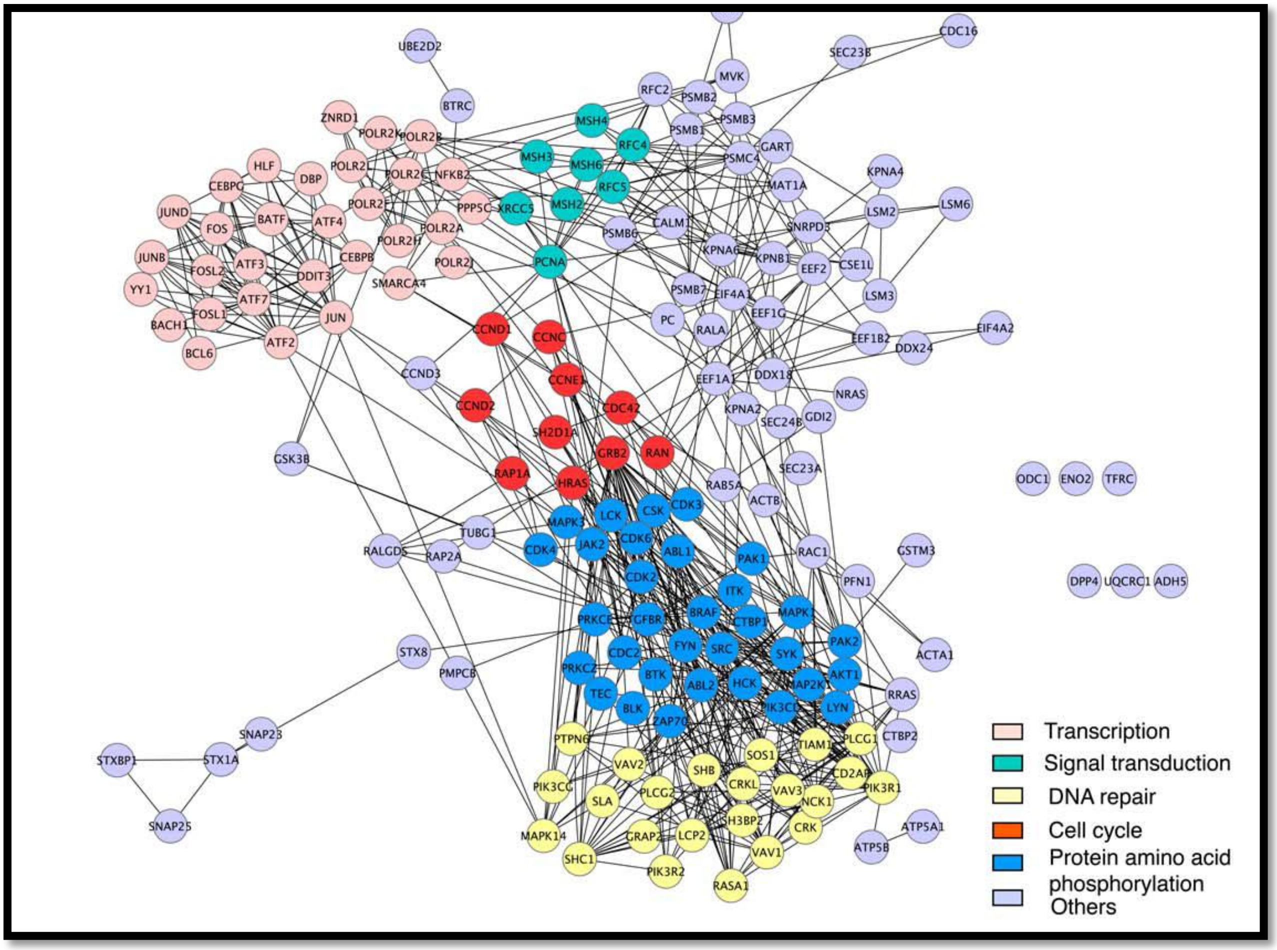


Figure 3: Image of protein-protein interaction network (Zhang *et al.*, 2011).

Thus, there is an increasing interest in these networks, as their analysis helps in understanding the relationship between proteins and how they are positioned in the whole system as proteins associated with a particular phenotype or function are not usually positioned randomly in the PPI network but exhibit high connectivity or may cluster together in central network locations (Nguyen *et al.*, 2011).

PPI networks can also be used to differentiate between a healthy and diseased patient (Steward *et al.*, 2003). This can be made possible by building interaction networks under different conditions. Thus, PPI network can offer various useful applications with respect to disease such as identification of a new disease, studying the network properties of already existing disorders and interpret various mechanisms related to disease. Further PPI networks can be used to identify disease related sub networks such as hubs and hot spots (essential genes) and can help in classifying diseases based on networks (Zhang *et al.*, 2011).

Studies have shown 39,000 or more interactions in human cell where disease genes are generally non essential and occupy peripheral positions in human network; hence these genes tend to cluster together and are highly interconnected (Legrain *et al.*, 2011).

Thus all in all, protein interaction networks helps unraveling the molecular basis of disease based on various physical and chemical properties of the proteins.

To characterize the function of these interactions, with respect to their physical and chemical properties, structural details play an important role. Structural profile of the proteins provides the opportunities in understanding the finer details of cellular functioning in terms of structural scaffolds and interfaces which facilitate the underlying molecular recognition events. Proteins interact with each other through their binding sites, thus interface properties and the topology of the proteins involved is very important in deciphering their function in the network. The strength of interaction is mostly determined by various parameters such as amino acid composition, gap volume, area in contact, the formation of salt bridges and hydrogen bonds (Gonzalez *et al.*, 2012). These interface properties can also determine whether the binding will be promiscuous or specific, how proteins in hub is different from non-hub proteins, and how a single protein can interact with multiple proteins with different binding affinity. In a broad group of disorders such as metabolic disorders, there are various diseases under this category. Each disease can be due to alterations in many proteins. This interaction is important in characterizing the important structural feature such as planarity, gap volume index, ASA, polar residues, non polar residues, charged residues in an interface.

3.1 <u>Protein interfaces – The regions where proteins interact</u>

The region where two protein chains come into contact is called as binding site, and for both sides involving residues of both proteins, it is known as interface. Protein interfaces tend to be planar, well packed and have great residue conservation. Hydrogen bonds, electrostatic interactions, van der waals forces, salt bridges and hydrophobic interactions determine the stability and specificity of the interacting proteins. Also, different complexes have different residue composition such as transient complexes rely more on salt bridges and hydrogen bonds, whereas the obligate complexes rely more on hydrophobic interactions (Figure 4). The interaction between proteins through interface also determines, whether the binding will be promiscuous or specific (Tuncbag *et al.*, 2009). This physical interaction is mainly governed by shape, chemical complementarities, flexibility of molecules involved and environmental conditions such that if two proteins are interacting with through large interface with high complementarity, they will probably interact with high affinity and high specificity (Kar *et al.*, 2009; Tuncbag *et al.*, 2009).

A prominent example of interface is revealed in cancer related proteins, where the analysis revealed that the interface for cancer proteins were smaller in size, more planar, less tightly packed and more hydrophilic than those of non-cancer proteins. Studies showed that cancer-related proteins tend to interact through multi-interface hubs and are longer and have larger surface areas. Thus to participate in multiple PPI, the proteins interact with distinct interfaces to different proteins (Kar *et al.,* 2009).

It has been determined that interfaces have more hydrophobic residues and fewer hydrophilic residues. Interfaces with hydrophobic residues are critical for the stabilization of protein–protein complexes. Various studies have shown that the contacts between residues with opposite charges, the contacts between hydrophobic residues and Cys–Cys contacts are more frequent across protein–protein interfaces. Hydrophobic interactions have been widely accepted to be the main stabilizing force for two proteins to interact. Further, interactions between pairs of hydrophilic residues are predominantly important; whereas hydrophobic interactions are important at longer distances. Cys-Cys pairs forms disulfide bond which contributes to the stability of protein-protein interaction. Apart from above interactions that help in strengthening the PPI, the presence of aromatic residues also helps in tighter packaging and better geometric fit as they form strong hydrophobic interactions between the bulky hydrophobic side chains (Yan *et al.,* 2008).

Protein–protein associations provide a useful link between structure and function of bimolecular systems thus allowing the characterization of the energetic of molecular complexes. Hence, detection of specific amino acid residues which add specificity and strength to protein interactions is of utmost importance. Therefore, the properties of interfaces are studied in detail:

3.2 Physical and chemical properties of Protein-Protein Interfaces

The interfaces are characterized by six properties: size and shape, complementarity, residue interface propensities, hydrophobicity, segmentation, secondary structure and conformational changes.

- **Size**
  PPIs are very complex and can be characterized based on their structural features such as size, shape and surface complementarity. The hydrophobic, electrostatic interactions and the flexibility of molecules involved are very important in establishing a good fit between the molecules (Moreira *et al.,* 2007). The chemical nature of protein interface is similar to the average protein surface (Tuncbag *et al.,* 2009). The standard size of the interface is approximately 1200-2000 Å^2. Interfaces with size around 1150-1200 Å^2 are low stability complexes generally short lived i.e. transient complexes. It is assumed that protein-protein binding energy is directly related to the buried hydrophobic surface area. Most of the protein heterodimer interfaces are larger than 600 Å^2, this cutoff corresponds to

minimum area required to make a water tight seal around energetically favorable interactions (Moreira *et al.,* 2007).

- **Chemical character**

  Hydrophobic interactions in the protein are the leading forces in PPIs and also in stability of interface. Interfaces are frequently hydrophobic in nature and bury large extent of non polar surface area. These interactions occur through the van der Waals contact between the nonpolar regions of their amino acid residues which results in the tight packing of residues organized as patches. These patches help in expulsion of water molecules in the interface, thus increasing the entropy that favors complex formation.

  Electrostatic forces are another important driving force for complex formation, as electrostatic complementarity of interacting protein surfaces promotes complex formation and defines the lifetime of the complex. In protein interfaces, 76% of the hydrogen bonds are formed by side chains of amino acids and other hydrogen bonds are formed between protein contact surfaces and the surrounding water molecules (Moreira *et al.*, 2007).

- **Conservation of the interfaces**

  Interface residues are more conserved than the rest of the protein surface (Tuncbag *et al.*, 2009).

- **Hot spots**

  The energetic contribution of amino acid residues are not distributed uniformly and only a few key residues contributes to the binding free energy of protein-protein complexes called “hot spots”. Hot spot residues are identified via Alanine Scanning Mutagenesis method. This method has been widely used to amp epitopes and residues as alanine substitution remove the side chain atom without introducing any additional conformational freedom. The principle behind alanine substitution is the role of side-chain functional groups at specific positions and the energetic contributions of individual side chains to protein binding can be inferred such that if a residue has significant drop in binding affinity when mutated to alanine, then the reissue is considered as hot spot. Glycine can also be used instead of alanine to study the function but glycine introduces conformational changes and thus not commonly used. There is a relationship between conserved residues and hot spots and these regions are buried and tightly packed resulting in densely packed clusters of network called “hot regions”. These regions contribute to the majority of binding affinity within the protein.

  The high propensity for interaction with diversity of partners is explained as; the same hot spot adapts to the same residues, showing high functional and structural adaptivity. Thus hot spots are helpful in understanding the binding sites for protein dimer and also protein multimer (Tuncbag *et al.*, 2009; Moreira *et al.*, 2007). Very few hot spots are at the edge of an interface rather they are centralized and compact residues essential for protein association.

- **Amino acid composition**

  The fundamental amino acids involved in interface are tryptophan, tyrosine and arginine. Tryptophan acts as a hot spot residue owing to its large size and aromatic nature. Tryptophan stabilizes the interface by its aromatic-p interactions, large hydrophobic surface and as hydrogen bonding donor. Additionally in Alanine Scanning Mutagenesis substitution of tryptophan creates a large cavity that causes destabilization of the complex. Tyrosine residue is also considered as hot spot due to high conservation propensity as well as the ability of participating in hydrogen bonding and aromatic-p interactions.

  Finally, arginine is prominent in interface residues due to its ability to form multiple types of favorable interactions such as forming five hydrogen bonds simultaneously and forming salt bridges within the interface. On the other hand, serine, leucine, threonine and valine are not favored and are usually absent as hot spots (Moreira *et al.,* 2007).

- **Complementarity – Clusters of hot spots**

  Hot spot of one protein against the hot spot of another protein establishes the region determining complex binding and tight fitting. These regions are characterized by complementary pockets scattered through central region of the interface and rich in structurally conserved residues. Complementary is defined by large complementarity both in shape and in juxtaposition of hydrophobic and hydrophilic hot spots. This is possible by formation of salt bridges by buried charged residues and fitting of hydrophobic residues from one surface into the small nooks on the other surface. Alignment of polar and non polar residues, number of buried waters, and size of buried surface and the packing densities of atoms involved in protein-protein interface are the factors affecting complemenatrity.

  Complemented pockets enriched in conserved residues are formed by residues across the protein-protein interface. Tryptophan is often found on the wall of complemented packet and occludes interactions from solvent within the pocket. Glycine is also important in some structural motifs as it lacks the side chain thus helps in tight packing and coupling with polar, aromatic and hydrophobic residues across the interface (Moreira *et al.,* 2007).

The parameters of interfaces for analyzing interactions in the present study are as follows:

- **Accessible Surface Area (ASA)** – this is the mean ASA buried by each domain of the complex.

- **Complementarity of interface- Gap volume** – provides a measure of complimentarity and closeness of packing of the interface between two interacting proteins by measuring the volume of empty space between them (Kar *et al.*, 2009).

- **Gap volume index** – the gap volume index between two protein domains is calculated as ratio of gap volume to the interface area. It estimates the volume enclosed between any two molecules, delimiting the boundary by defining a maximum allowed distance from both interface (Kar *et al.*, 2009).

- **Planarity** – the planarity of interfaces analyzes the shape of the interface. It is defined as the rmsd (root mean square deviation) of the interface atoms from the least-squares plane fitted through all interface atoms. The larger the planarity index, the less planar the interface. Moreover there is a correlation between the planarity of the interfaces and their ASAs. As the ASAs of the interfaces increase, the planarity index also increases, and the interfaces become less planar (Kar *et al.,* 2009).

- **Hydrogen bonding** – Hydrogen bonds play key role in specificity of interactions between two proteins. The average number of hydrogen bonds is proportional to the area of subunit surfaces i.e. ASA: one bond per each 100–200 Å^2 (Moreira *et al.*, 2007).

- **Sequence segmentation** – in interface this is defined as interface residue separated by more than five residues are counted in different segment (Jones *et al.,* 2000).

- **Residue propensity** – an interface residue propensity more than 1 indicates that a residue type is more prevalent in interface than rest of the protein surface (Jones *et al.,* 2000).

- **Number of salt bridges** – the more the number of salt bridges, there will be more electrostatic interactions resulting in tighter packing within the residues.

Protein interface has been studied and stored in databases such as PiBASE, InterPare, SCOWLP, 3DID, SCOPPI, PRINT and 2P2Idb (Tuncbag *et al.*, 2009). In the present study, 2P2Idb is used to find the interface parameters.

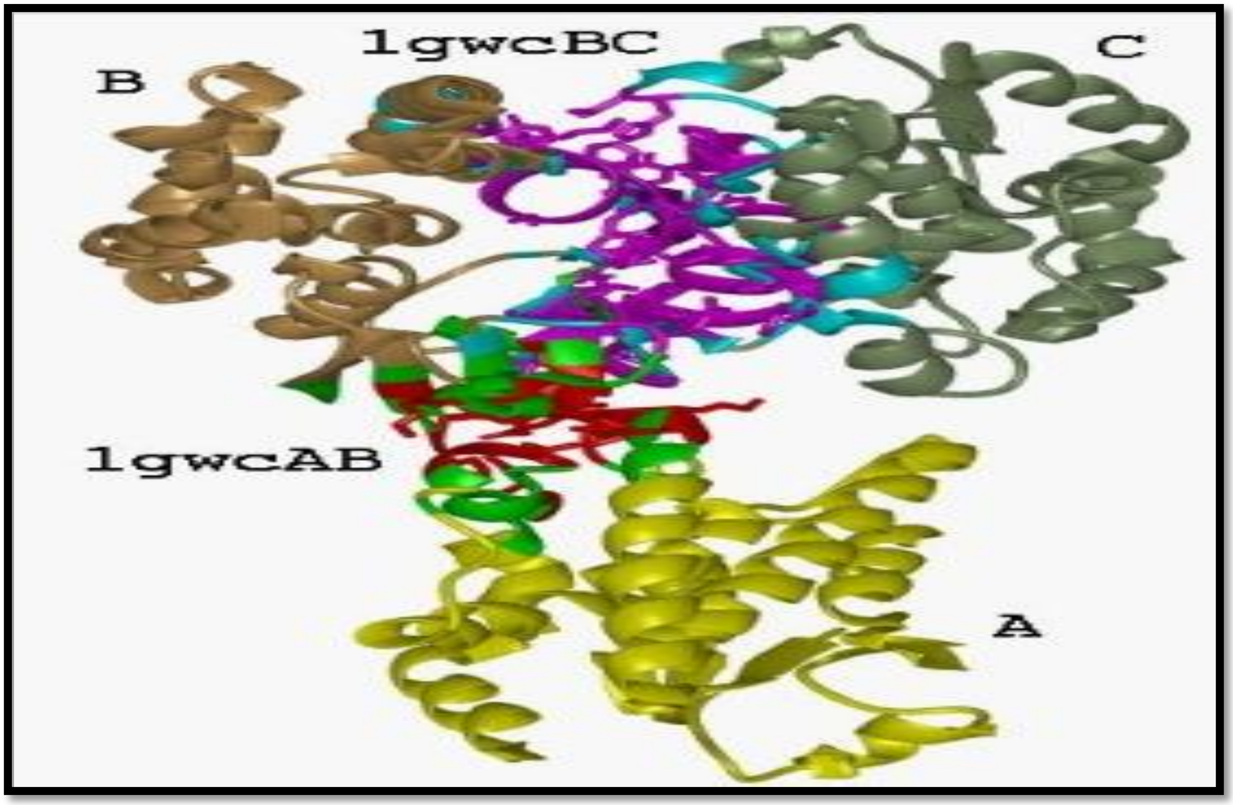


Figure 4: Depicts the interfaces between three chains of the protein glutathione s transferase. The PDB code is 1gwc. There are two interface between chains AB and BC. Chains A and C are not close enough to form interface. In the BC interface, the magenta is the contacting residues and cyan is the nearby residues. In the AB interface, red residues are the interacting residues and the yellow residues near the red residues are the neighboring (nearby) residues. The side chains of the interacting residues are also displayed in the figure.

3.3 Protein networks and diseases

3.3.1 Genetic diseases/ Genetic Basis of Disease

A disease is an abnormal condition impairing normal homeostasis, affecting a part or the entire organism. A genetic disease or disorder is a result of change or mutation in one's DNA. A mutation is the change in the genetic code that makes up the gene in the form of deletions, insertion, translocation and many others changes. Genes code for the proteins, the molecules that perform most of the life functions and make majority of cellular structures. Hence when a gene is mutated, its protein can longer perform the normal function thus disturbing the balance, resulting in the disorder. Genetic disorders can be caused by mutation in one gene, multiple genes, combination of genetic changes and environmental factors or by damage to the chromosome. As we know more about human genome, we come to know that almost all the diseases have genetic component. Some diseases are caused my mutations that are inherited from parents and present since birth such as color blindness, while other disease may develop during the life time of the individual where mutations occur randomly or due to environmental factors, such as cancers. These diseases are not inherited.

Further, genetic disorders have been categorized in three broad groups in previous studies; Monogenic disorders, also caused as mendelian disorders are caused my mutation in a single gene and are usually rare for example, sickle cell disease, cystic fibrosis. Multi-factorial inheritance disorder, are caused by mutations in multiple genes often acting together with environmental factors. Heart disease, diabetes, and cancers are the example of multi-factorial diseases. Third category of genetic disorders list chromosomal disorders which are caused by excess or deficiency of genes or structural changes within chromosomes for example Down syndrome.

The correlation between mutation and symptom is often not clear even in the mendelian disorders. Recent studies have shown that influence of other genes or environment and pleiotropy amongst genes plays a great role in defining the phenotype of the diseased individual. Pleiotropy occurs when a single gene produces multiple phenotypes. Mutation in these genes complicates disease elucidation as they can cause multiple syndromes or can affect various biological processes the gene mediates (Gonzalez *et al.*, 2012).

Usually, diseases are seen as similar based on their clinical appearance with no emphasis on underlying molecular process. Phenotypes of various diseases often overlap, recognition of this overlap brought the concept of ‘syndrome families’, accounting the common features shared between diseases (Sam *et al.,* 2007). It has long been known that mutations at different loci in the genome can lead to the same genetic disease and this genetic heterogeneity has its roots at the PPI level, suggesting that other genes associated with the phenotype also have some functional role. Therefore, it is possible that functional properties of shared molecular networks reflect phenotypic overlap of diseases. Thus, PPI networks provide unique opportunities for exploring disease pathways (Sam *et al.,* 2007).

3.3.2 <u>Molecular Basis of Disease</u>

Documentation of genes with the associated disease is not enough to understand the biological details of pathogenesis and disease progression. It is important to identify various molecules and mechanism triggering, participating and controlling the biological process. This understanding of molecular mechanism is a complex process and not much is known about the mechanism (Gonzalez *et al.,* 2012).

Further, by recent studies we know that a molecular mechanism leading to diseases is dependent on interlinked proteins and networks, but we are far from unraveling the difference in molecular interactions in healthy and diseased organism. Mechanism of genetic malfunctioning, that leads to one or several diseases can be understood when the molecular level of the protein interactions are known. Understanding the interactions and hence binding between proteins is critical for the rational design of new therapeutic agents targeted to disrupt the interactions that cause disorders (Figure 5) (Kann *et al.*, 2007).

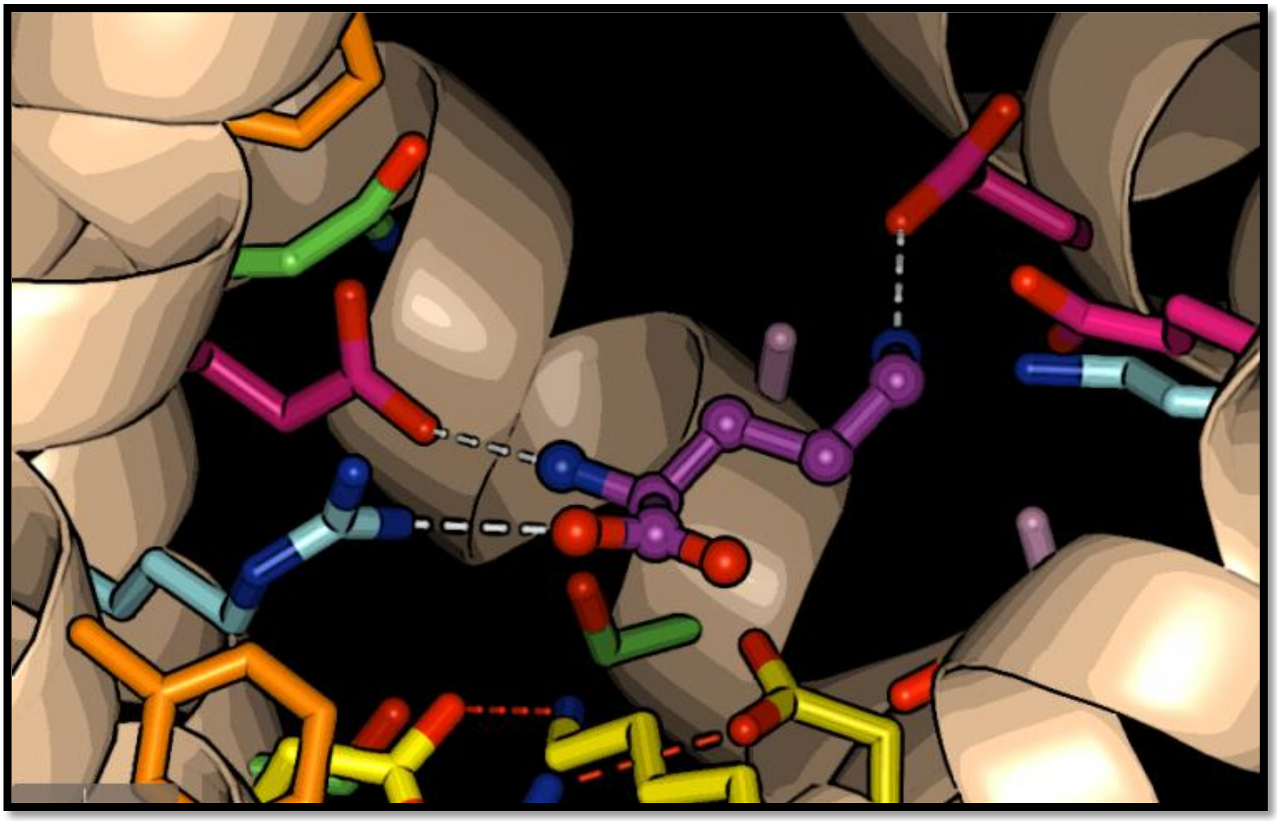


Figure 5: The above image shows how side chains of protein interact with each other to result in a phenotype. If this arrangement gets disturbed then it leads to disease.

3.3.3 Protein structure, protein complexes and disease

There has been an increasing emphasis on the structure of proteins, where NMR, X-ray crystallography have been used to generate high throughput structural resolution. The aim of structural genomics is to provide three dimensional structural models of the proteins encoded by entire genome. Many structures have been elucidated and have been stored in Protein Data Bank (PDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. Understanding the shape of these macromolecules helps to understand how it works. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome). Although around 40,000 known structures are deposited in PDB, but these structures might not be related to any human disease. Hence to deduce protein structures involved in diseases, various computational homology studies have been implemented to deduce function from protein sequence and structure information (Kann *et al.,* 2007).

A protein interaction provides a vast source of molecular information as their interactions are involved in various metabolic, signaling, immune, and gene-regulatory networks which are essential for normal homeostasis of the body. Now, since they are involved in normal functioning, they are also the key targets for molecular studies of biological diseased state (Gonzalez *et al.,* 2012)

Recent studies have shown that when structural data is combined with the information about mutation, the molecular mechanism of the disease can be unveiled, for example, a study by Thronton and collaborators showed that disease-related mutations are more likely to be buried in the protein structure than what would be expected for the average protein residues. Few examples where protein structural analysis have helped to elucidate the molecular basis of disease are; Von Hippel-Lindau syndrome (VHL) in which mutation from Tyrosine to Histidine at residue 98 (which is part of the binding site) disrupts the binding of the VHL protein to protein hypoxia-inducible factor (HIF) which leads to HIF accumulation resulting in angiogenic growth factor and local proliferation of blood vessels (Butte *et al.*, 2006). In some cases, during the binding process, interactions between two proteins might involve order-disorder transitions in partially disordered regions of the interacting proteins. These unstructured or disordered regions have been found to be involved in many disease mechanisms. For instance, the cancer suppressor BRCA1 has been shown to contain intrinsically disordered regions through which it binds to several proteins (Kann *et al.,* 2007).

Several studies have been carried out with respect to particular disease, for example in Huntington's disease protein-protein interaction network was generated with all proteins related to HD. HD is caused by the repeat expansion of CAG in the Huntingtin (Htt) gene which causes aggregation of the mutant Htt in insoluble neuronal inclusion bodies which consequently leads to neuronal degeneration. Now, the network revealed many new interactions and functional annotation was carried out for several uncharacterized proteins. Importantly, an interaction of the Htt with GIT1, a GTPase-activating protein which seems to be required for the Htt aggregation was discovered (Goehler *et al.*, 2007). Thus targeting GIT1 protein can be useful in designing drugs and other therapeutical measures. Other studies involving ataxias and Purkinje cell degradation have shown that most of the proteins interact directly or indirectly with each other (Kann *et al.,* 2007). These examples show that proteins involved in a disease are likely to interact with proteins already known to cause similar diseases.

Another important characteristic of studying PPI is the discovery of disease markers, which can be used to access the outcome of various diseases. Studies have been carried out where the gene expression profiles combined with PPI network has been used to predict the outcome of cancer patients that is if they are under the risk of benign or malignant tumor.

A systematic computational study of a subset of proteins related to cancer was performed in which the authors found that the network topology of the cancer related proteins is quite different from those not involved in disease. They found that cancer proteins are highly interconnected with other cancer related proteins than other proteins (Jonsson *et al.,* 2006).

Other than the interference in protein-protein interaction network, several diseases are caused due to disruption in protein-DNA interaction, protein folding mechanism, protein-RNA disruption, can enable pathogen host protein interaction or can lead to new undesired protein interactions (Gonzalez *et al.,* 2012). New interactions can alter homeostasis due to misfolding and aggregation which leads to the loss of vital cellular functions and can cause toxicity (Duennwald *et al.,* 2006). The disruptions that lead to the establishment of Pathogen-host protein interactions also play a key role in bacterial and viral infections by facilitating the hijacking of the host's metabolism for microbial need. For example the infection caused by Human papillomavirus (HPV) generates lesions of anogenital tract and leads to cancer. HPV infection bypasses the immune system by interacting with important negative cell regulatory proteins to target them for degradation and inactivation. This leads to cellular transformation, immortalization of host cells and proliferation of tumorigenically-transformed cells (Scheffner *et al.,* 2003).

Further it is important to understand that pathways are different from PPI networks. Pathways comprises of series of sequential biochemical reactions involving metabolic, genetic and signaling process, where substrates are changed in a linear fashion into products whereas PPI networks only map the functional and physical interaction between protein pairs resulting in a complex grid of connections. Now, pathway analysis cannot be used to decipher the molecular basis of disease as most of genes involved in a particular process have not been assigned to a pathway. PPIs not only identify disease-associated interacting proteins but also the potentially interesting disease-associated gene candidates (genes coding for interacting proteins are putative disease causing genes). This assumption has been used in the present work where all the cardiovascular disease associated genes have been retrieved and their interacting partners have been identified. This methodology has been used to identify novel disease genes by finding interaction partners of known disease associated proteins. This has been confirmed by studies that found that mutations on the genes of interacting proteins lead to similar disease phenotype because of their functional relationship (Gonzalez *et al.,* 2012).

In reality, proteins are continuously being synthesized and degraded in the body, this kinetics of the process and network dynamics need to be considered to have complete understanding of how disruptions of protein and their interaction leads to disease (Kann *et al.,* 2007).

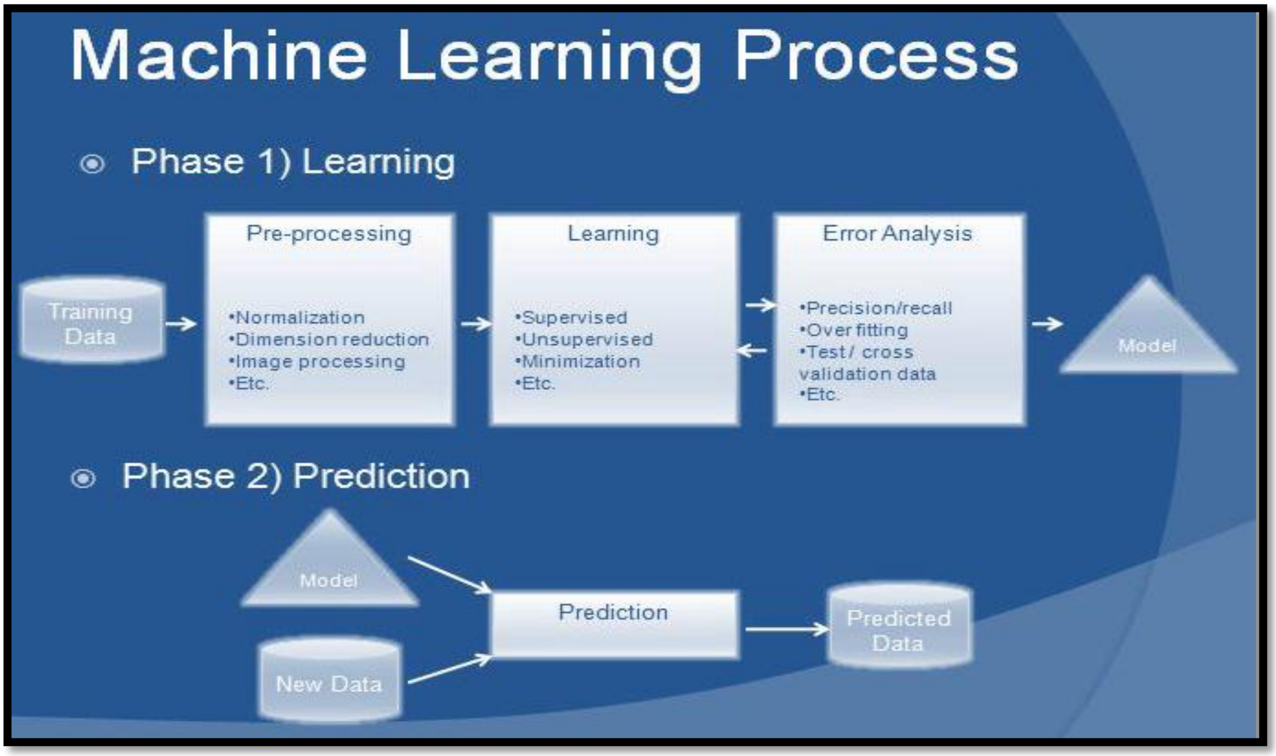
3.4 Machine learning and methods

Machine learning, a branch of artificial intelligence, is a system that acquires and integrate knowledge through training, experience and analytical observation and used to make predictions and classification of compounds based on its learning (Figure 6). It can be defined as : "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997). Simply, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.

In machine learning a known sample is provided first, which trains the algorithm and then the corresponding knowledge acquired is used to test, analyze and interpret the unknown data.

Machine learning algorithms can be broadly classified into two groups:

- Supervised learning generates a set of function that screens the inputs into desired outputs (labels). In this, the data is pre-assigned to particular classes that train the model. Also called inductive learning.
- Unsupervised learning labels are not known during training. No pre-assignment of the data into classes.



Machine Learning Process

Phase 1) Learning

Training Data

Pre-processing
•Normalization
•Dimension reduction
•Image processing
•Etc.

Learning
•Supervised
•Unsupervised
•Minimization
•Etc.

Error Analysis
•Precision/recall
•Over fitting
•Test / cross validation data
•Etc.

Model

Phase 2) Prediction

Model

New Data

Prediction

Predicted Data

Figure 6: The machine learning process showing learning and prediction as its two phases.

3.4.1 Supervised learning:

Supervised machine learning algorithms discover patterns in the data that relate data attributes with a target (class) attribute. These patterns are then utilized to predict the values of the target attribute in test data instances. The classes used for training are pre-determined and based on the patterns searched the mathematical models are constructed (Figure 7).These models then are evaluated on the basis of their predictive capacity in relation to measures of variance in the data itself. Supervised learning is mostly performed for classification tasks (Manchanda *et al.*, 2007).

Different supervised learning processes include decision trees, Bayesian Classification, Neural networks, Support Vector Machines, Genetic algorithm etc.

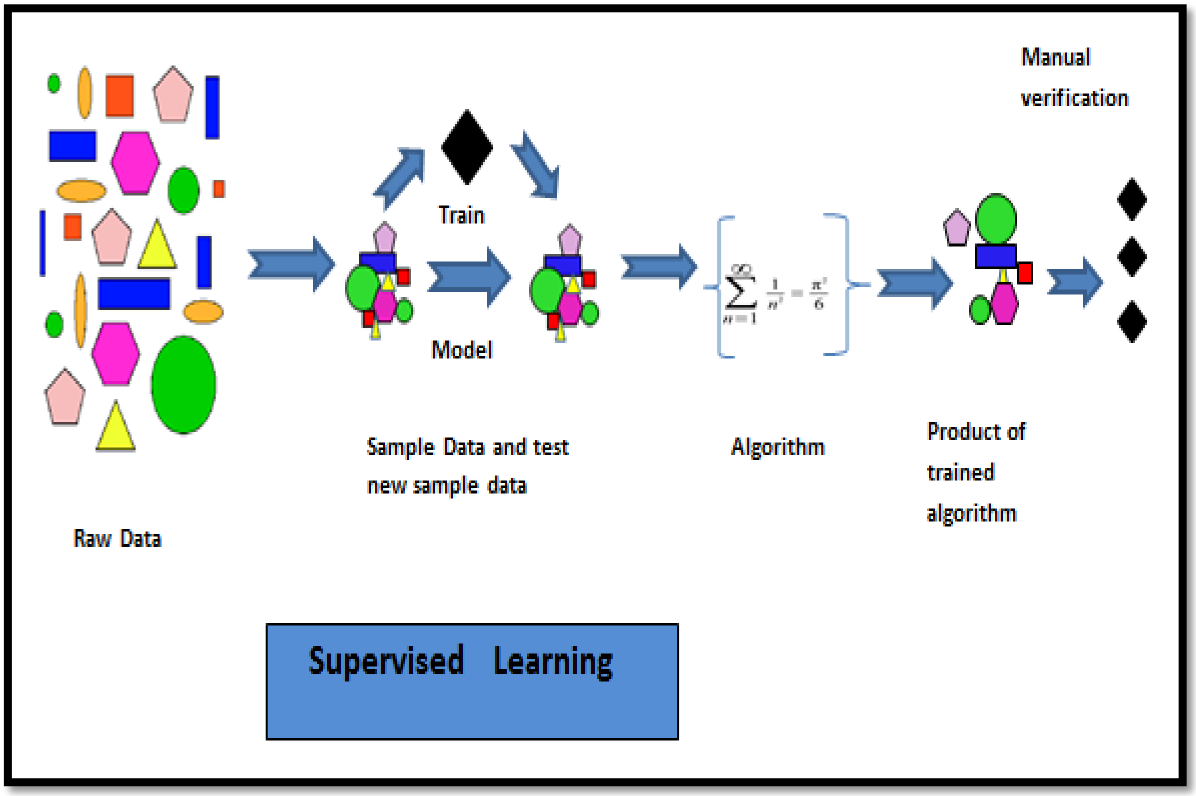

Manual verification
Train
Model
$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$
Sample Data and test new sample data
Algorithm
Product of trained algorithm
Raw Data
Supervised Learning

Figure 7: Process of supervised learning**.**

Random Forest:

The algorithm is based on decision trees. Random Forests are a combination of tree predictors in which multiple classification trees are constructed from an independent identically distributed random input vector. It is trained in such a way that each object is classified based on certain decisions made on the node of the tree which is dependent on certain pre-defined variables. Individual trees are constructed using bootstrapping, each with different attributes. Each random redistribution is generated by randomly drawing with replacement *N* examples where *N* is the size of the training set. A tree is grown on a fixed-size subset of attributes (smaller than the total number of attributes) randomly drawn on each round (Figure 8). Multiple random trees are constructed by repeating this method. After a large number of trees are generated, each tree in the forest gives a classification or votes for a class and the most popular class gives the final classification (Breiman, 2001). The misclassification error is calculated to predict the performance of the model.

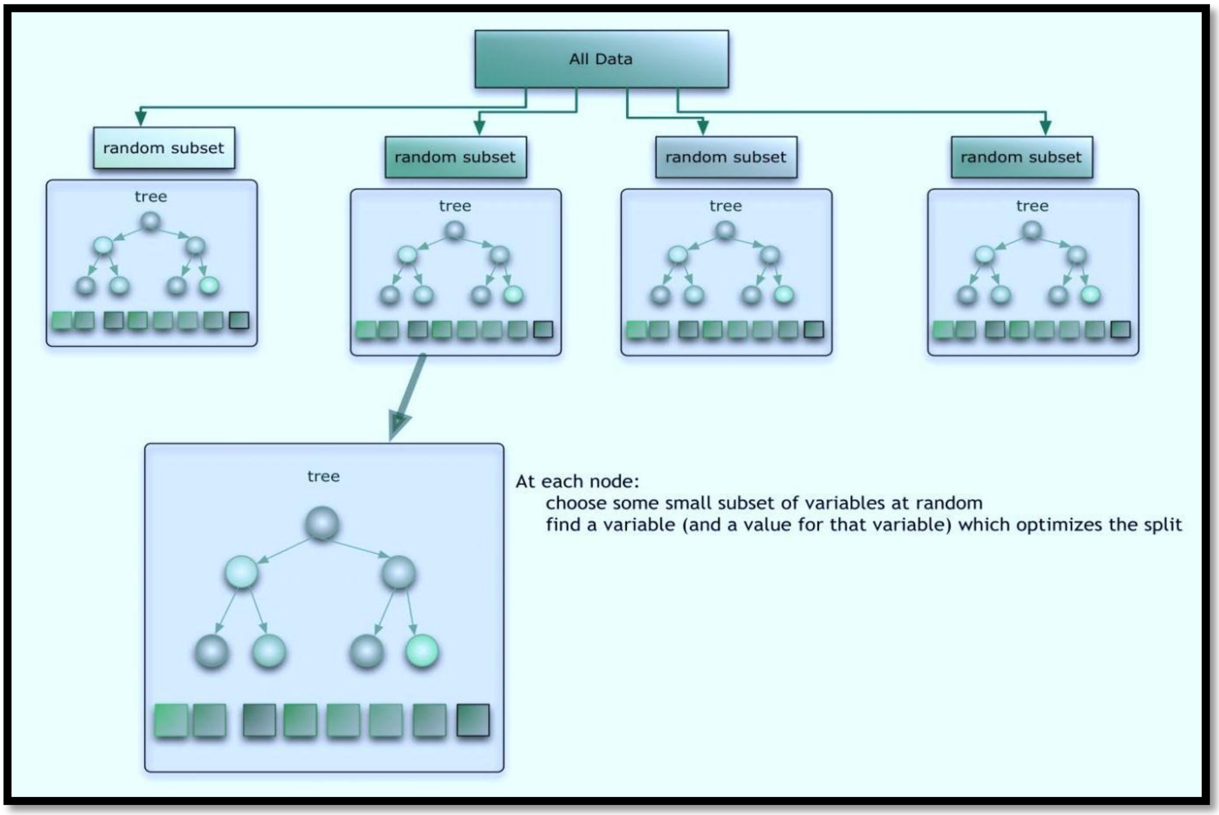

All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 8: Construction of ensemble of trees in random forest algorithm.

Training by Random Forest algorithm for some number of trees T:

1. Sample N cases at random with replacement to create a subset of the data (see top layer of figure above). The subset should be about 66% of the total set.
2. At each node:
   i. For some number m (see below), m predictor variables are selected at random from all the predictor variables.
   ii. The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
   iii. At the next node, choose another m variable at random from all predictor variables and do the same.

Depending upon the value of m, there are three slightly different systems:

- Random splitter selection: m =1
- Breiman‘s bagger: m = total number of predictor variables
- Random forest: m << number of predictor variables. Brieman suggests three possible values for m: ½√m, √m, and 2√m.

When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority (Figure 9).

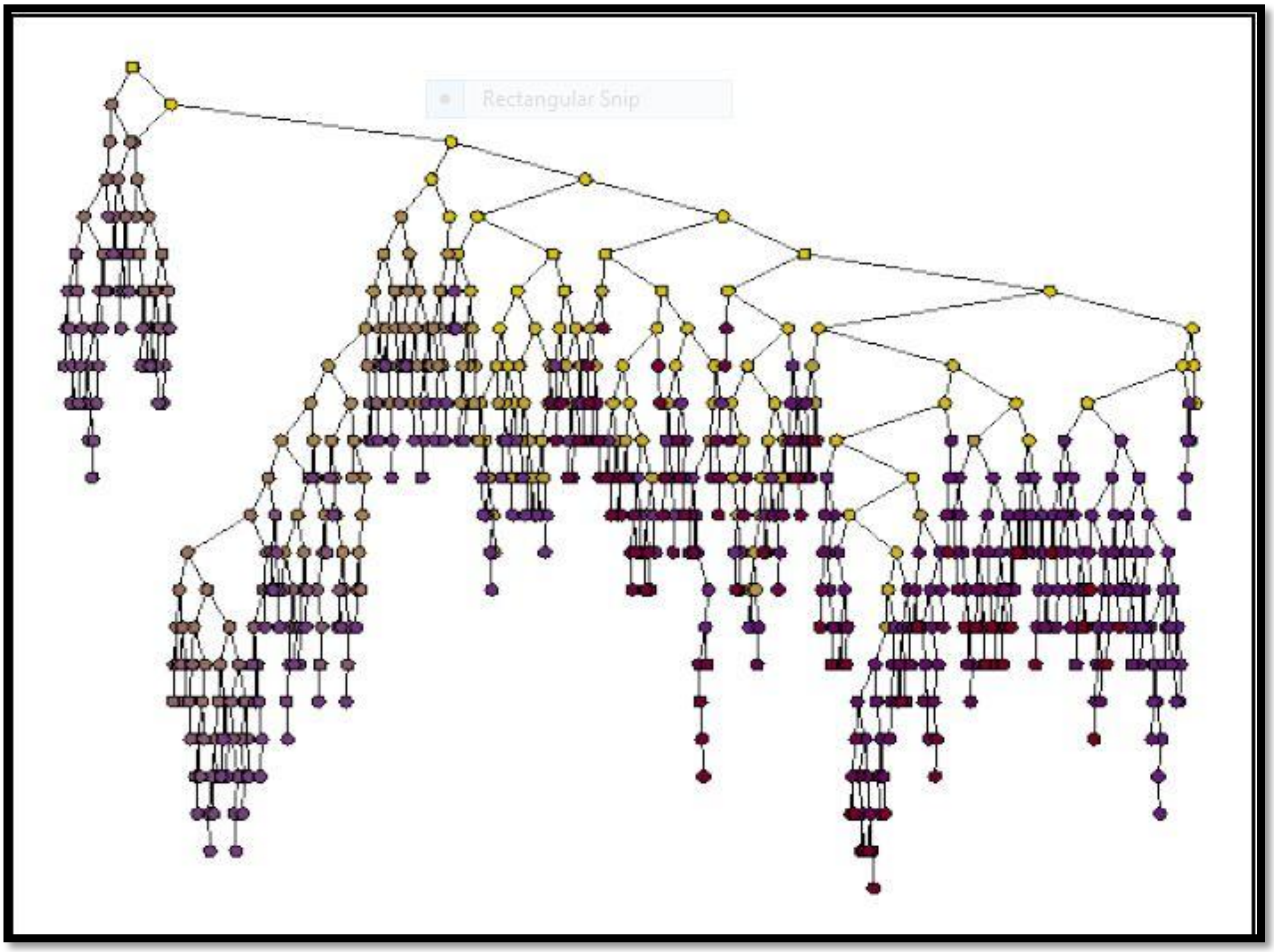


Figure 9: Manual view of Random Forest.

Strengths:

- Random forest runtimes are quite fast, and
- They are able to deal with unbalanced and missing data.
- Capable of handling of large input variables without over-fitting.
- The accuracy is maintained on larger sets.

Weaknesses:

- When used for regression they cannot predict beyond the range in the training data.
- They may over-fit data sets that are particularly noisy.

3.4.2 Unsupervised learning:

The data have no target attribute. It is explored to find some intrinsic structures in them. Unsupervised learners are not provided with classifications. So, the basic task of unsupervised learning is to develop classification labels. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. These groups are termed clusters, and there is a whole family of clustering machine learning techniques. Clustering groups the data instances that are similar to each other in one cluster and data instances that are very different from each other into different clusters (Figure 10). Hence, clustering is often called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given.

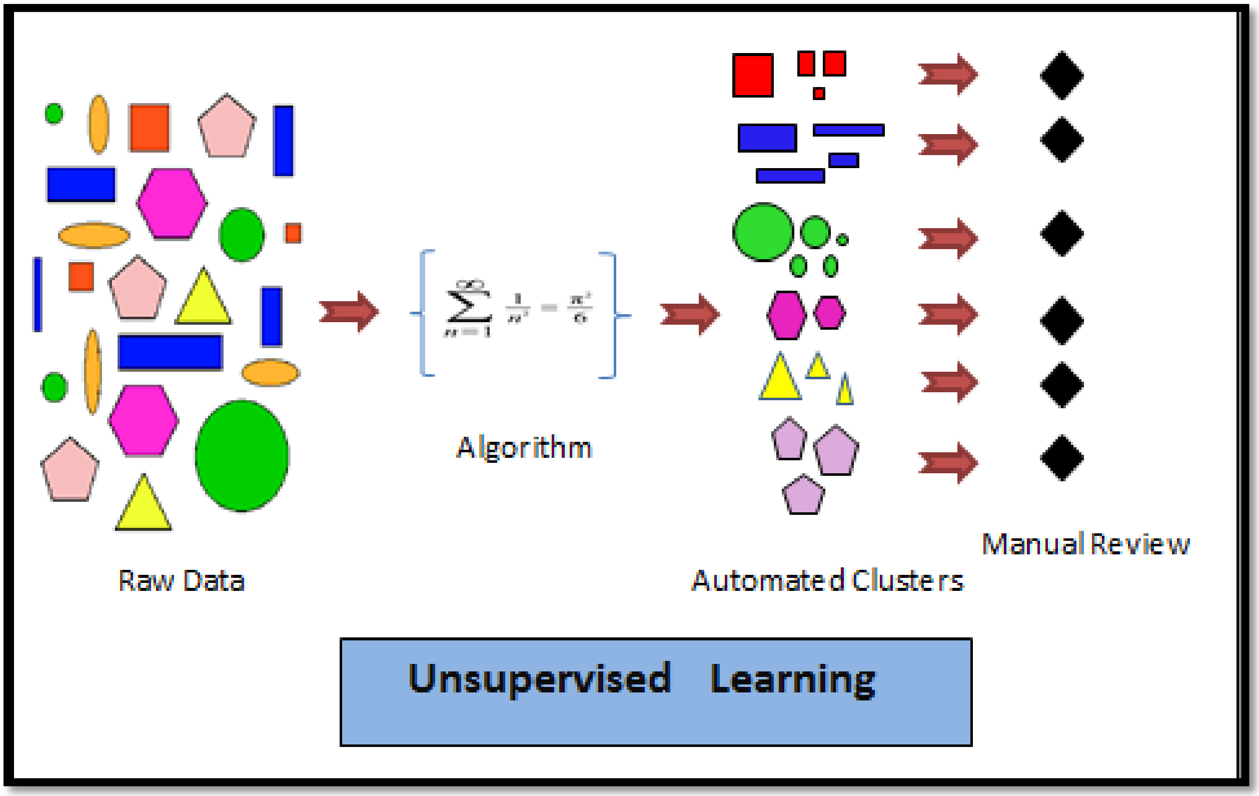


Figure 10: Process of unsupervised learning.

Different types of clustering algorithms are known in machine learning: k-means clustering, hierarchical clustering, Cobweb, overlapping clustering etc.

Hierarchical Clustering

Hierarchical Clustering algorithm produces a nested sequence of clusters, a tree, also called Dendogram. The base of the hierarchy gives the initial structures and subsequent levels provide smaller to larger clusters.

Types of hierarchical clustering:

Agglomerative (bottom up) clustering:

It builds the dendrogram (tree) from the bottom level, each data point forms a cluster (also called a node) and merges the most similar (or nearest) pair of clusters or nodes. It stops when all the data points are merged into a single cluster i.e., the root cluster (Figure 11). It is more popular then divisive methods.

Algorithm:

1. Make each data point in the data set D a cluster.
2. Compute all pair-wise distances of x1,x2,….., xn € D.
3. Repeat
4. Find two clusters that are nearest to each other.
5. Merge the two clusters and form a new cluster c.
6. Compute the distance of c from all other clusters.
7. Repeat, until there is only one cluster left.

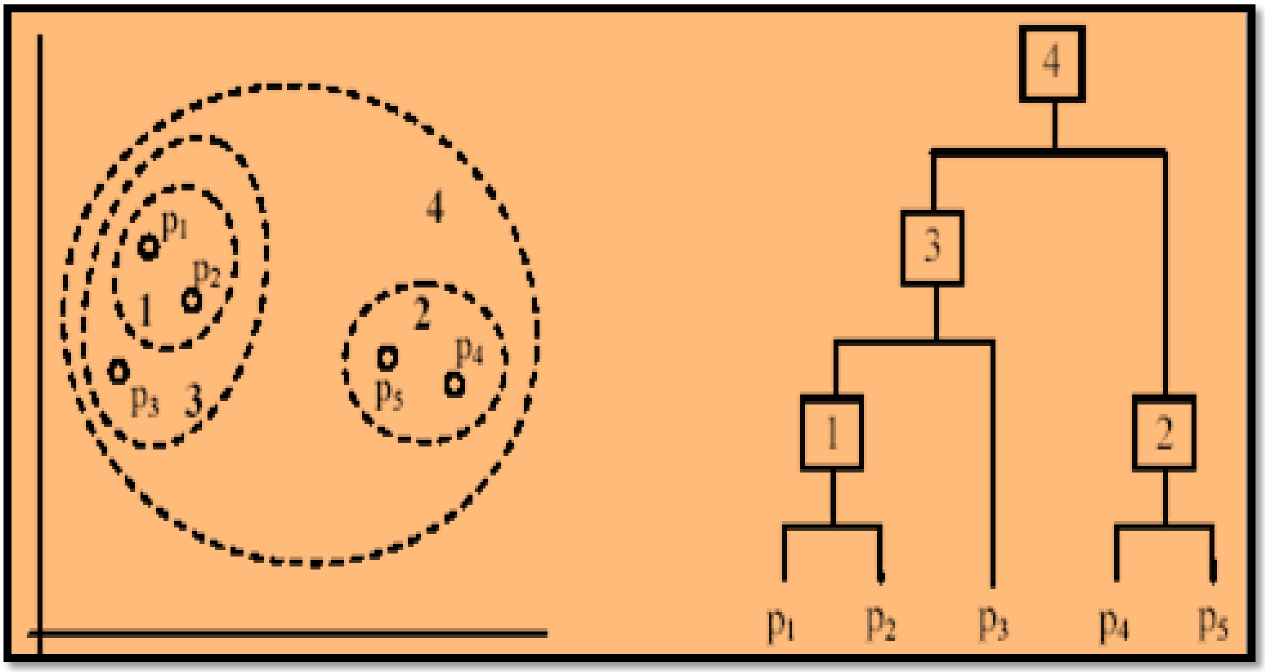


Figure 11: Output of hierarchical clustering algorithm showing nested clusters (left) and dendogram (right).

<u>Divisive (top down) clustering:</u>
It starts with all data points in one cluster, the root. The main root then splits into a set of child clusters. Each child cluster is recursively divided further. The iteration stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point.

In this study supervised learning using WEKA (a machine learning tool kit) using classifiers - Random Forest is used to predict the disease-associated protein based on the interface properties.

The impact of prediction methods is considered to be huge. Prediction method uses two criterions-:

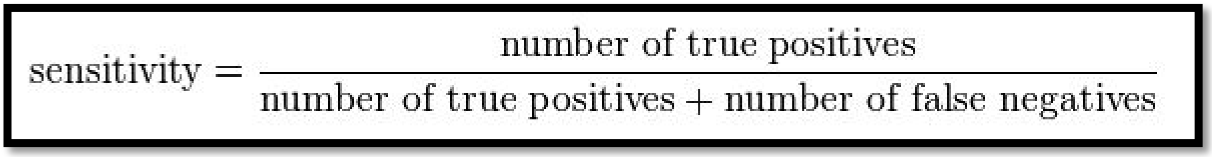
- **Precision-**: it is expressed as percentage of correctly predicted disease protein over all predicted disease protein.
- **Recall-**: it is the ability of method to detect disease-associated proteins in the test set. It is expressed as percentage of correctly predicted disease-associated proteins over all predicted disease-associated protein.

High precision means that any predicted protein is likely to be disease-associated. High recall means that method is able to correctly predict a large portion of proteins in a test set. A trade-off is required to get high precision and recall.

Precision= tp/(tp + fp), Recall= tp/(tp + fn)

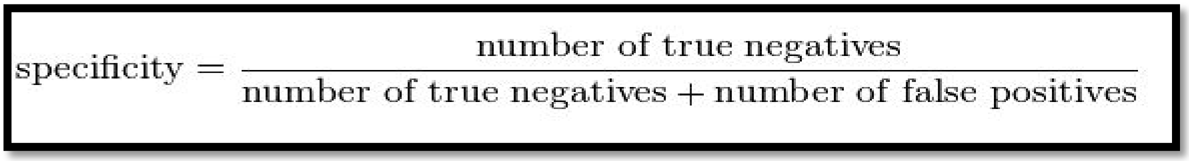
Sensitivity and specificity are statistical measures of the performance of a binary classification test.

Sensitivity (called the true positive rate, or the recall rate) measures the proportion of actual positives which are correctly identified as such.



$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Specificity (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such.



$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Tp= true positive (correctly predicted disease protein)
Tn= true negative (non disease protein predicted as non-disease protein)
Fp= false positive (non disease protein that is predicted as an disease associated protein)
Fn= false negative (disease associated protein predicted as non-disease protein)

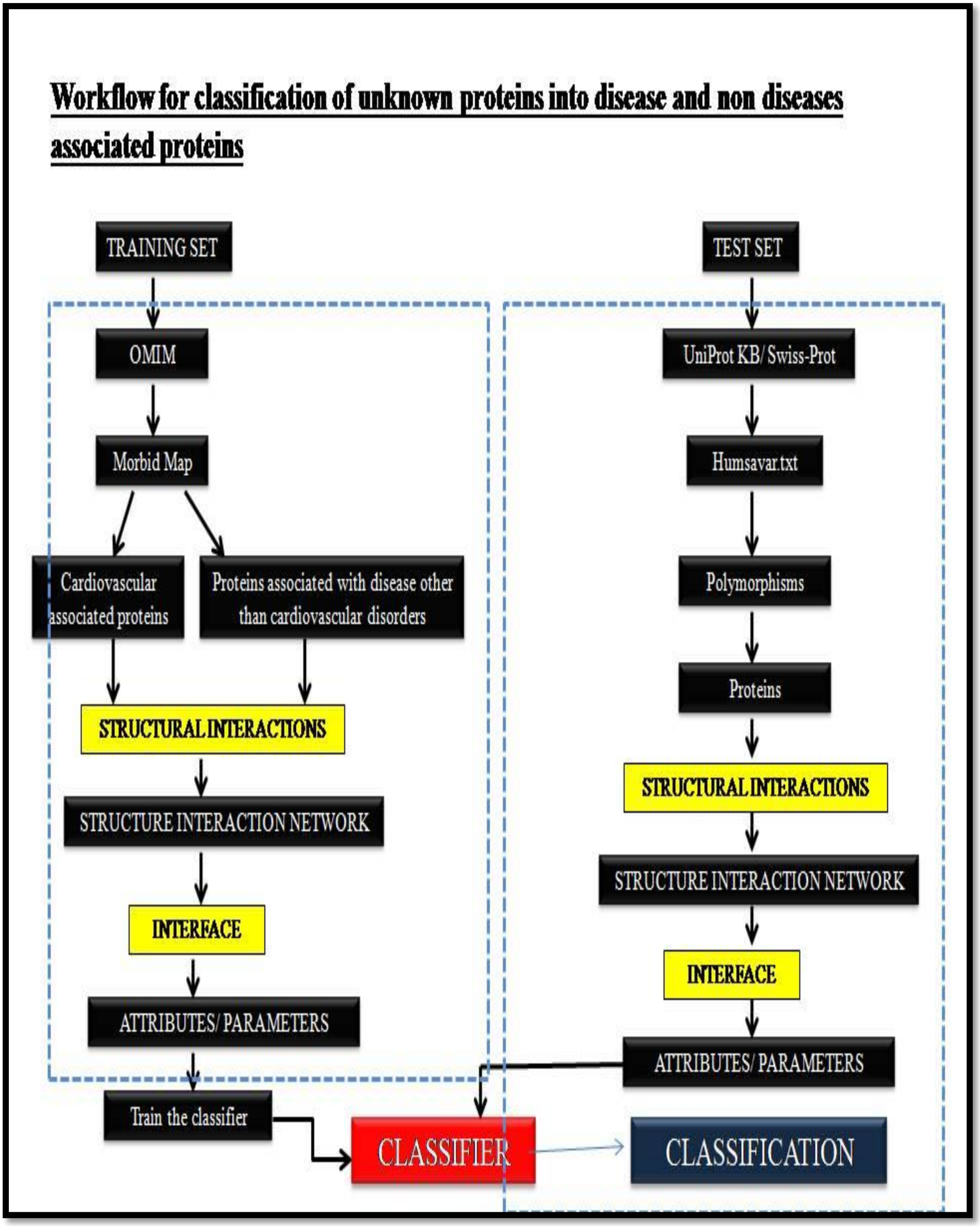

Workflow for classification of unknown proteins into disease and non diseases associated proteins
TRAINING SET
TEST SET
OMIM
Morbid Map
Cardiovascular associated proteins
Proteins associated with disease other than cardiovascular disorders
STRUCTURAL INTERACTIONS
STRUCTURE INTERACTION NETWORK
INTERFACE
ATTRIBUTES/ PARAMETERS
Train the classifier
CLASSIFIER
UniProt KB/ Swiss-Prot
Humsavar.txt
Polymorphisms
Proteins
STRUCTURAL INTERACTIONS
STRUCTURE INTERACTION NETWORK
INTERFACE
ATTRIBUTES/ PARAMETERS
CLASSIFICATION

Figure 12: Workflow of the present study.

4. METHODOLOGY

4.1 Building training sample:

OMIM Morbid Map was used to obtain training and test sample.

Training sample: consisted of a positive set and a negative set. Positive set: list of all the proteins associated with the cardiovascular disorders. There are a total of 124 diseases in the class of cardiovascular disorders. For each disease, concerned proteins were listed. Negative set: list of protein set of diseases other than cardiovascular diseases.

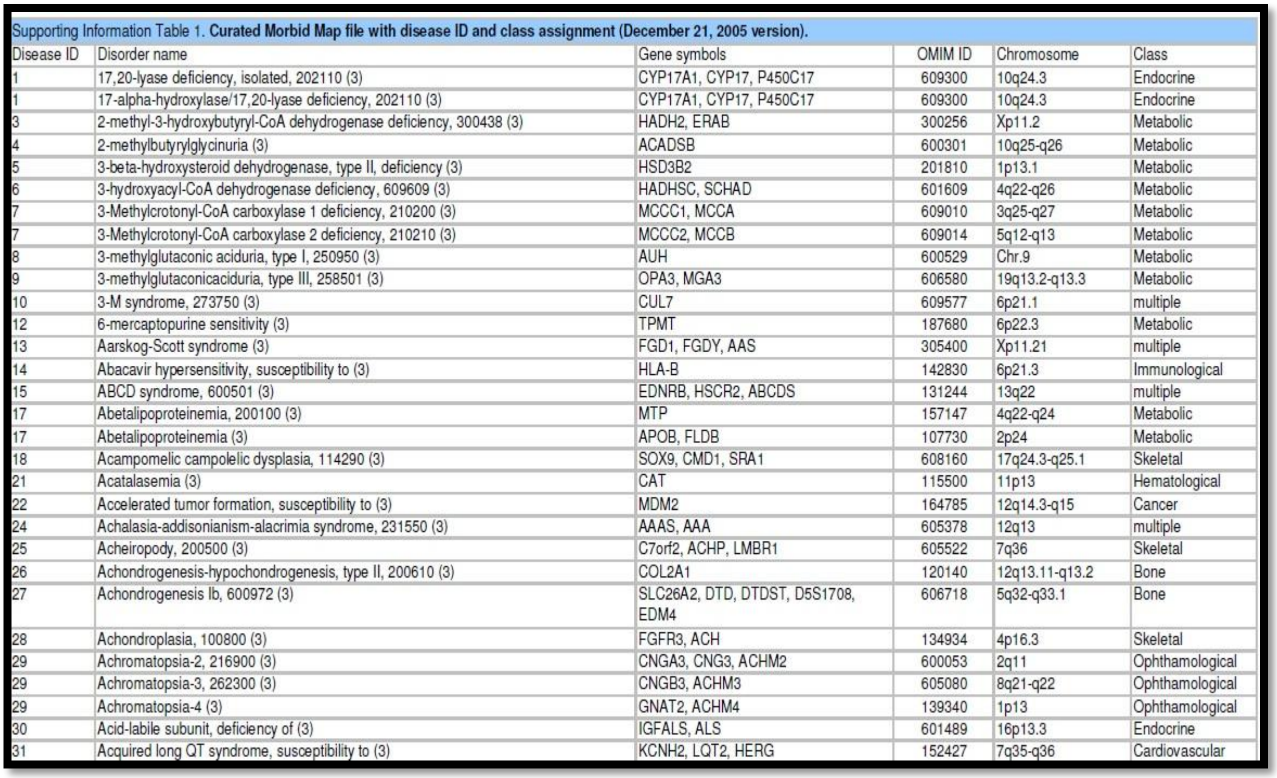
Test sample: list of all the proteins which are not known to be involved in the disease. These are the proteins with polymorphisms but their association with disease is not yet established. Proteins were retrieved from humsavar.txt – UniProtKB. It is text file describing human polymorphism and disease mutation and contains disease variants, polymorphisms and unclassified variants. In the present study, we used protein variants under polymorphism as test set.

4.2 Retrieval of dataset

There are many databases for information on protein-protein interacting partners such as HPRD, DIP, IntAct, MIPS, but none of the database provides the information on structural interactions amongst proteins. To retrieve such interactions, various databases were used:

- OMIM
- UNIPROT
- PDB
- NCBI

We used morbid map from OMIM (Online Mendelian Inheritance in Mammals) database to identify all the genes involved in cardiovascular class of disorders (Table 1). The OMIM database is one of the largest catalogs of human genes and disorders and focuses on inheritable and heritable disease within which morbid map provides the most complete and best curated list of known disorder-gene associations. In morbid map, there are four fields; the name of the disorder, the associated gene, corresponding OMIM id and the chromosomal location. In a study by Kar *et al.*, they classified the disorders into classes based on the physiological system affected. Using the information in morbid map, an excel sheet was made with cardiovascular disorders listed along with the genes.



Supporting Information Table 1. **Curated Morbid Map file with disease ID and class assignment (December 21, 2005 version).**

Disease ID	Disorder name	Gene symbols	OMIM ID	Chromosome	Class
1	17,20-lyase deficiency, isolated, 202110 (3)	CYP17A1, CYP17, P450C17	609300	10q24.3	Endocrine
1	17-alpha-hydroxylase/17,20-lyase deficiency, 202110 (3)	CYP17A1, CYP17, P450C17	609300	10q24.3	Endocrine
3	2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency, 300438 (3)	HADH2, ERAB	300256	Xp11.2	Metabolic
4	2-methylbutyrylglycinuria (3)	ACADSB	600301	10q25-q26	Metabolic
5	3-beta-hydroxysteroid dehydrogenase, type II, deficiency (3)	HSD3B2	201810	1p13.1	Metabolic
6	3-hydroxyacyl-CoA dehydrogenase deficiency, 609609 (3)	HADHSC, SCHAD	601609	4q22-q26	Metabolic
7	3-Methylcrotonyl-CoA carboxylase 1 deficiency, 210200 (3)	MCCC1, MCCA	609010	3q25-q27	Metabolic
7	3-Methylcrotonyl-CoA carboxylase 2 deficiency, 210210 (3)	MCCC2, MCCB	609014	5q12-q13	Metabolic
8	3-methylglutaconic aciduria, type I, 250950 (3)	AUH	600529	Chr.9	Metabolic
9	3-methylglutaconicaciduria, type III, 258501 (3)	OPA3, MGA3	606580	19q13.2-q13.3	Metabolic
10	3-M syndrome, 273750 (3)	CUL7	609577	6p21.1	multiple
12	6-mercaptopurine sensitivity (3)	TPMT	187680	6p22.3	Metabolic
13	Aarskog-Scott syndrome (3)	FGD1, FGDY, AAS	305400	Xp11.21	multiple
14	Abacavir hypersensitivity, susceptibility to (3)	HLA-B	142830	6p21.3	Immunological
15	ABCD syndrome, 600501 (3)	EDNRB, HSCR2, ABCDS	131244	13q22	multiple
17	Abetalipoproteinemia, 200100 (3)	MTP	157147	4q22-q24	Metabolic
17	Abetalipoproteinemia (3)	APOB, FLDB	107730	2p24	Metabolic
18	Acampomelic campolelic dysplasia, 114290 (3)	SOX9, CMD1, SRA1	608160	17q24.3-q25.1	Skeletal
21	Acatalasemia (3)	CAT	115500	11p13	Hematological
22	Accelerated tumor formation, susceptibility to (3)	MDM2	164785	12q14.3-q15	Cancer
24	Achalasia-addisonianism-alacrimia syndrome, 231550 (3)	AAAS, AAA	605378	12q13	multiple
25	Acheiropody, 200500 (3)	C7orf2, ACHP, LMBR1	605522	7q36	Skeletal
26	Achondrogenesis-hypochondrogenesis, type II, 200610 (3)	COL2A1	120140	12q13.11-q13.2	Bone
27	Achondrogenesis Ib, 600972 (3)	SLC26A2, DTD, DTDST, D5S1708, EDM4	606718	5q32-q33.1	Bone
28	Achondroplasia, 100800 (3)	FGFR3, ACH	134934	4p16.3	Skeletal
29	Achromatopsia-2, 216900 (3)	CNGA3, CNG3, ACHM2	600053	2q11	Ophthamological
29	Achromatopsia-3, 262300 (3)	CNGB3, ACHM3	605080	8q21-q22	Ophthamological
29	Achromatopsia-4 (3)	GNAT2, ACHM4	139340	1p13	Ophthamological
30	Acid-labile subunit, deficiency of (3)	IGFALS, ALS	601489	16p13.3	Endocrine
31	Acquired long QT syndrome, susceptibility to (3)	KCNH2, LQT2, HERG	152427	7q35-q36	Cardiovascular

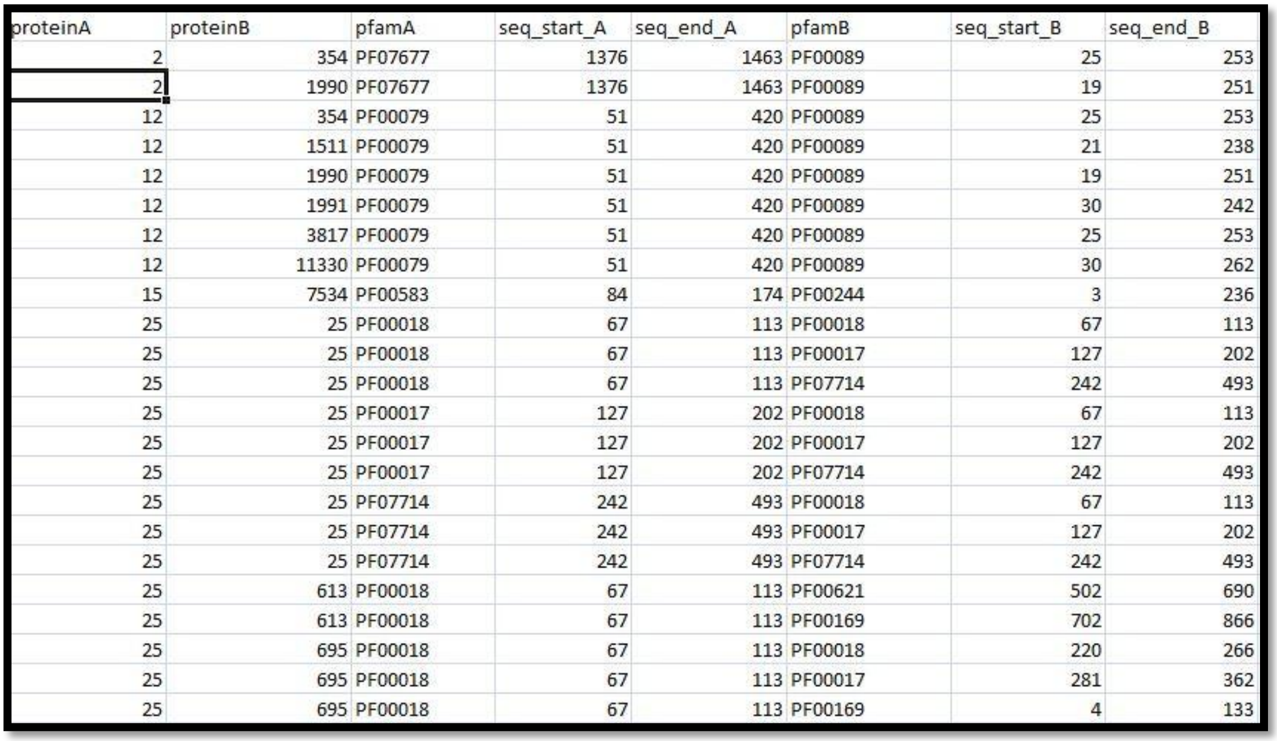
Table 1: OMIM Morbid Map.

The respective proteins of the genes and their gene ids were retrieved from NCBI (National Centre for Biotechnology Information). The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. It is a collection of 44 databases. The Protein database is a collection of sequences from several sources such as-:
GenBank, RefSeq, SwissProt, PDB.

4.3 <u>Retrieval of the interacting proteins</u>:

Wang *et al.*, 2012; found high quality binary protein-protein interactions along with the atomic resolution interface. They combined reliable literature-curated binary interactions and well verified yeast two-hybrid interactions to produce human protein interaction network (Wang *et al.*, 2012). Further, they structurally resolved the interfaces of these interactions using iPfam and 3did utilizing homology modeling approaches. This resulted in a human structural interaction network (hSIN). This information is stored in the database with the gene id and Pfam id.

hSIN was used to determine the structurally relevant interactions to the proteins involved in the cardiovascular interactions. Gene ids for the disease associated proteins were retrieved from NCBI, and the interacting partners with reliable structure were obtained from hSIN (Table 2).

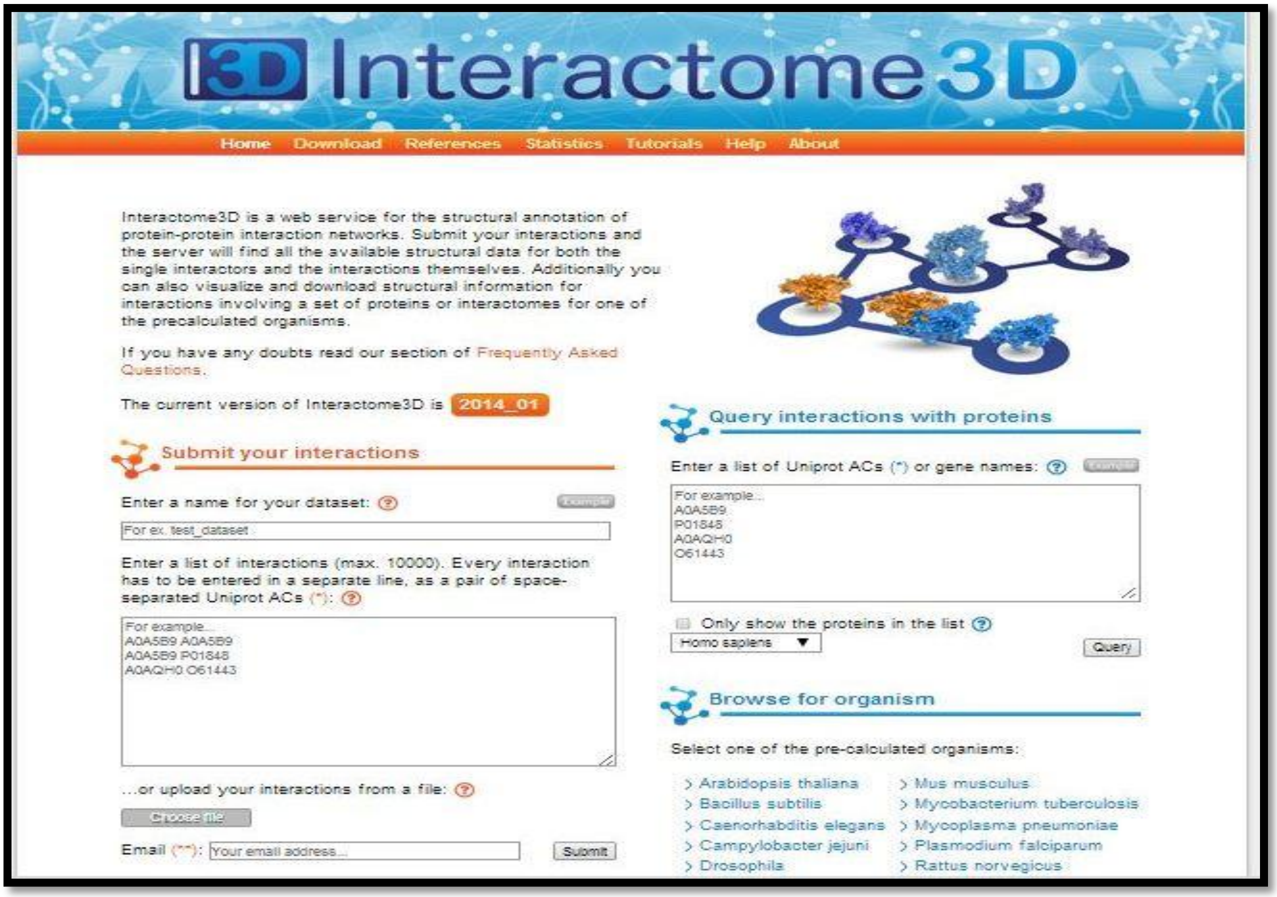


proteinA	proteinB	pfamA	seq_start_A	seq_end_A	pfamB	seq_start_B	seq_end_B
2	354	PF07677	1376	1463	PF00089	25	253
2	1990	PF07677	1376	1463	PF00089	19	251
12	354	PF00079	51	420	PF00089	25	253
12	1511	PF00079	51	420	PF00089	21	238
12	1990	PF00079	51	420	PF00089	19	251
12	1991	PF00079	51	420	PF00089	30	242
12	3817	PF00079	51	420	PF00089	25	253
12	11330	PF00079	51	420	PF00089	30	262
15	7534	PF00583	84	174	PF00244	3	236
25	25	PF00018	67	113	PF00018	67	113
25	25	PF00018	67	113	PF00017	127	202
25	25	PF00018	67	113	PF07714	242	493
25	25	PF00017	127	202	PF00018	67	113
25	25	PF00017	127	202	PF00017	127	202
25	25	PF00017	127	202	PF07714	242	493
25	25	PF07714	242	493	PF00018	67	113
25	25	PF07714	242	493	PF00017	127	202
25	25	PF07714	242	493	PF07714	242	493
25	613	PF00018	67	113	PF00621	502	690
25	613	PF00018	67	113	PF00169	702	866
25	695	PF00018	67	113	PF00018	220	266
25	695	PF00018	67	113	PF00017	281	362
25	695	PF00018	67	113	PF00169	4	133

Table 2: An excel sheet representing details of human structural interaction network.

4.4 **Structural interaction network:**

Once the structurally interacting partners have been obtained, a structure interaction network was formed by using Interactome3D (Mosca *et al.*, 2013). It is a web service for the structural annotation of protein-protein interaction networks. Interacrtome3D provides structural details at atomic resolution for over 12.000 protein-protein interaction in eight model organisms ranging from *Escherichia coli* to yeast to human (Figure 13). This web service is fully automated computational approach, which handles two types of input data: a set of interactions provided by the user or a list of organism for modeling of their entire protein-protein interactome. It works by collecting all the important information about the protein and its binary interactions and finds out any experimentally verified structure in PDB (Protein Data Bank), or else does homology modeling by selecting the best template itself. The results obtained are categorized into three: complete experimental structure i.e. covering more than 80% of the length of the protein with 100% sequence identity, homology models created by Modeller with more than 80% coverage and partial models or structures in which fragments are grouped together to cover maximum possible length of the protein (Mosca *et al.*, 2013).



Interactome3D is a web service for the structural annotation of protein-protein interaction networks. Submit your interactions and the server will find all the available structural data for both the single interactors and the interactions themselves. Additionally you can also visualize and download structural information for interactions involving a set of proteins or interactomes for one of the precalculated organisms.

If you have any doubts read our section of Frequently Asked Questions.

The current version of Interactome3D is 2014_01

Submit your interactions

Enter a name for your dataset: ? Example

For ex. test_dataset

Enter a list of interactions (max. 10000). Every interaction has to be entered in a separate line, as a pair of space-separated Uniprot ACs (*): ?

For example...
A0A5B9 A0A5B9
A0A5B9 P01848
A0AQH0 O61443

...or upload your interactions from a file: ?

Choose file

Email (**): Your email address... Submit

Query interactions with proteins

Enter a list of Uniprot ACs (*) or gene names: ? Example

For example...
A0A5B9
P01848
A0AQH0
O61443

☐ Only show the proteins in the list ?

Homo sapiens ▼ Query

Browse for organism

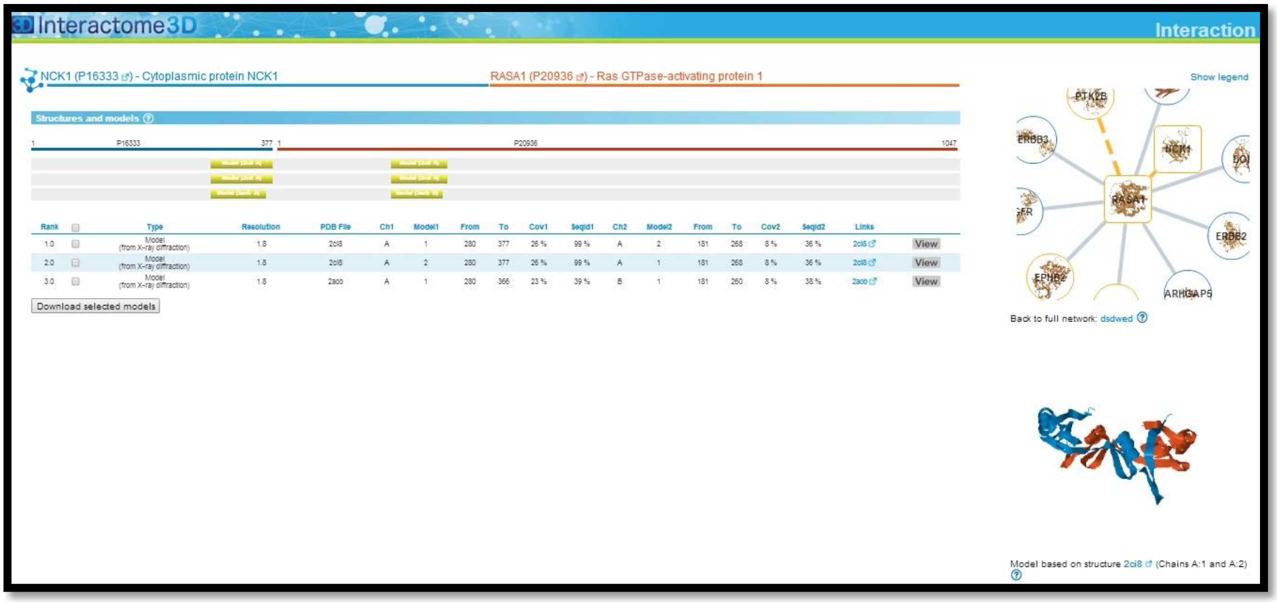
Select one of the pre-calculated organisms:

- Arabidopsis thaliana
- Bacillus subtilis
- Caenorhabditis elegans
- Campylobacter jejuni
- Drosophila
- Mus musculus
- Mycobacterium tuberculosis
- Mycoplasma pneumoniae
- Plasmodium falciparum
- Rattus norvegicus

Figure 13: Screenshot of Interactome3D web service.

4.5 Identification of interface structure using Intercatome3D

The network of the interacting proteins mapped with their structures provides essential information about individual proteins and their interacting partners. Structural protein-protein interaction network consist of nodes (proteins) and edges (interactions). For obtaining the interface structure, each interaction is analyzed and the edge between them depicts the structure of interface. By clicking on the link of two interacting proteins, PDB structures along with interacting chains were obtained (Figure 14).



Interactome3D

Interaction

NCK1 (P16333) - Cytoplasmic protein NCK1

RASA1 (P20936) - Ras GTPase-activating protein 1

Show legend

Structures and models

1 P16333 377 1 P20936 1047

Rank	□	Type	Resolution	PDB File	Ch1	Model1	From	To	Cov1	Seqid1	Ch2	Model2	From	To	Cov2	Seqid2	Links	
1.0	□	Model (from X-ray diffraction)	1.8	2ci8	A	1	280	377	26 %	99 %	A	2	181	268	8 %	36 %	2ci8	View
2.0	□	Model (from X-ray diffraction)	1.8	2ci8	A	2	280	377	26 %	99 %	A	1	181	268	8 %	36 %	2ci8	View
3.0	□	Model (from X-ray diffraction)	1.8	2aob	A	1	280	366	23 %	39 %	B	1	181	260	8 %	38 %	2aob	View

Download selected models

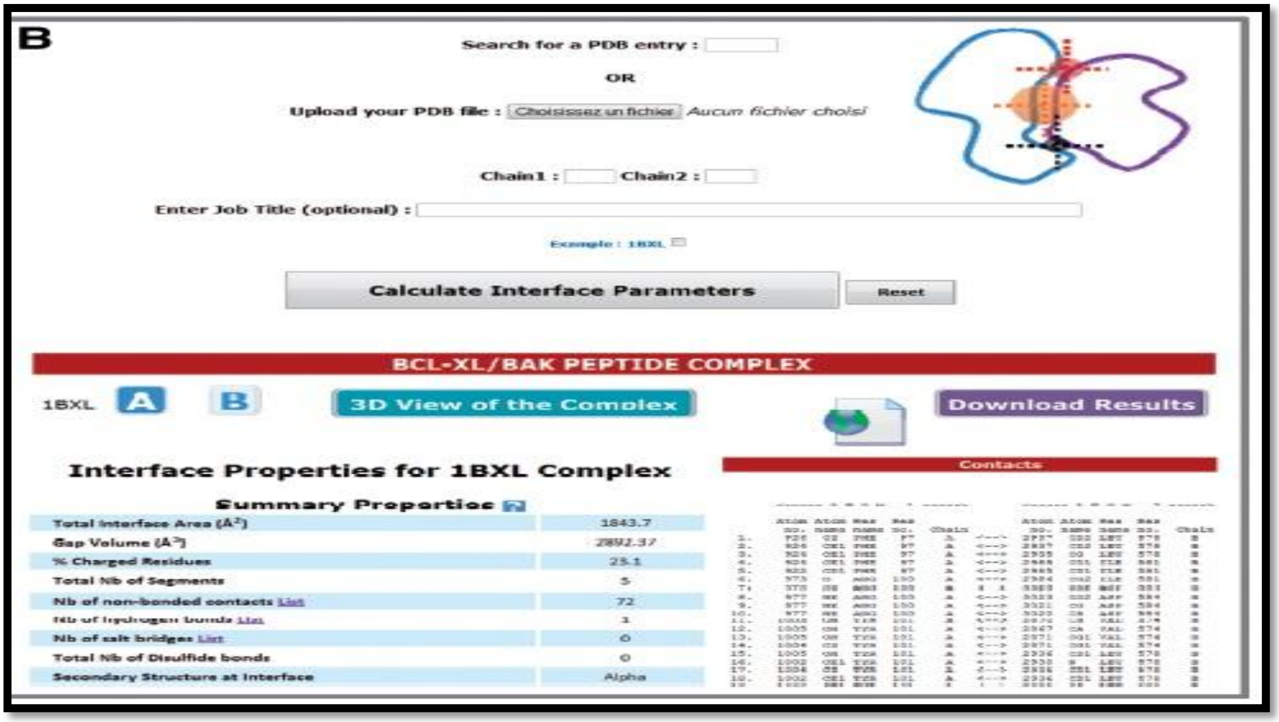
Back to full network: dsdwed

Model based on structure 2ci8 (Chains A:1 and A:2)

Figure 14: The image depicts the PDB ID for the complex between NCK1 and RASA1. The PDB structure is also shown along with the chains interacting.

4.6 Interface property analysis

For interface analysis, 2P2Iinspector (a protein–protein interface analysis tool) was used that invokes VMD, NACCESS and SURFNET. It is a complete tool that computes the interaction properties from 3D structure of interacting complexes. The various descriptors provided by this tool includes ASA (accessible Surface Area), Gap Volume, percentage charged residues, secondary structure contribution, number of hydrogen bonds, number of salt bridges and number of disulphide bonds. All these parameters are important in determining the specificity and the strength of the interface. Along with parameter details, this tool also provides the visualization of the protein-protein complex with a Jmol applet and also, the residue and atomic composition of the interface. Input can be a PDB code or the PDB file along with the chains involved in the complex (Basse *et al.*, 2013) (Figure 15).


B
Search for a PDB entry :
OR
Upload your PDB file : Choisissez un fichier Aucun fichier choisi
Chain1 : Chain2 :
Enter Job Title (optional) :
Example : 1BXL □
Calculate Interface Parameters
Reset
BCL-XL/BAK PEPTIDE COMPLEX
1BXL
A
B
3D View of the Complex
Download Results

Interface Properties for 1BXL Complex

Summary Properties

Total interface Area ($Å^2$)	1843.7
Gap Volume ($Å^3$)	2892.37
% Charged Residues	23.1
Total Nb of Segments	5
Nb of non-bonded contacts List	72
Nb of hydrogen bonds List	1
Nb of salt bridges List	0
Total Nb of Disulfide bonds	0
Secondary Structure at Interface	Alpha

Contacts

```
        Atom  Atom  Res   Res                   Atom  Atom  Res   Res
        no.   name  name  no.   Chain           no.   name  name  no.   Chain
  1.    925   CZ    PHE   97    A     <-->      2937  CD2   LEU   578   B
  2.    926   CE1   PHE   97    A     <-->      2937  CD2   LEU   578   B
  3.    926   CE1   PHE   97    A     <-->      2935  CG    LEU   578   B
  4.    926   CE1   PHE   97    A     <-->      2965  CD1   ILE   581   B
  5.    922   CD1   PHE   97    A     <-->      2965  CD1   ILE   581   B
  6.    973   O     ARG   100   A     <-->      2964  CG2   ILE   581   B
  7.    [illegible]
  8.    977   NE    ARG   100   A     <-->      3023  OD2   ASP   584   B
  9.    977   NE    ARG   100   A     <-->      3021  CG    ASP   584   B
 10.    977   NE    ARG   100   A     <-->      3022  [illegible]  ASP   584   B
 11.    [illegible]                             [illegible]  VAL   574   B
 12.    1005  OH    TYR   101   A     <-->      2967  CA    VAL   574   B
 13.    1005  OH    TYR   101   A     <-->      2971  CG1   VAL   574   B
 14.    1004  CZ    TYR   101   A     <-->      2971  CG1   VAL   574   B
 15.    1005  OH    TYR   101   A     <-->      2936  CD1   LEU   578   B
 16.    1002  CE1   TYR   101   A     <-->      2933  [illegible]  LEU   578   B
 17.    1004  CZ    TYR   101   A     <-->      2936  CD1   LEU   578   B
 18.    1002  CE1   TYR   101   A     <-->      2936  CD1   LEU   578   B
```

Figure 15: Screenshot of 2P2Iinspector tool.

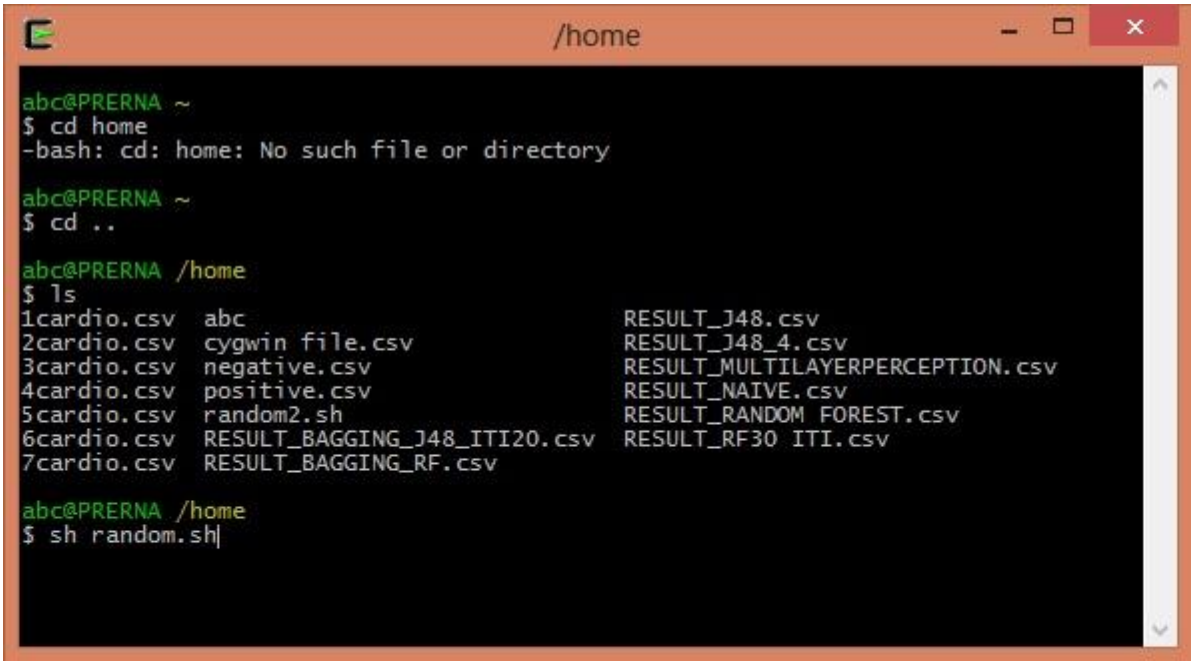
4.7 Classification analysis – algorithm and analysis

Comparison between the proteins involved in cardiovascular disorders and those diseases which are not known to be involved in diseases were made using interfacial features which describe the protein complexes. By using machine learning algorithm, an automatic classifier capable of identifying genes more likely to be involved in cardiovascular disorders based on the interfacial patterns was made (Xu *et al.,* 2006).

4.7.1 Processing of dataset before classification analysis.

Before running Weka (hall *et al.,* 2009), Cygwin tool was used (Red Hat, *et al.,*), which is a popular GNU development tools for Microsoft Windows. With cygwin it is possible to easily port many Unix programs without the need for extensive changes to the source code. This includes configuring and building most of the available GNU software (including the packages included with the Cygwin development tools themselves) as well as lots of BSD tools and packages (including OpenSSH). Even if the development tools are of little to no use to you, you may have interest in the many standard POSIX utilities provided with the package. They can be used from one of the provided UNIX shells like bash, tcsh or zsh, as well as from the standard Windows command shell if you have to for some sad reason. It is a free tool.

With cygwin, a series of random file were created by combining the positive and negative data set (Figure 16). These files were then used in Weka for classification and finding the prediction probability of test or unknown dataset.



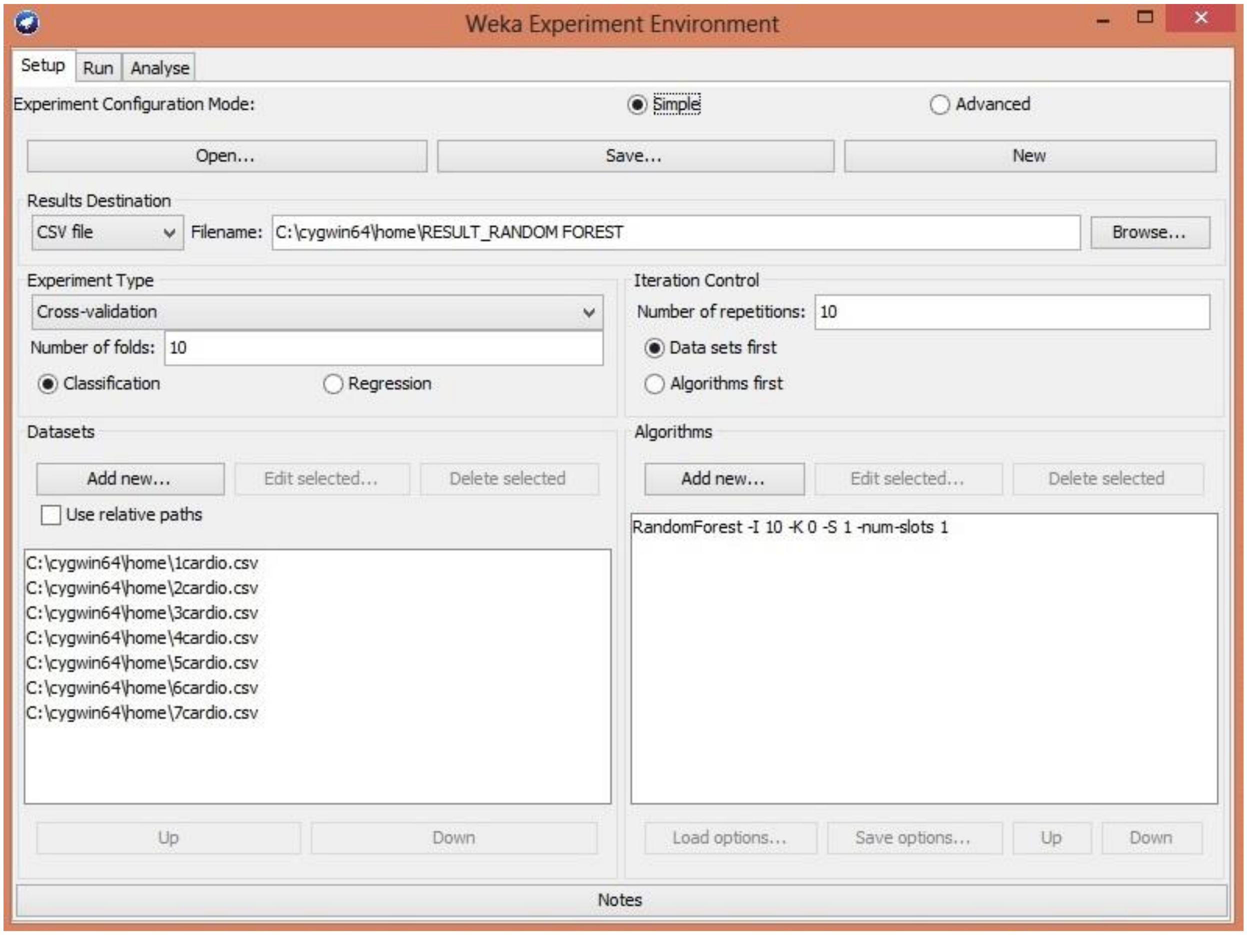
```
abc@PRERNA ~
$ cd home
-bash: cd: home: No such file or directory

abc@PRERNA ~
$ cd ..

abc@PRERNA /home
$ ls
1cardio.csv  abc                           RESULT_J48.csv
2cardio.csv  cygwin file.csv               RESULT_J48_4.csv
3cardio.csv  negative.csv                  RESULT_MULTILAYERPERCEPTION.csv
4cardio.csv  positive.csv                  RESULT_NAIVE.csv
5cardio.csv  random2.sh                    RESULT_RANDOM FOREST.csv
6cardio.csv  RESULT_BAGGING_J48_ITI20.csv  RESULT_RF30 ITI.csv
7cardio.csv  RESULT_BAGGING_RF.csv

abc@PRERNA /home
$ sh random.sh
```

Figure 16: Screenshot of Cygwin GUI.

4.7.2 <u>Model building and prediction</u>

Weka was used for model building and using the model for the prediction of the test set. All classification and analyses was performed on the Weka workbench (Bouckaert *et al.*, 2010). Weka (Waikato Environment for Knowledge Analysis) is popular open source Java based software that contains implementations of a diverse range of classification and clustering algorithms. It provides a simple GUI supporting the data from various sources and in different file formats. It has multiple algorithms (including that of regression, association rule mining, clustering, classification etc.) and pre-processing tools that allow comparison of different methods. The workbench is used for both supervised as well as unsupervised algorithms. The data visualization facilities help in easy access and analysis of results. We used Weka 3.7.11 for generating our models. The input files were converted in CSV form compatible with Weka. The models were built using Random Forest algorithm.

Experimenter was used to determine which algorithm gives the best and accurate result (Figure 17). Various algorithms such as Naïve Bayes, Random Forest, J48, and Multilayer Perceptron were used. Parameters such as true positive, true negative, false positive, false negative, precision and recall were used to determine the best algorithm. Random forest with iteration 30 was used to generate the model for prediction.



Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination
CSV file
Filename: C:\cygwin64\home\RESULT_RANDOM FOREST

Experiment Type
Cross-validation
Number of folds: 10
● Classification ○ Regression

Iteration Control
Number of repetitions: 10
● Data sets first
○ Algorithms first

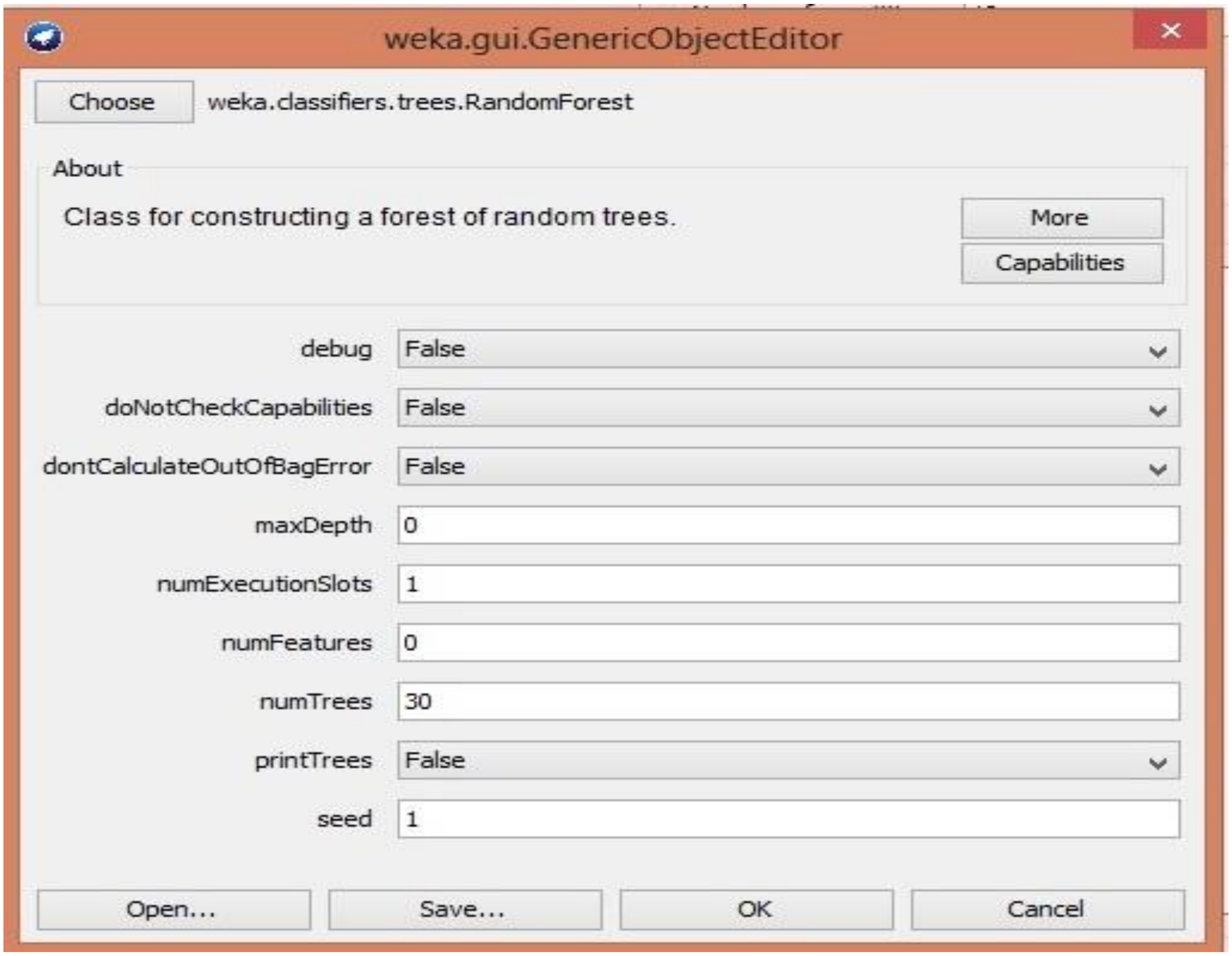
Datasets
☐ Use relative paths

C:\cygwin64\home\1cardio.csv
C:\cygwin64\home\2cardio.csv
C:\cygwin64\home\3cardio.csv
C:\cygwin64\home\4cardio.csv
C:\cygwin64\home\5cardio.csv
C:\cygwin64\home\6cardio.csv
C:\cygwin64\home\7cardio.csv

Algorithms
RandomForest -I 10 -K 0 -S 1 -num-slots 1

Figure 17: Weka experimenter to select the classifier algorithm.

In Random Forest, the numbers of trees were increased from 10 to 30, to obtain more precise results (Figure 18). Better results are obtained by using more tress as Random Forest takes average over many trees, but this should not exceed a particular tree otherwise prediction performance decreases.



weka.classifiers.trees.RandomForest

About

Class for constructing a forest of random trees.

debug False

doNotCheckCapabilities False

dontCalculateOutOfBagError False

maxDepth 0

numExecutionSlots 1

numFeatures 0

numTrees 30

printTrees False

seed 1

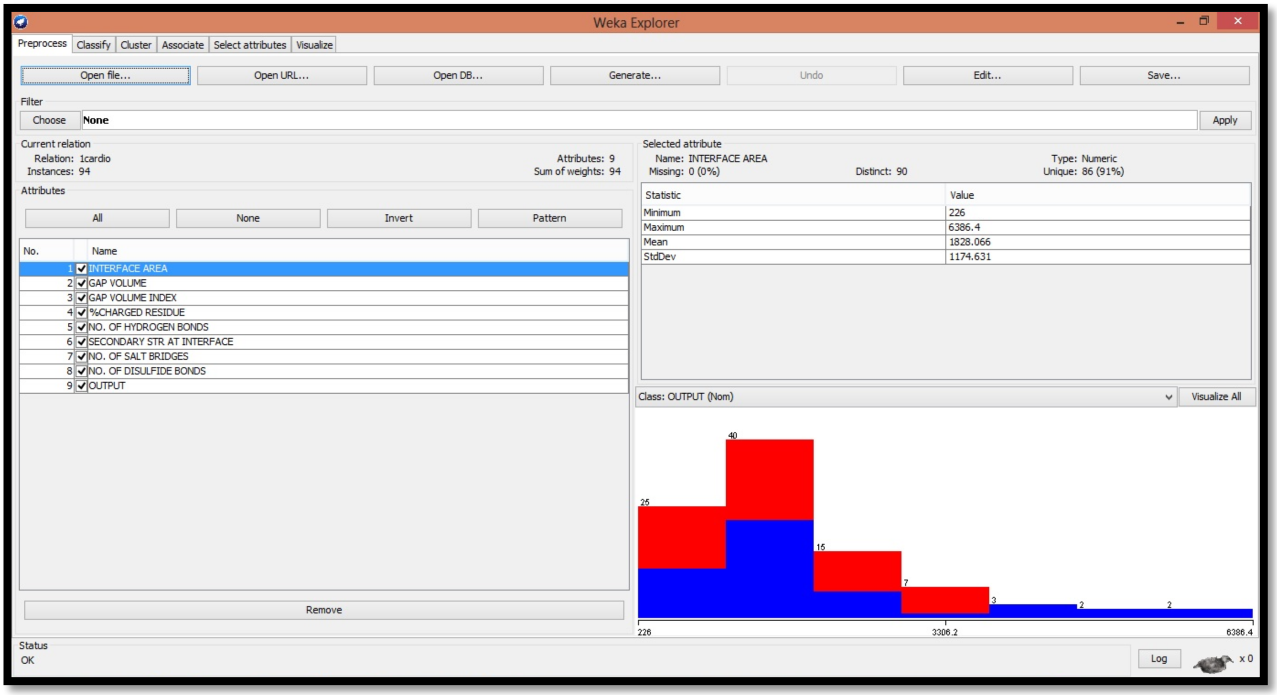
Figure 18: Random Forest sheet showing number of trees as 30.

Once the algorithm is decided, Explorer is used to form a model based on Random Forest classification algorithm and predictions are made using this model.

4.7.2.1 Model generation

- Weka explorer was opened and train dataset was uploaded in the pre-process tab.
- All the attributes were selected.
- From the classify tab, _Choose ‘then _trees’ then _Random Forest’ was selected.
- Random Forest was clicked and number of trees increased to 30.
- Cross validation value was set to 10.

- Build Model was clicked (Figure 19).



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose None | Apply

Current relation
Relation: 1cardio
Instances: 94
Attributes: 9
Sum of weights: 94

Attributes: All | None | Invert | Pattern

No.		Name
1	☑	INTERFACE AREA
2	☑	GAP VOLUME
3	☑	GAP VOLUME INDEX
4	☑	%CHARGED RESIDUE
5	☑	NO. OF HYDROGEN BONDS
6	☑	SECONDARY STR AT INTERFACE
7	☑	NO. OF SALT BRIDGES
8	☑	NO. OF DISULFIDE BONDS
9	☑	OUTPUT

Remove

Selected attribute
Name: INTERFACE AREA
Missing: 0 (0%)
Distinct: 90
Type: Numeric
Unique: 86 (91%)

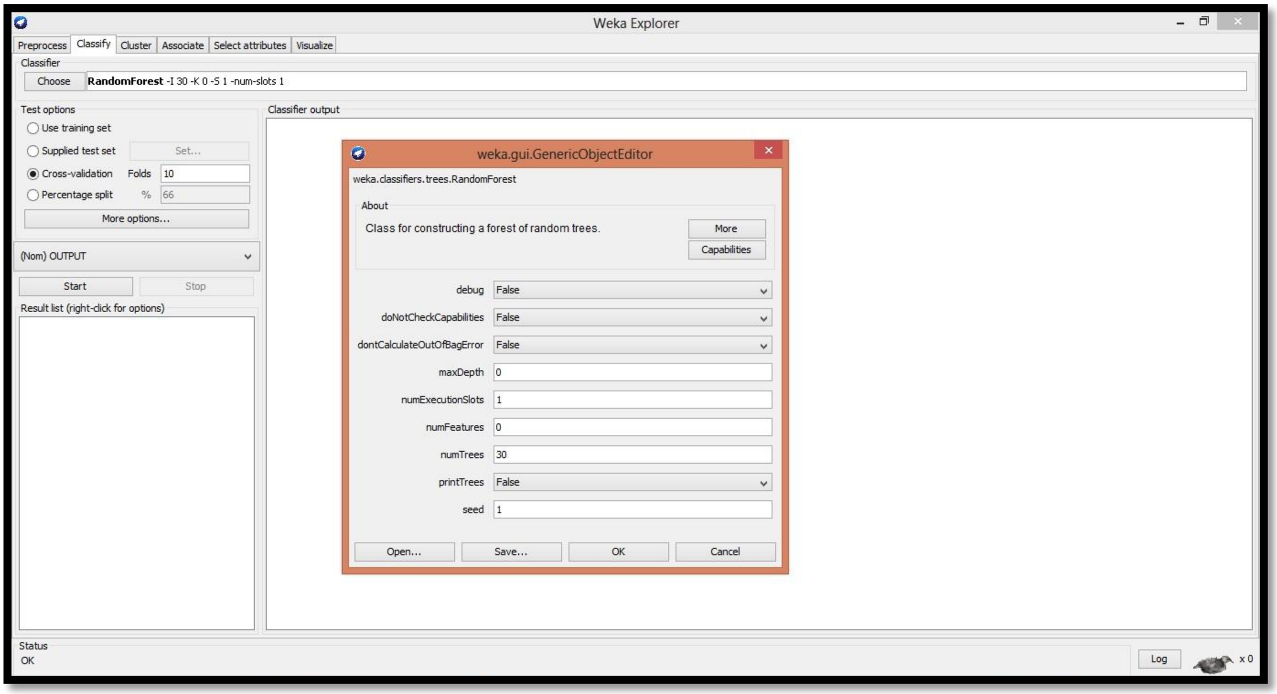
Statistic	Value
Minimum	226
Maximum	6386.4
Mean	1828.066
StdDev	1174.631

Class: OUTPUT (Nom) | Visualize All

Status
OK

Log | x 0

Figure 19: Weka explorer to generate model based on classifier algorithm.



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier
Choose **RandomForest** -I 30 -K 0 -S 1 -num-slots 1

Test options
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66
More options...

(Nom) OUTPUT

Start | Stop

Result list (right-click for options)

Classifier output

weka.gui.GenericObjectEditor

weka.classifiers.trees.RandomForest

About
Class for constructing a forest of random trees. | More | Capabilities

debug	False
doNotCheckCapabilities	False
dontCalculateOutOfBagError	False
maxDepth	0
numExecutionSlots	1
numFeatures	0
numTrees	30
printTrees	False
seed	1

Open... | Save... | OK | Cancel

Status
OK

Log | x 0

Figure 20: Weka explorer to cross validate the dataset.

This results in a table showing detailed accuracy by class and a 2×2 matrix.

4.7.2.2 Cross Validation

The technique is implied during training of the classifiers. K-fold cross validation is one of the most popularly used methods of cross-validation of the accuracy of a model. In k-fold cross validation, the entire data is divided into k subsets (folds) of equal sizes and training is done for (k-1) sets and testing is done on one set. The process is repeated k number of times so that each set is tested at least once. The process is shown in Figure 21. The average error rate is computed for all tests. We have used (k=10) or a 10-fold cross validation here. The resulting model from the cross-validation is applied to the test set.

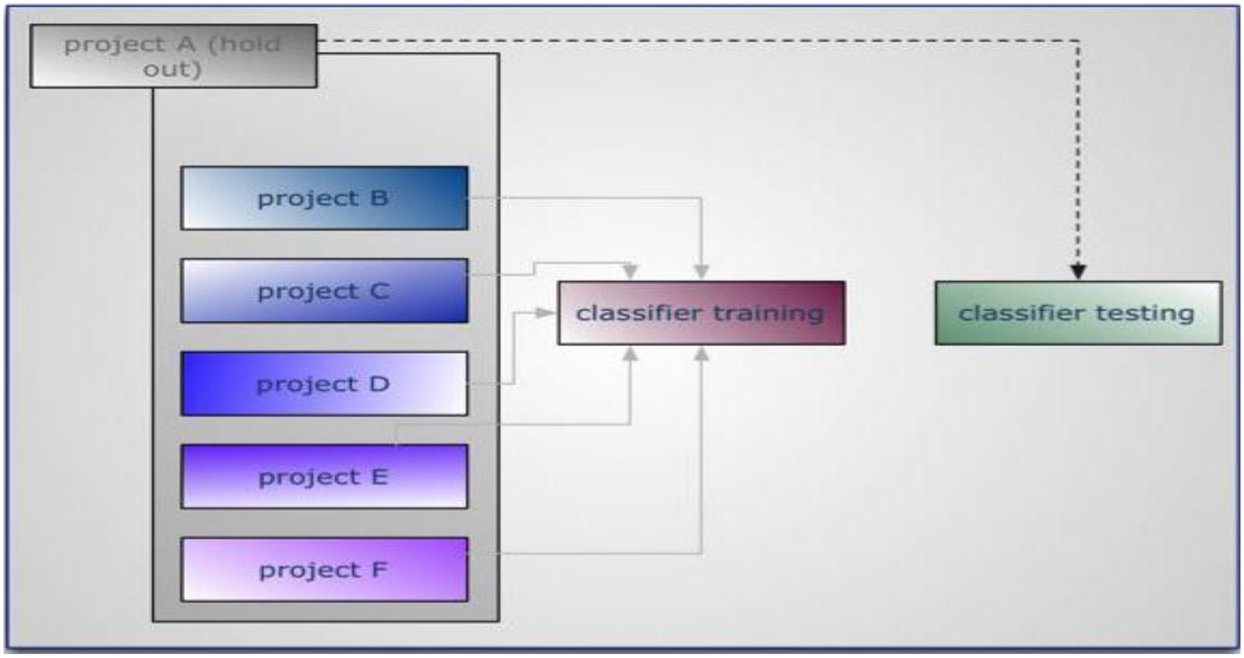


Figure 21: K-fold cross validation, one subset is used for training the model generated by rest of subsets as train set. The action is repeated in such a way that each subset becomes a test set at least once. The average of all is the final model.

4.7.2.3 Model performance evaluation

After the model generation, test sample is supplied and model is re-evaluated on current test set.

- In the classify tab of Weka itself, Click on _ 'Supplied Test Set' and the test set was uploaded by browsing.
- The generated model was clicked and 're-evaluate model on current test set' was chosen.

A CSV file was generated using prediction value, which was analyzed further.

4.8 DAVID (Database for Annotation, Visualization and Integrated discovery) analysis:

DAVID analysis of the predicted proteins was carried out to analyze their relationship with cardiovascular diseases. DAVID is a web-accessible program that provides integrated information about functional genomics annotations and their graphic summaries. It contains annotated gene or protein identifiers that share categorical information on protein domains, biochemical pathway membership, Gene Ontology, and etc. It includes functionally annotated data for genomes such as humans, rats, fly, mouse (Huang *et al.*, 2009).

5. RESULTS

5.1 Proteins associated with cardiovascular disorders

A total of 124 diseases under cardiovascular class of disorder were extracted from OMIM Morbid Map. Proteins associated with the specific diseases were identified from UniProtKB. An excel spreadsheet was created to associate proteins as well as genes with the respective disease.

For each protein, interactions which are structurally relevant i.e. in which both the interacting proteins have Pfam id, were considered and listed in excel sheet (Table 3).

DISEASE ID	DISEASE NAME	GENE SYMBOLS	UniProt ID	INTERACTIONS	OMIM GENE ID	CHROMOSOME NO.	CLASS
31	Acquired long QT syndrome, susceptibility to (3)	KCNH2, LQT2, HERG	Q12809	Q15669, P17612	152427	7q35-q36	Cardiovascular
130	Aortic aneurysm, ascending, and dissection (3)	FBN1, MFS1, WMS	P35555	O95967, Q9UBX5, P28300, P35556	134797	15q21.1	Cardiovascular
144	Arrhythmo geNAc right ventricular dysplasia 2, 600996 (3)	RYR2, VTSIP	Q92736	NO INTERACTION FOUND	180902	1q42.1-q43	Cardiovascular
144	Arrhythmo geNAc right ventricular dysplasia 8, 607450 (3)	DSP, KPPS2, PPKS2	NA	NI	125647	6p24	Cardiovascular
144	Arrhythmo geNIc right ventricular dysplasia, familial, 9,	PKP2, ARVD9	NA	NI	602861	12p11	Cardiovascular

163	Atherosclerosis, susceptibility to (3)	ALOX5	P09917	Q14019	152390	10q11.2	Cardiovascular
166	Atrial fibrillation, familial, 4, 607554 (3)	KCNE2, MIRP1, LQT6	NA	NI	603796	21q22.1	Cardiovascular
166	Atrial fibrillation, familial, 3, 607554 (3)	KCNQ1, KCNI9, LQT1, KVLQT1	P51787	P15382	607542	11p15.5	Cardiovascular
166	Atrial septal defect-2, 607941 (3)	GATA4	P43694	Q9Y2Y9	600576	8p23.1-p22	Cardiovascular
166	Atrial septal defect 3 (3)	MYH6, ASD3, MYHCA	NA	NI	160710	14q12	Cardiovascular
166	Atrial septal defect with atrioventricular conduction defects, 108900 (3)	NKX2E, CSX	NA	NI	600584	5q34	Cardiovascular
168	Atrioventricular block, idiopathic second-degree (3)	NKX2E, CSX	NA	NI	600584	5q34	Cardiovascular
168	Atrioventricular septal defect, 600309 (3)	GJA1, CX43, ODDD, SDTY3, ODOD	P17302	Q07157, Q02487-1	121014	6q21-q23.2	Cardiovascular

Table 3: List of all diseases associated with cardiovascular class of disorders along with the proteins and the interacting partners. NA – protein id was not available. NI – structural interactions were not found.

There are a total of 124 entries. In the table above, both the proteins and their respective interactions are listed. Colored rows indicate the diseases for which respective protein has no structural interaction. Using this table, structural interaction network was created.

5.2 Protein-Protein Structural Interaction Network:

A list of proteins and the interacting partners were provided to the web service Intercatome3D, which resulted in network along with structures, where proteins are denoted by nodes and link between interacting partners as edges.

Positive set:

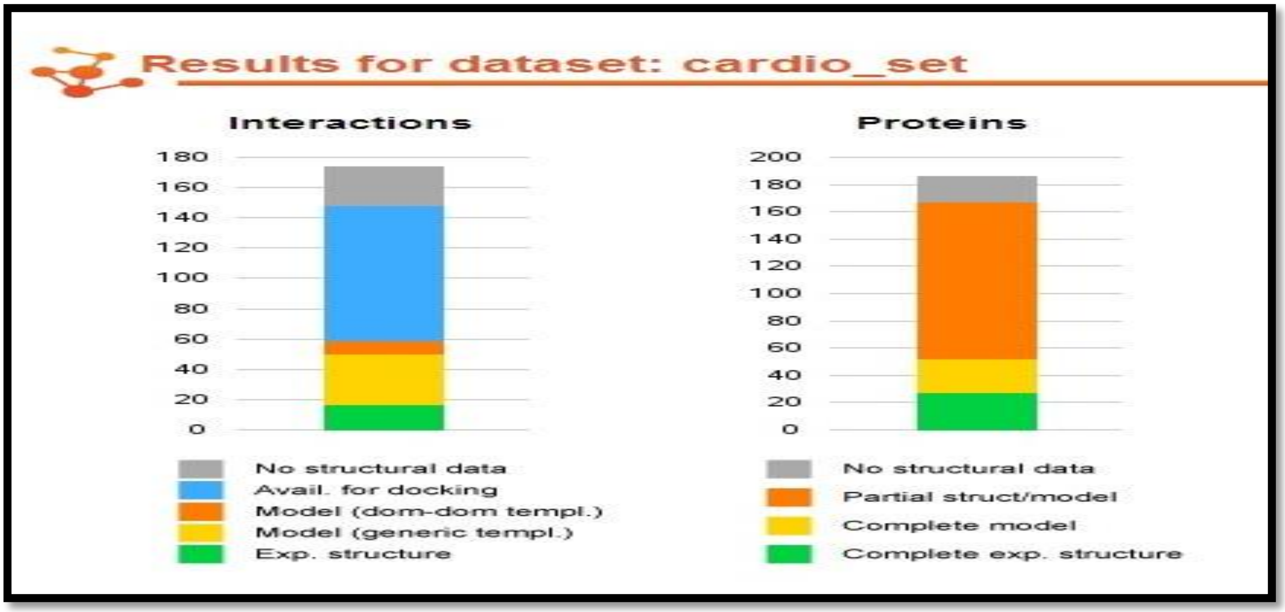
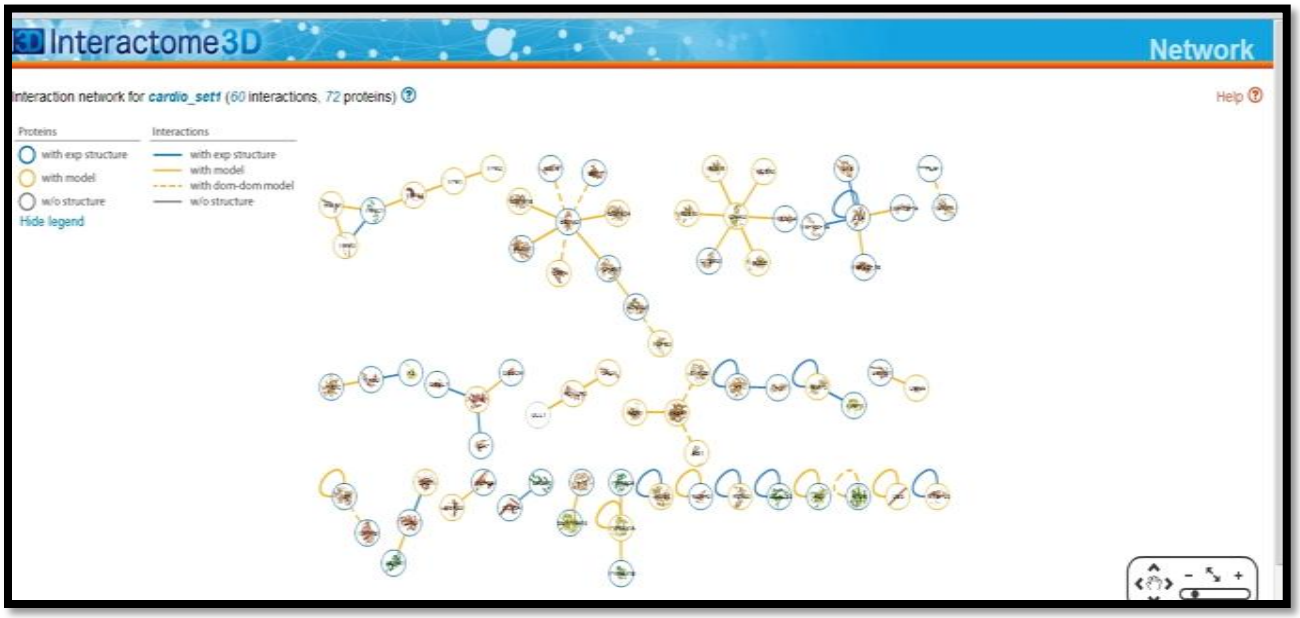


Figure 22: Bar graphs showing the proteins and their interactions with color indications of experimental structures, models, and no structural information.



Interaction network for *cardio_set1* (60 interactions, 72 proteins)

Figure 23: Interaction network with the mapped structures. Colour legend is provided on the top left corner.

Negative set:

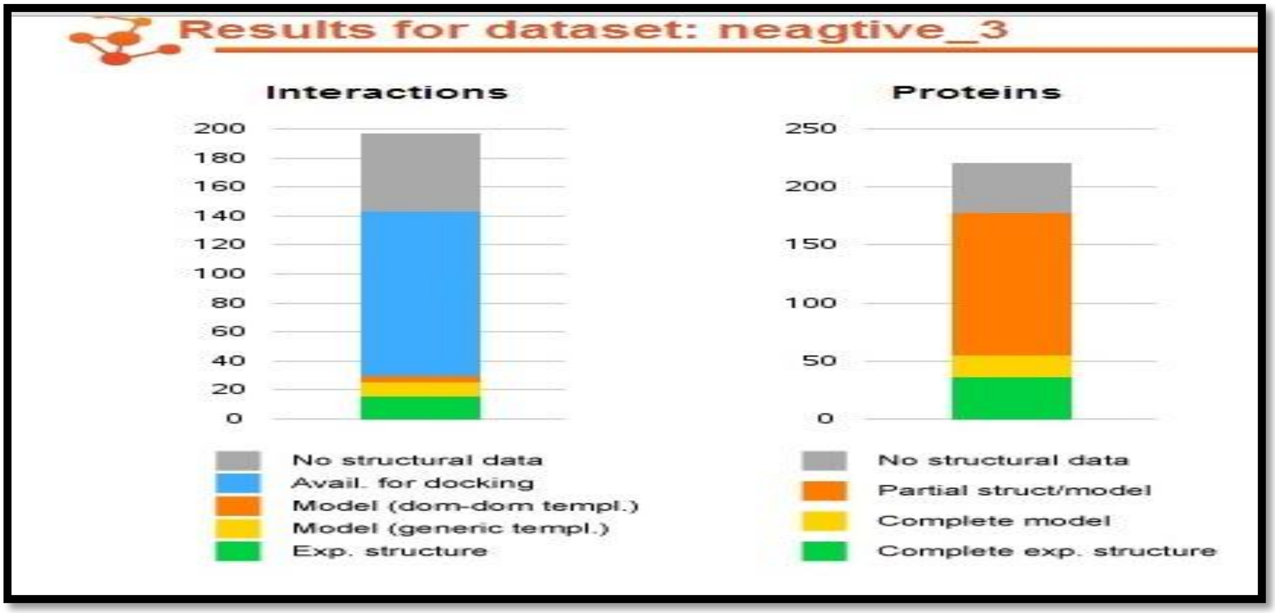
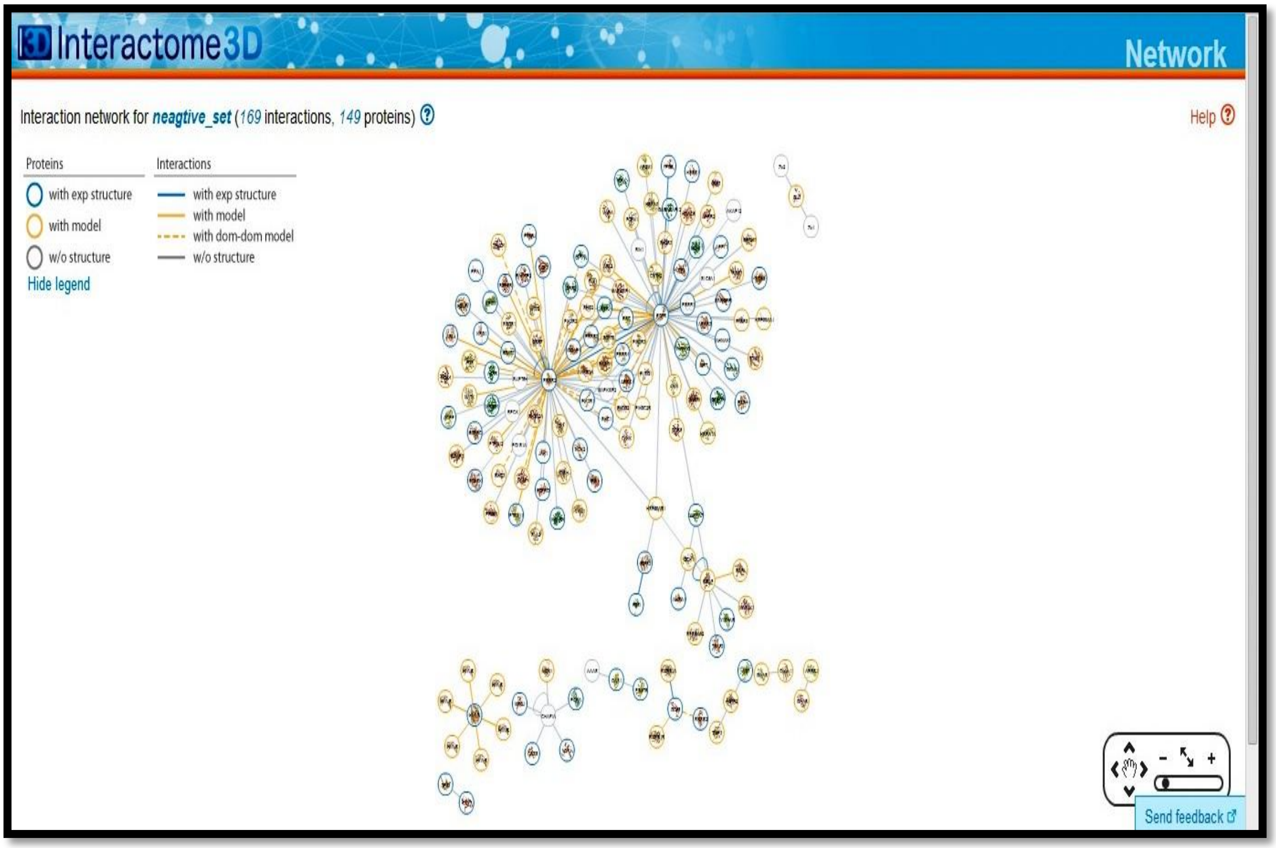


Figure 24: Bar graph for negative set.



Interaction network for *neagtive_set* (169 interactions, 149 proteins)

Figure 25: Structural interaction network for negative set.

Test set:

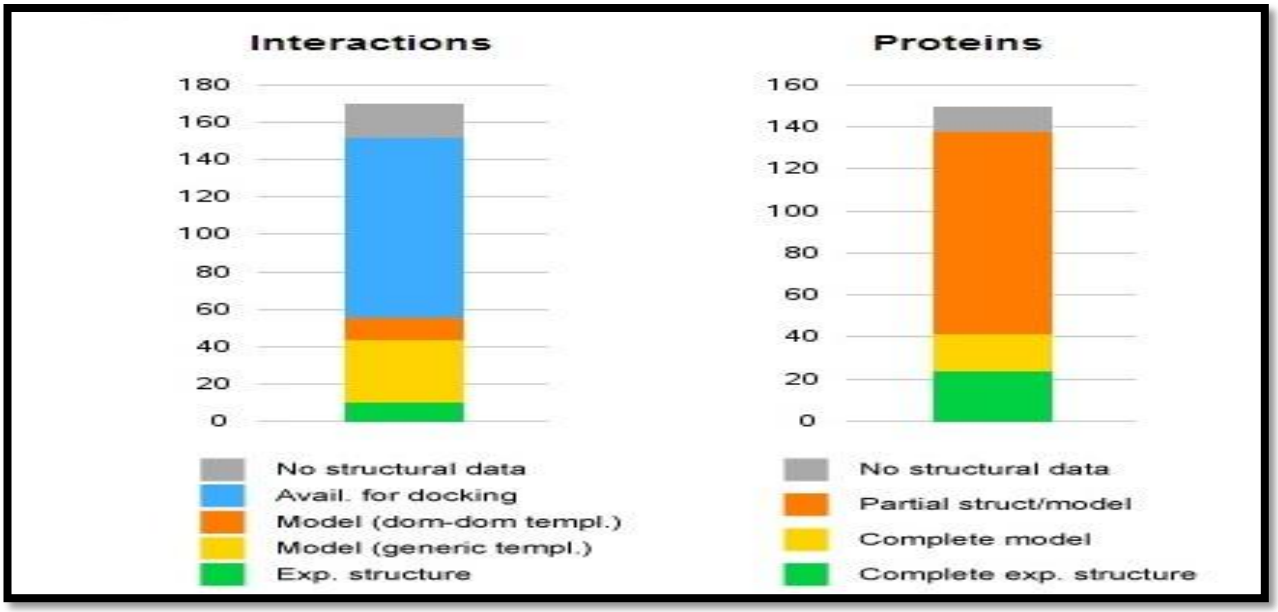


Figure 26: Bar graph for unknown set

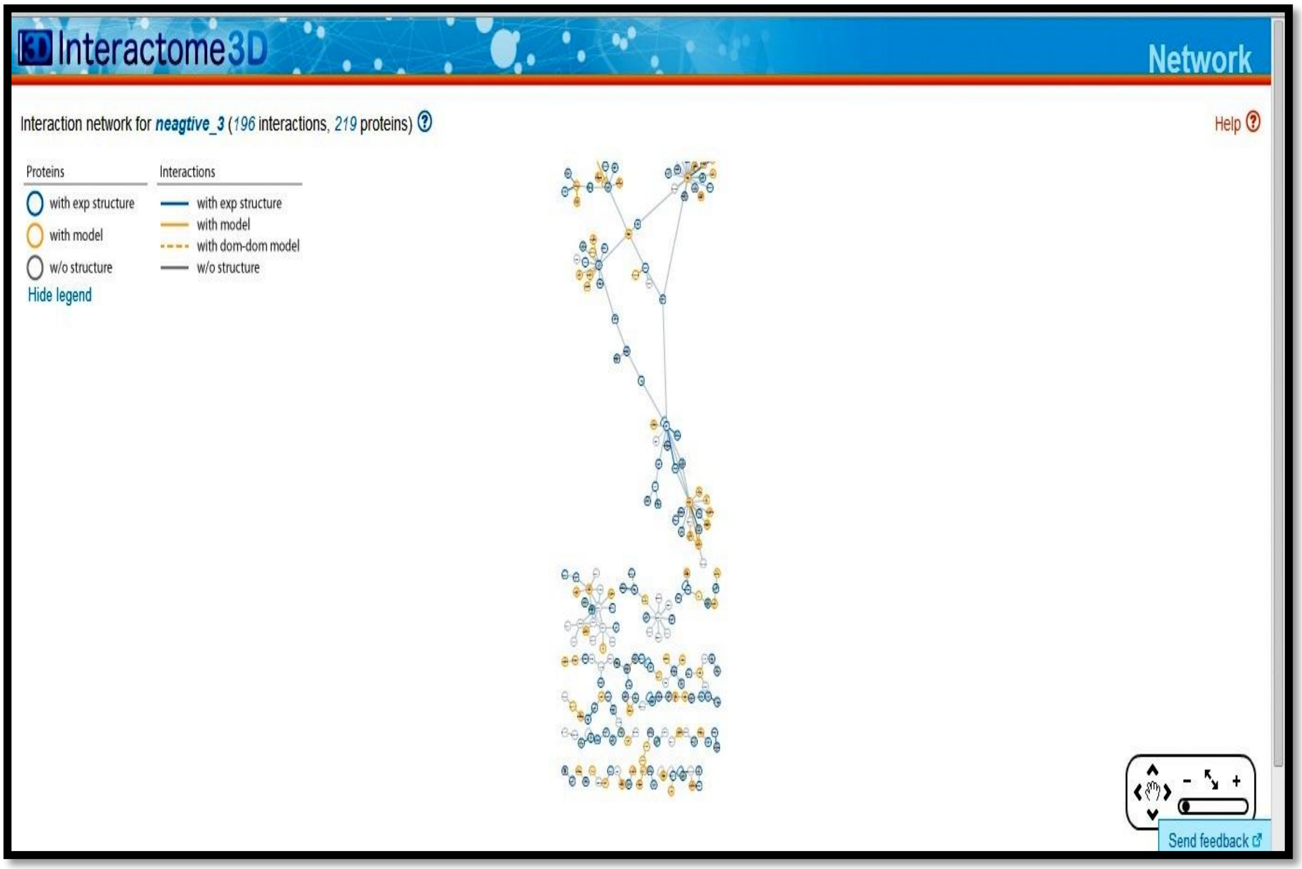

Interaction network for *neagtive_3* (*196* interactions, *219* proteins)

Figure 27: Structural Interaction network for unknown polymorphic proteins

As described in methodology, the link between individual interacting partners is clicked which opens in a new window showing the best three models for the proteins, and the PDB structure with the chains involved in interaction. This information about PDB structure and interacting chains was used to determine the interface parameters.

5.3 Deducing interface parameters using 2P2Iinspector

2P2Iinspector is a tool for determining the interface attributes. For example, for PDB id 1RY7, a detailed analysis of interface is given in the form of Figure 29.

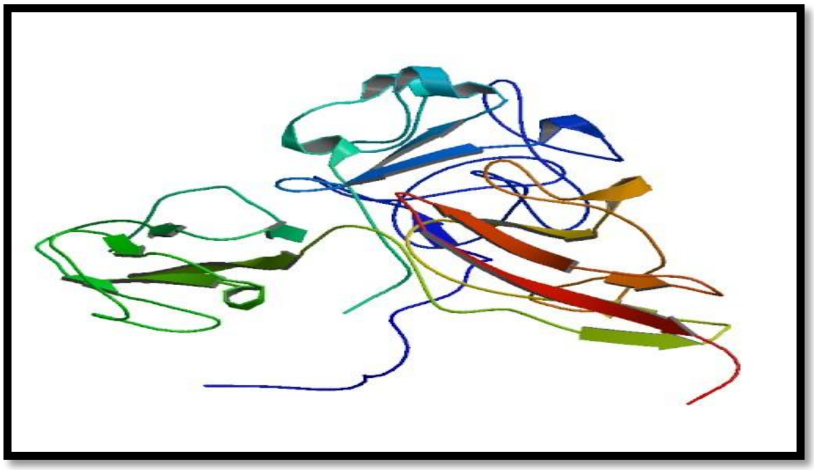
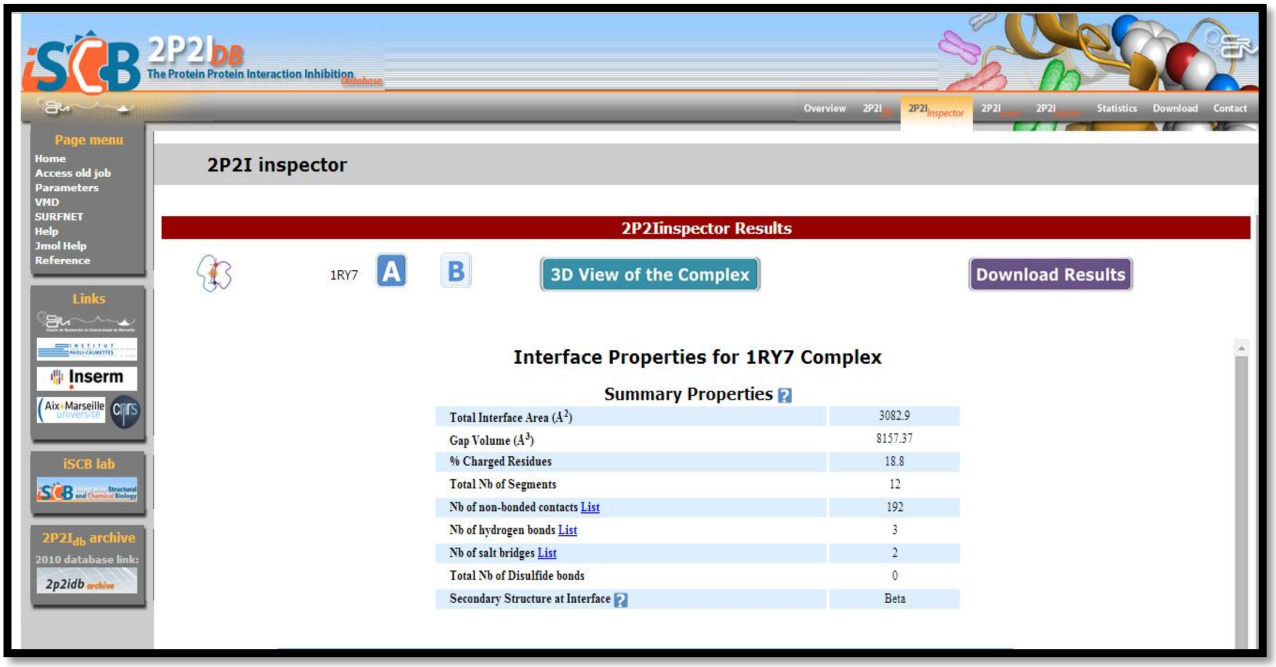


Figure 28: Biological Assembly Image for 1RY7. Crystal Structure of the 3 Ig form of FGFR3c in complex with FGF1. Protein chains are colored from the N-terminal to the C-terminal using a rainbow (spectral) color gradient



2P2I inspector

2P2Iinspector Results

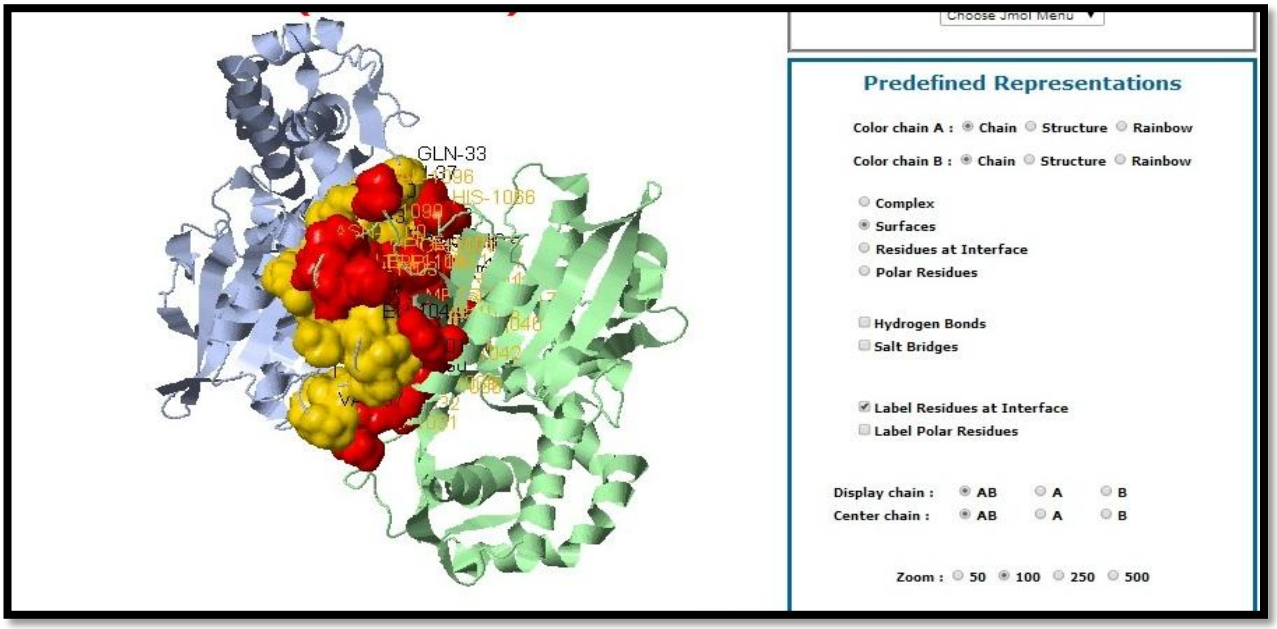
Interface Properties for 1RY7 Complex

Summary Properties ?

Total Interface Area ($Å^2$)	3082.9
Gap Volume ($Å^3$)	8157.37
% Charged Residues	18.8
Total Nb of Segments	12
Nb of non-bonded contacts List	192
Nb of hydrogen bonds List	3
Nb of salt bridges List	2
Total Nb of Disulfide bonds	0
Secondary Structure at Interface ?	Beta

Figure 29: Interface Properties

Along with the interface analysis between two protein chains, detailed analysis of single chain is also provided which can be used in various studies and elaborate structure visualization with various options to label interface residues, polar residues, hydrogen bonds and salt bridges (Figure 30).



Choose Jmol Menu ▼

Predefined Representations

Color chain A : ● Chain ○ Structure ○ Rainbow

Color chain B : ● Chain ○ Structure ○ Rainbow

○ Complex
● Surfaces
○ Residues at Interface
○ Polar Residues

□ Hydrogen Bonds
□ Salt Bridges

☑ Label Residues at Interface
□ Label Polar Residues

Display chain : ● AB ○ A ○ B
Center chain : ● AB ○ A ○ B

Zoom : ○ 50 ● 100 ○ 250 ○ 500

Figure 30: Jmol visualization of the interface showing interface residues.

The interfaces properties are summarized in excel spreadsheet (Table 4, 5, 6):

Training sample: Positive set

DISEASE NAME	PROTEIN A	PROTEIN B	PDB ID	ASA	GV	GVI	%CR	HB	SECSTRU	SB	DB
Myocardial infarction	LTA (P01374 link)	TNFRSF1A (P19438 link)	1tnr link (Chains A:1 and R:1)	1208.29	6149.25	5.089216993	40	3	BETA	0	0

	LTA (P01374 link)	TNFRSF1B (P20333 link)	3k51 link (Chains A:2 and B:2)	932.3	5575.5	5.980371125	33.3	1	Coil	1	0
	LTA (P01374 link)	LTA (P01374 link)	4mxv link (Chains B:1 and A:1)	1826.7	4839.75	2.649449828	21.4	3	BETA	0	0
	LTA (P01374 link)	LTB (Q06643 link)	4mxw link (Chains A:1 and D:1)	1828.1	4377.37	2.394491549	5.9	2	BETA	0	0
	LTA (P01374 link)	TNFRSF14 (Q92956 link) member 14	3alq link (Chains C:1 and S:1)	1212.2	6314.62	5.209222901	0	2	Beta	0	0

Table 4: Interface parameters for positive set. ASA- Accessible Surface Area, GV- Gap Volume, GVI- Gap Volume Index, %CR- charged residues, HB- hydrogen bonds, Sec stru- Secondary structure, SB – number of salt bridges, DB- number of disulphide bonds.

There were total 47 proteins in the list of whom the interface properties were calculated.

Negative set:

PROTEIN 1	PROTEIN 2	PDB ID	ASA	GV	GVI	%CR	HB	SEC STRU	SB	DB
FGF1 (P05230 link)	FGFR3 (P22607 link)	1ry7 link (Chains A:1 and B:1)	3082.9	8157.37	2.646005385	18.8	3	1	2	0
BRAF (P15056 link)	BRAF (P15056 link)	3ny5 link (Chains B:1 and A:1)	1940.7	3631.5	1.87123203	28.6	6	1	1	0
BRAF (P15056 link)	RAF1 (P04049 link)	4ehe link (Chains A:1 and B:1	2337.2	6706.12	2.869296594	25	4	0	0	0
GDF5 (P43026 link)	BMPR1A (P36894 link)	3qb4 link (Chains D:1 and C:1)	1812.1	4012.87	2.214485956	13.3	5	0	0	0
GDF5 (P43026 link)	BMPR1B (O00238 link)	3evs link (Chains C:1 and B:2)	743.6	3550.5	4.774744486	42.9	0	1	0	0

Table 5: Interface parameters for negative set. ASA- Accessible Surface Area, GV- Gap Volume, GVI- Gap Volume Index, %CR- charged residues, HB- hydrogen bonds, Sec stru- Secondary structure, SB – number of salt bridges, DB- number of disulphide bonds.

400 interacting proteins were considered in negatives set, which was used to form training set.

Test Set:

PROTEIN 1	PROTEIN 2	PDB ID	ASA	GV	GVI	%CR	HB	SEC STRU	SB	SB
AANAT	YWHAZ	1IB1, AE	2861.3	8100	2.830881068	43.3	11	ALPHA	3	0
A2M	CELA1	3HS0, AB	7475.3	15406.87	2.061037015	22.2	30	BETA	1	0
ABCB8	ABCB8	1G9X, AB	392.8	3813.75	9.709139511	33.3	0	COIL	0	0
ABCB8	ABCB8	1XF9, AB	1494.2	5764.5	3.85791728	41.7	0	COIL	1	0
ABCB8	ABCB8	2BBS, AB	437.4	12035.25	27.5154321	66.7	0	ALPHA	0	0
ABCB8	ABCB8	2IXF, AB	1666.3	11431.12	6.86018124	11.8	5	ALPHA	0	0
ABCB8	ABCB8	3C41, JK	1738.2	10246.5	5.894891267	26.7	1	COIL	0	0
ABCB8	ABCB8	3G61, AB	2993.3	12544.87	4.190983196	22.7	0	ALPHA	2	0
ABCB8	ABCB8	2HYD, AB	14303	30543.75	2.135478571	19.8	12	ALPHA	4	0
ABCD1	ABCD1	1G29, AB	2855	10337.62	3.620882662	33.3	5	ALPHA	2	0

Table 6: Interface properties for test set. ASA- Accessible Surface Area, GV- Gap Volume, GVI- Gap Volume Index, %CR- charged residues, HB- hydrogen bonds, Sec stru- Secondary structure, SB – number of salt bridges, DB- number of disulphide bonds.

600 proteins from a total of 1500 polymorphic proteins were determined to have structures and their interface properties are listed in the table above.

The training set included both the positive set (i.e. cardiovascular disease associated proteins with their interacting partners) and negative set (i.e. proteins and their interacting partners associated with diseases other than cardiovascular disorders) files and the test set has a list of polymorphic proteins not associated with any disease. A total of 8 descriptors were defined namely Accessible Surface Area, Gap Volume, Gap Volume Index, % charged residues, number of hydrogen bonds, number of salt bridges, number of disulphide bonds and secondary structure at interface. These files were used as input in Weka.

5.4 Classification and Prediction analysis

The descriptors files containing the parameters and the protein ids of both proteins along with the output were converted into CSV (Comma Separated File) which is required for Weka format. Weka experimenter was used to find the algorithm with the best output. A set of famous algorithms such as Naïve Bayes, Random Forest, J48, Random Forest with iteration 30, Bagging with Random Forest and Multilayer Perceptron were run on the dataset and comparisons were done between them (Table 7). Various statistical values were used in determining the optimum classifier for our training sample which is:

True Positive: Proteins associated with cardiovascular disease correctly predicted as cardiovascular disorders associated.

True Negative: Proteins associated with disease other than cardiovascular correctly predicted as not related to cardiovascular disorder.

False Positive: Non cardiovascular-disease associated protein predicted as cardiovascular-disease associated.

False Negative: Cardiovascular-disease associated protein predicted as non-cardiovascular-disease related.

Precision: Percentage of correctly predicted cardiovascular-disease protein over all predicted cardiovascular-disease associated disease protein.

Recall: Percentage of correctly predicted cardiovascular-disease protein over all cardiovascular-disease associated disease protein.

Classifier	True Positive Rate	False Positive Rate	True Negative Rate	False Negative Rate	Precision	Recall
Random Forest_10	0.77	0.4	0.59	0.23	0.67	0.77
Naïve Bayes	0.6	0.41	0.59	0.4	0.6	0.56
J48	0.74	0.39	0.61	0.26	0.67	0.74
Random Forest_30	0.77	0.38	0.62	0.23	0.7	0.77
Multilayer Perceptron	0.71	0.41	0.6	0.28	0.65	0.71
Bagging_Random Forest	0.75	0.37	0.63	0.25	0.68	0.6

Table 7: Comparisons between different classifiers.

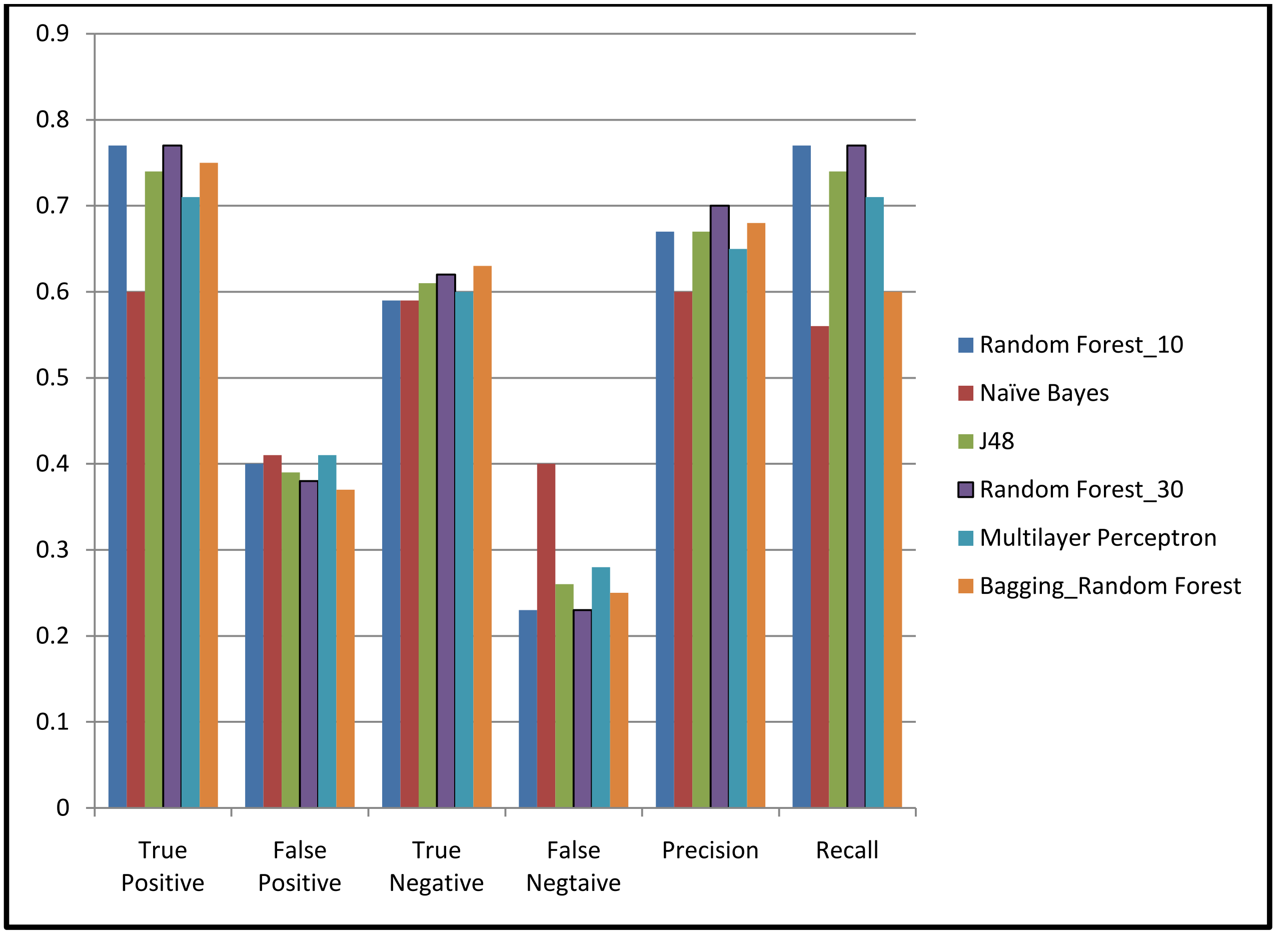
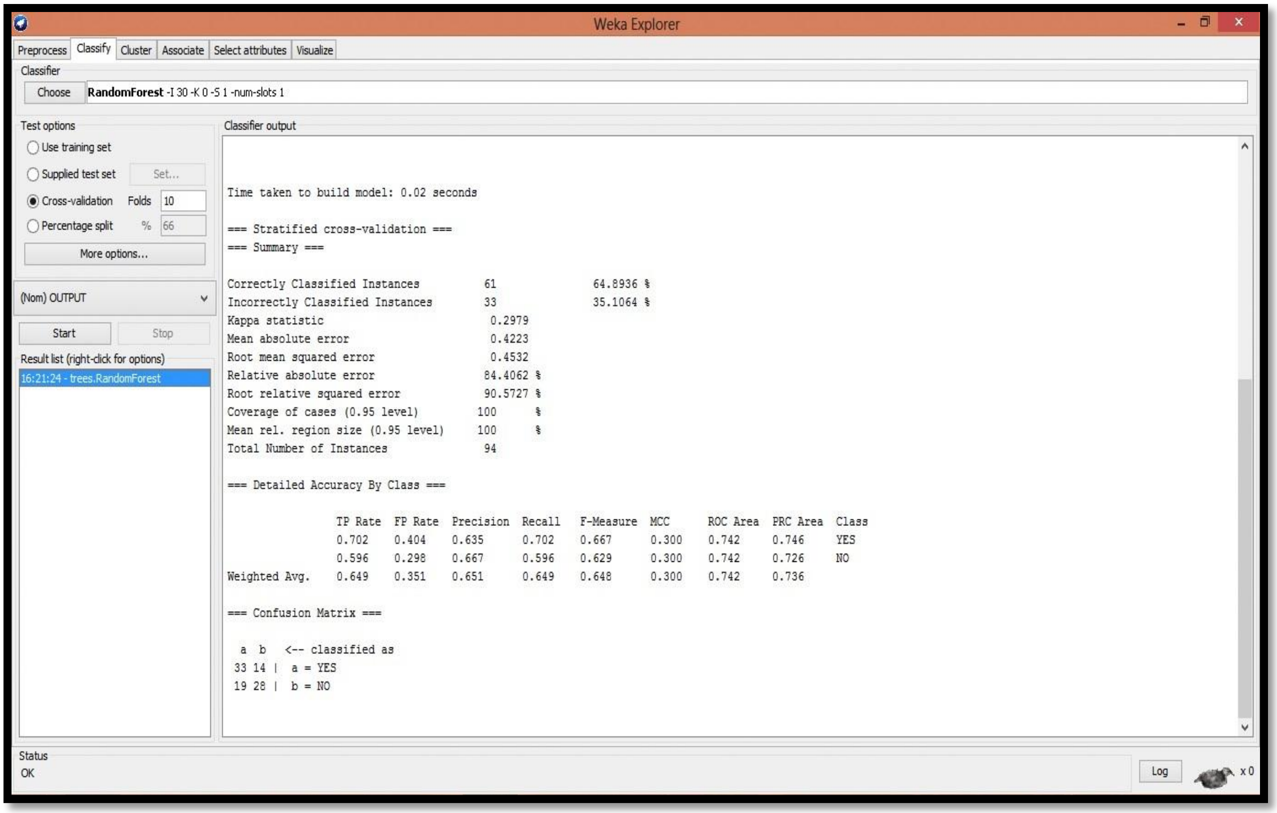


Figure 31: Bar graph with different classifiers and the statistical values.

In Figure 31, true positive rate, true negative rate, precision and recall were highest for Random Forest with iterations 30 (i.e. the number of trees were increased from 10 to 30). In the columns of false positive and false negative, Random Forest_30 scores the least. Thus, Random Forest_30 is the optimum classifier for the model building.

5.4.1 Evaluation of Model generated

Explorer in Weka was used for generating model with the training sample using Random Forest_30 (Figure 32).



Classifier

RandomForest -I 30 -K 0 -S 1 -num-slots 1

● Cross-validation Folds 10

(Nom) OUTPUT

16:21:24 - trees.RandomForest

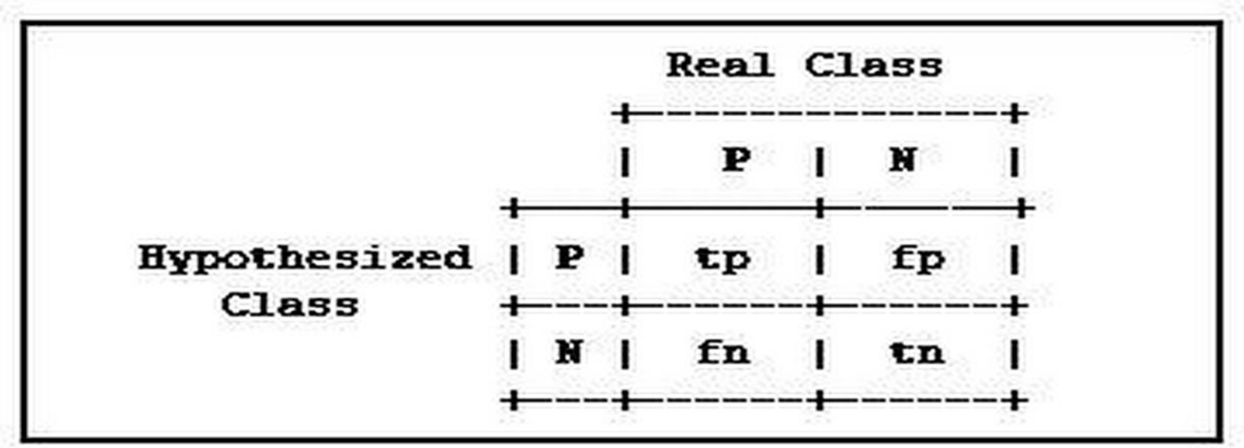
Status

OK

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          61               64.8936 %
Incorrectly Classified Instances        33               35.1064 %
Kappa statistic                          0.2979
Mean absolute error                      0.4223
Root mean squared error                  0.4532
Relative absolute error                 84.4062 %
Root relative squared error             90.5727 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.702    0.404    0.635      0.702    0.667      0.300    0.742     0.746     YES
                 0.596    0.298    0.667      0.596    0.629      0.300    0.742     0.726     NO
Weighted Avg.    0.649    0.351    0.651      0.649    0.648      0.300    0.742     0.736

=== Confusion Matrix ===

  a  b   <-- classified as
 33 14 |  a = YES
 19 28 |  b = NO
```

Figure 32: Cross validation results of model generated using Weka.

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.



		Real Class	
		P	N
Hypothesized Class	P	tp	fp
	N	fn	tn

Where tp is the number of 'positive' examples classified as 'positive',
tn is the number of 'negative' examples classified as 'negative',
fp is the number of 'negative' examples classified as 'positive',
fn is the number of 'positive' examples classified as 'negative',

Accuracy

It simply measures the ratio of the test examples a system classifies correctly. So, simply

$$\text{Accuracy} = \frac{\text{Correctly classified examples}}{\text{All examples in the sample}}$$

For confusion matrix that would be equal to

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

The true positive rate (TP) or **sensitivity** is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{True positive rate} = \frac{tp}{tp + fn}$$

The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$\text{False positive rate} = \frac{fp}{fp + tn}$$

The true negative rate (TN) or **specificity** is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$\text{True negative rate} = \frac{tn}{tn + fp}$$

The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$\text{False negative rate} = \frac{fn}{fn + tp}$$

Precision and Recall

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{recall} = \frac{tp}{tp + fn}$$

Precision and recall values are not symmetrical.

Precision corresponds to ratio of correctness in the examples classified as positive, recall measures ratio of examples classified as positive among all positive examples. In a sense, remeasures precision measures fidelity, while recall measures completeness. In most cases, tuning machine learning system to improve one of these measures result in a drop in the other. In some cases high precision may be more important, and in some cases high recall may be more important. However, in most cases, we aim at improving both values. The combination of these values are called f-score, and in most common form, it is the harmonic mean of both

$$\text{f-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Receiver Operator Characteristics (ROC) Analysis

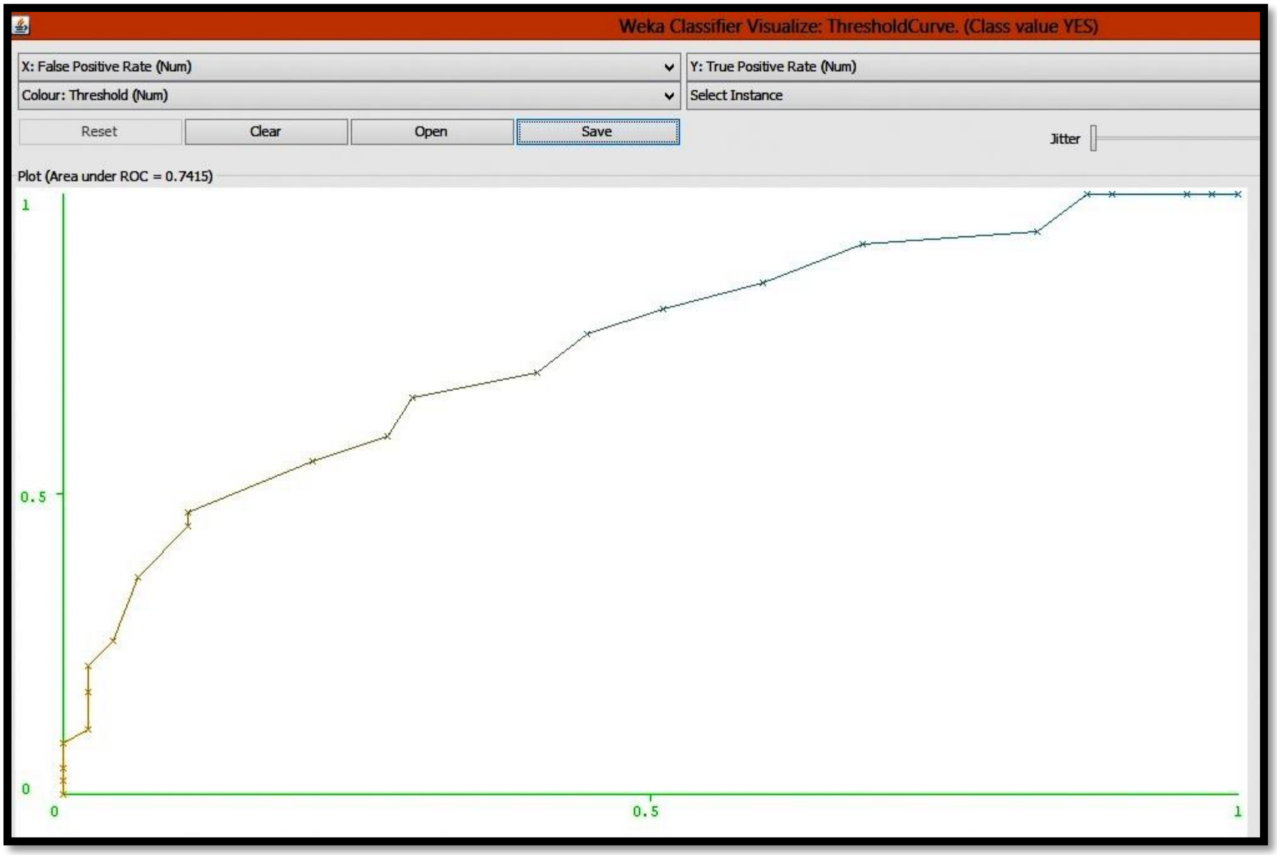
ROC analysis is done for evaluation and model selection. In ROC analysis we plot tp ratio (true positives divided by all positives) against fp ratio (false positives divided by all negatives).

In ROC graph, the ideal classifier appears at the upper left corner (tpr=1, fpr=0). An ROC curve demonstrates several things:

1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
4. The area under the curve is a measure of accuracy of prediction.

For Random Forest_30, the area under curve was 0.74 (Figure 33).



Weka Classifier Visualize: ThresholdCurve. (Class value YES)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7415)

Figure 33: ROC curve for Random Forest_30.

5.4.2 Prediction Probability for Test set:

Now, once the model has been evaluated, predictions were made using the test sample to predict potential cardiovascular-disorder related proteins from a set of polymorphic proteins. Further analysis of these can provide insight into unknown molecular mechanism involved in the occurrence of particular disease.

List of all the polymorphic protein along with their prediction probability was obtained. Threshold value of >0.8 was set for considering those proteins which could be potential proteins associated with cardiovascular-disease.

A total of 42 proteins out of 600 polymorphic proteins had prediction probability more than 0.8. The relevant association of these unknown polymorphic proteins with cardiovascular disease was analyzed using DAVID and included in the list. As described, these proteins are not explicitly involved in causing cardiovascular disease but may be involved in various pathways and mechanisms leading to diseases. Further analysis of these proteins provides insight into the molecular mechanism yet to be known. The table 8 below shows the proteins with prediction value more than 0.8.

PROTEIN 1	PROTEIN 2	PDB ID	Prediction Probability	Information about the potential protein
DSG1	DSG1	3IFQ , AC	0.924	Mutations in the desmoglein-2 (DSG2) gene have been reported in patients with arrhythmogenic right ventricular cardiomyopathy (ARVC) but clinical information regarding the associated phenotype is at present limited (Syrris *et al.*, 2006). KEGG PATHWAY ENTRY - ko05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)
ABI1	EPS8	1GRI, AB	0.894	Abi1 is a central regulator of actin polymerization through interactions with multiple protein complexes. Mice lacking Abi1 or α4 exhibit midgestational lethality with abnormalities in placental and cardiovascular development (Ring *et al.*, 2007).
BCAR1	BCAR 1	1OV3 , AB	0.894	Gertow *et al.*, 2012; identified rs4888378 in the BCAR1-CFDP1-TMEM170A locus as a novel genetic determinant of cIMT and coronary artery disease risk in individuals of European descent.
CASP1	CASP1	3E4C, AB	0.892	CASP1 haplotype carrying the A(in6) allele was associated with a lower mRNA expression. These results indicate that caspase-1 levels are predictive of future cardiovascular death in patients with coronary artery disease. The role of CASP1 genetic variations in the susceptibility to myocardial infarction requires further investigation (Blankenberg *et al.*, 2006).
ACTN2	ACTN 2	1HCI , AB	0.877	Studies indicate that HF is associated with two different types of remodeling of α-actinin and only one of those was reversed after Cardiac Rescynchronization therapy(Justin *et al.*, 2012).

A2M	CELA1	3HS0, AB	0.861	A total of 35 genes were differentially expressed in cases with CHD versus controls at false discovery rate<0.5, including GZMB, TMEM56, and GUK1. Cluster analysis revealed 3 gene clusters associated with CHD, 2 linked to increased erythrocyte production and a third to reduced natural killer and T cell activity in cases with CHD (Joehans *et al.*, 2013)
ACY1	ACY1	2ZOG , AB	0.861	No significant result found
CAND 1	CAND 1	1U6G, AB	0.861	Obesity is one of the most serious health problems of the 21st century. It is associated with highly increased risk of type 2 diabetes, high blood pressure, cardiovascular disease as well as several cancers. CHOP stability is controlled by a CSN-CRL3Keap1 complex, which is crucial for adipogenesis (Huang *et al.,* 2012).
FARSA	FARSB	3L4G, AB	0.861	No significant result found
EGFL6	EGFL6	2BO2, AB	0.859	With increasing rates of obesity driving the incidence of type 2 diabetes and cardiovascular diseases to epidemic levels, understanding of the biology of adipose tissue expansion is a focus of current research (oberauer *et al.*, 2010)
BHMT	BHMT	1LT8 , AB	0.838	Hyperhomocysteinemia, a risk factor for cardiovascular disease, can be caused by genetic mutations in enzymes of homocysteine metabolism. Homocysteine remethylation to methionine is catalyzed by folate-dependent methionine synthase, or by betaine-homocysteine methyltransferase (BHMT), which utilizes betaine as the methyl donor (Weisberg *et al.*, 2003)
DCP2	DCP2	2DSD , AB	0.827	No significant result found
DHPS	DHPS	1RQD, AB	0.827	No significant result found
BCR	BCR	2B3R , AB	0.826	UPR signaling in cardiovascular disease and its related therapeutic potential (Minamino *et al.*, 2010)
BHMT	BHMT	1LT7 , AB	0.824	Betaine-homocysteine S-methyltransferase (BHMT) uses betaine to catalyze the conversion of homocysteine (Hcy) to methionine. There are common genetic polymorphisms in the BHMT gene in humans that can

				alter its enzymatic activity (Teng *et al.*, 2007).
FGG	FGG	2FFD, AB	0.813	No significant result found
ACVR1	ACVR 1	3O96, AB	0.811	BmpR1a and AcvR1 are needed for normal heart development, in which they play some non-redundant roles, and refine our understanding of the genetic and morphogenetic processes underlying Bmp-mediated heart development important in human congenital heart disease (Thomas *et al.*, 2014).
ABL1	ABL1	2HYY , AB	0.809	The pathophysiological dysfunction of protein kinase signaling pathways underlies the molecular basis of many cancers and of several manifestations of cardiovascular disease, such as hypertrophy and other types of left ventricular remodeling, ischemia/reperfusion injury, angiogenesis, and atherogenesis (Force *et al.*, 2004).
FBP1	FBP1	3KC1, AB	0.809	Insulin-like growth factors (IGFs) are peptide hormones that have significant structural homology with insulin. IGF binding proteins (IGFBPs), in particular IGFBP-1, are important determinants of IGF activity such as enhancing peripheral glucose uptake, decreasing hepatic glucose output and modifying lipid metabolism. Herein factors which alter IGFBP-1 and the utility of measuring IGFBP-1 are considered as the role of IGFBP-1 is explored within the context of insulin resistance and the development of cardiovascular disease (Mehta *et al.*, 2012).
ABLIM 1	ABLI M1	2DFY , CX	0.806	No significant result found
BCL6	BCL6	3LBZ , AB	0.806	The Bcl6 gene encodes a sequence-specific transcriptional repressor and is ubiquitously expressed in adult murine tissues including heart muscle. The objective of this study was to examine the role of Bcl6 in cardiac myocytes (Yoshida *et al.*, 1999).
BAIAP 2	BAIAP 2	1Y2O , AB	0.802	No significant result found

Table 8: Proteins predicted as potentially associated with cardiovascular disorders and their associations described.

6. CONCLUSION

In the present study, the main aim was to identify and distinguish proteins based on the interface parameters of the structure involved. It started with genes associated with disease to the structures associated with diseases. A classifier using Random Forest with iteration 30 algorithm was used for training based on the training set which included interface properties of proteins (with their interacting partners) associated with cardiovascular diseases and proteins (with their interacting partners) associated with other diseases than cardiovascular disease. Based on the training, a list of 42 proteins from a set of 600 unknown polymorphic proteins was obtained which are predicted to be potentially associated with cardiovascular disorders. These proteins may not be the causing factor of particular disease but can be involved in various pathways and mechanism yet unknown to us. Study of their interactions with other proteins can significantly improve our understanding in the molecular mechanism of diseases. Generally, mutations in proteins affect the structure of protein and hence it is becoming important to shift the focus from genetic studies to molecular studies. Significant work has been done recently to incorporate structures and this work is an addition to the previous studies. The wider scope of this study is the characterization of all the hereditary disorders based on their structural properties to gain better understanding in the molecular mechanism (such as effect of mutation on protein structure, which residue is being affected, affect of all the residues including charged, polar, uncharged, how hydrogen bonding between complexes are important and so on) behind these diseases.

7. DISCUSSION AND FUTURE PERSPECTIVE

Main concern of the present study is to present interface analysis of cardiovascular-disorder related proteins to shed lights on details of interactions and to emphasize the importance of using structures in network studies (Kar *et al.,* 2009). The positive results showed that the data on interface properties can be accessed for classification purposes be it between two different diseases, a set of disease having same phenotypic effect with other set of disease with different phenotype, or for all the hereditary diseases known. This study provided insights into the usage fine structural details in disorders related to cardiac system. A further exploration of this study can be utilized to study whether the classifier could work equally efficient when used for classifying all the inherited disease.

These results can be used as evidence when searching for candidate gene in predefined disease loci and also for predictions, to identify novel genes involved in the mechanisms. Studies have been carried out where sequence features such as cDNA, number of exons, protein size were investigated between set of genes known to be involved in hereditary disease and those not (Adie *et al.,* 2005; Lopez-Bigas and Ouzounis, 2004). Functional annotation shared between known diseases genes had also been used to design a classifier (Perez-Iratxeta *et al.*, 2002; Turner *et al.*, 2003). The classification based on genomic sequences have disadvantages associated with them, as algorithms which are based on functional annotations are inherently biased towards a particular known subset of genes. The present study utilized human PPI network with structural properties to compare genes involved in diseases or not. It is assumed that structural properties provide a much wider coverage of the understanding of mechanism of disease than sequence analysis. A promising approach for disease gene discovery is to combine all the evidences available such as PPIs, structures, sequences, transcriptional expression, functional annotation to make a classifier that would be possible to predict accurately without any ambiguities (Calvo *et al.,* 2006).

Protein-Protein interaction network provide a vast area of study in which different criteria, different features can be utilized to devise new diagnostic tool or to understand a novel mechanism. Mutations leading to interaction disruptions or creation of new interaction in disease stare are possible using PPI networks with structural perspective. Further, new prognostic tools can be created by identification of pathways or disease sub networks that get activated only in diseases states. For instance, a recent study integrated protein networks with cancer expression profile and identified pathway that get activated only during tumor progression discriminating metastasis better than earlier described markers (Gonzalez *et al.,* 2012).

Drug target identification and drug design can be made more comprehensive by including disease networks and structural details, for example, structural information on allosteric site or binding site can be used to design potential drug to affect protein function. Further, rebuilding different interaction networks such as metabolic, signaling etc, prediction about hub protein can be made which are involved in various pathways for proper functioning of the cell. This knowledge can be used while designing the drugs, i.e. whether to target hub proteins or not (Gonzalez *et al.,* 2012).

Some studies have shown the association between age-related cardiovascular diseases and osteoporosis with common etiology such as increased risk of hip fracture in women. Recent studies have shown that drugs have common effect for cardiovascular diseases and osteoporosis, for example bone antiresorptive drug reduces the risk of cardiovascular diseases. Similarly, positive effects of statins, antihypertensive drugs have been shown on bone mass. These studies points to the common physiopathological molecular pathways between both the diseases. Statins for example, reduces cardiovascular mortality through reduction of LDL cholesterol levels, these have also been associated to increase bone mineralization in mice and reduction of fractures. All these examples suggest a link between the vascular and skeletal systems; therefore it is the main concern to understand the exact physiopathological mechanism shared between both the diseases as well as to determine common risk factors and genetic determinants (Marini *et al.,* 2010).

8. REFERENCES

1. Adie EA; Adams RR; Evans KL; Porteous DJ; Pickard BS. (2006). SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics. **22(6):**773-4.

2. Aditya Rao; Gopalakrishnan Bulusu; Rajgopal Srinivasan and Thomas Joseph. (2012). Protein-Protein Interactions and Disease; Protein Interactions; Dr. Jianfeng Cai (Ed.); ISBN: **978-953**-51-0244-1; InTech.

3. Aung; Z. (2006). Computational analysis of 3D protein structures. (Doctoral dissertation; School of Computing; National Institute of Singapore).

4. Basse MJ; Betzi S; Bourgeas R; Bouzidi S; Chetrit B; Hamon V; Morelli X; Roche P. (2013). 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. Nucleic Acids Res. **41**(Database issue):D824-7.

5. Butte AJ; Kohane IS. (2006). Creation and implications of a phenome-genome network. Nat Biotechnol. **24(1)**:55-62.

6. Calvo S; Jain M; Xie X; Sheth SA; Chang B; Goldberger OA; Spinazzola A; Zeviani M; Carr SA; Mootha VK. (2006). Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet. **38(5):**576-82.

7. Choura M; Rebaï A. (2011). Structural analysis of hubs in human NR-RTK network. Biol Direct. **6**:49.

8. Duennwald ML; Jagadish S; Giorgini F; Muchowski PJ; Lindquist S. (2006). A network of protein interactions determines polyglutamine toxicity. Proc Natl Acad Sci USA. **103:**11051–11056.

9. Goh KI; Cusick ME; Valle D; Childs B; Vidal M; Barabási AL. (2007). The human disease network. Proc Natl Acad Sci USA. **104(21)**:8685-90.

10. Gonzalez MW; Kann MG. (2012). Chapter 4: Protein interactions and disease. PLoS Comput Biol. **8**(12).

11. Hall M; Frank E; Holmes G; Pfahringer B; Reutemann P; Ian H. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations. **Volume 11**; Issue 1.

12. Huang da W; Sherman BT; Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. **4(1):**44-57.

13. Huang da W; Sherman BT; Lempicki RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. **37(1):**1-13.

14. Jones S; Marin A; Thornton JM. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng. **13(2)**:77-82.

15. Jonsson PF; Bates PA. (2006). Global topological features of cancer proteins in the human interactome. Bioinformatics. **22(18)**:2291-7.

16. Kann MG. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform. **8(5)**:333-46.

17. Kar G; Gursoy A; Keskin O. (2009). Human cancer protein-protein interaction network: a structural perspective. PLoS Comput Biol. **5**(12).

18. López-Bigas N; Ouzounis CA. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. **32(10):**3108-14.

19. Magrane M. and the UniProt consortium. (2011). UniProt Knowledgebase: a hub of integrated protein data Database. bar009.

20. Marini F; Brandi ML. (2010). Genetic determinants of osteoporosis: common bases to cardiovascular diseases? Int J Hypertens. **pii**: 394579.

21. Mitchell; T. (1997). Machine Learning; McGraw Hill. ISBN **0-07-042807-7**; p.2.

22. Moreira IS; Fernandes PA; Ramos MJ. (2007). Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins. **68(4)**:803-12.

23. Mosca R; Céol A; Aloy P. (2013). Interactome3D: adding structural details to protein networks. Nat Methods. **10(1)**:47-53.

24. Nguyen TP; Liu WC; Jordán F. (2011). Inferring pleiotropy by network analysis: linked diseases in the human PPI network. BMC Syst Biol. **5**:179.

25. Online Mendelian Inheritance in Man; OMIM®. McKusick-Nathans Institute of Genetic Medicine; Johns Hopkins University (Baltimore; MD). World Wide Web URL:http://omim.org/.

26. Sam L; Liu Y; Li J; Friedman C; Lussier YA. (2007). Discovery of protein interaction networks shared by diseases. Pac Symp Biocomput. **76-87**.

27. Scheffner M; Whitaker NJ. (2003). Human papillomavirus-induced carcinogenesis and the ubiquitin-proteasome system. Semin Cancer Biol. **13:**59–67.

28. Steward RE; MacArthur MW; Laskowski RA; Thornton JM. (2003). Molecular basis of inherited diseases: a structural perspective. Trends Genet. **19(9):**505-13.

29. Tuncbag N; Kar G; Keskin O; Gursoy A; Nussinov R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Brief Bioinform. **10(3)**:217-32.

30. Venselaar H; Te Beek TA; Kuipers RK; Hekkelman ML; Vriend G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. **11**:548.

31. Wang X; Wei X; Thijssen B; Das J; Lipkin SM; Yu H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. **30(2):**159-64.

32. Xu J; Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. (2006). Bioinformatics. **22(22)**:2800-5.

33. Yan C; Wu F; Jernigan RL; Dobbs D; Honavar V. (2008). Characterization of protein-protein interfaces. Protein J. **27(1)**:59-70.

34. Zhang KX; Ouellette BF. (2011). CAERUS: predicting CAncER oUtcomeS using relationship between protein structural information; protein networks; gene expression data; and mutation data. PLoS Comput Biol. **7**(3).

35. Zhang X; Zhang R; Jiang Y; Sun P; Tang G; Wang X; Lv H & Li X. (2011). The expanded human disease network combining protein-protein interaction information. European Journal of Human Genetics **19**:783-788.