Chapter 2

Review of Literature

2.1 THE PROGRESS OF CANCER

The unrestrained growth and proliferation of cells are the basic abnormality in the progress of cancer. The abnormalities in the multiple cell regulatory systems are the reason of the comprehensive failure in the division management of cancerous cells. It is known that the process of progress of cancer involves a number of steps that includes the malignancy of cells acquired from a sequence of modification which involves mutation. Most cancers are found to be developed at the late stages of life. As a result of genetic perturbations or mutations, the irregular division and growth of a distinct cell occurs and further it directs to the development of cancer cells. According to the arrival of further mutations in cells, the growth of cancer continues. A few of the genetic perturbations deliberate some discerning benefits to the cell. For instance, brisk development and the expressed dominance of the offspring of mutated cell among the cancer inhabitants known as clonal selection and it is continued throughout the growth of cancer. Mutations in the following three broad categories of genes and chromosomal aberrations are responsible for this loss of normal control mechanisms.

a) Proto-oncogenes, the product of them assist the development of cell division and growth. They are frequently concerned in processes such as signal transduction and implementation of mitogenic signals. But, the commencement of these genes leads to a tumor-inducing agent, an oncogene (Todd, 1999).

b) Tumor suppressor genes, which shields a cell from developing cancer and the mutated gene cause a loss or reduction in its function results in progress of cancer. The inactivation of tumor suppressor gene plays a vital role in configuring various types of cancers, than proto-oncogene/oncogene activation (Weinberg, 2014). And,

c) DNA repair enzymes, mutations in this category reduce the effectiveness of the process of DNA repair than the usual level. The reduction in the expression of DNA repair proteins resulted from the mutations in the responsible genes appears
to be significant in the genetic instability characteristic of cancers (Jacinto et al, 2007).

Many cancer mutations alter signal pathway components by transferring signals at the wrong time and regulate the cell proliferation unacceptably, and also there are numerous mutations that inhibit apoptosis (programmed cell death) which have been found in tumors. In B-cell lymphoma, the protein Bcl-2 which inhibits the programmed cell death is found to be the target of a chromosome translocation. As a result of the translocation, the survival of B lymphocytes occurs, instead of being damaged, as a result of the placement of the gene Bcl-2 in an authoritarian element. Most of the human cancer cells express telomerase which could be the reason for the uncontrolled division and it demonstrates its vital function in cancer. If a cell can reactivate the telomerase expression, it can also pause the catastrophic cycle, further to recover adequate chromosomal constancy for endure and its offspring can accede to an extremely anomalous chromosome set, which comprises numerous mutations. The accumulation of further mutations in these damaged cells can drive tumor progression. The steps in the development of tumor are, therefore associated with triggering explicit oncogenes and not triggering explicit tumor suppressor genes as a result of mutation (Alberts et al, 2002).

2.2 PROTO-ONCOGENES AND TUMOR SUPPRESSOR GENES

In the development of cancer, proto-oncogenes and tumor-suppressor genes have an important position. They codes for proteins responsible for the cell growth and proliferation. A mutated proto-oncogene or tumor-suppressor gene plays a vital role in the growth of cancer.
The expression of mutant forms of these proteins results in cancer: growth factors (I), growth factor receptors (II), signal-transduction proteins (III), transcription factors (IV), pro- or anti-apoptotic proteins (V), cell cycle control proteins (VI), and DNA repair proteins (VII). The mutations occurring in classes I – IV proteins commonly lead to governing active oncogenes, while the class VI proteins generally perform as tumor suppressors; as a result of pathogenic mutations, they restrain cells from control and supervision, significantly raising the possibility of the mutant cells to turn into tumor cells. Moreover, class VII mutations drastically boost the possibility of mutations in the other classes. The category of virus-encoded proteins that activate growth-factor receptors (Ia) also can bring on cancer. The inactivating mutations present in one or more proteins of majority of cancers normally function to restrain progression through the G1 stage of the cell cycle (e.g., Rb and p16), while it is not seen in colon cancers. The presence of inactivating mutations which are seen in most of the human tumors such as in protein p53, which generally present at critical spindle checkpoints, stopping the cycle if a preceding step has occurred wrongly or if there is a damage in DNA. Similarly, an active ras gene is seen in numerous human tumors of diverse source (Lodish et al, 2000).
2.2.1 PROTO-ONCOGENES TO ONCOGENES

From the identified oncogenes, a small number of them are resultants of regular cellular genes called as proto-oncogenes, and the resulted proteins of them take part in cellular growth-controlling pathways. The gene \( ras \) and its mutated version \( ras^D \) are the examples for proto-oncogene and oncogene. The normal resulted protein is involved in signal-transduction; and the oncoprotein grants an unrestrained growth-promoting signal (Lodish et al, 2000). Another example is \( Bcr-Abl \) gene seen in Philadelphia Chromosome, and its resulted protein tyrosine kinase involved in unrestrained cell division and growth. The fusion of the wrecked end of chromosome 22 which encloses the BCR gene, and a section of chromosome 9 that encloses ABL1 gene generates a new gene BCR-ABL. The resulted gene codes for a protein that exhibits elevated tyrosine kinase activity. The proteins of cell proliferation are activated by the unregulated expression of the resulted protein from BCR-ABL which in turn leads the cell to grow and divide uncontrollably. Consequently, the Philadelphia Chromosome is allied with Chronic Myelogenous Leukemia as well as other forms of Leukemia (Chial, 2008).

Since majority of the proto-oncogenes carry a vital role in many of the signal pathways, they comes under the category of conserved sequences. Mutations are the normal cause for the transformation of a proto-oncogene into an oncogene. Nevertheless, the following reasons can fabricate oncogenes from the subsequent proto-oncogenes.

a) Missense mutations in a proto-oncogene can result in a functional change on the protein product

b) Localized reduplication of a DNA segment which comprises a proto-oncogene, can be resulted in over expression of the resulted protein

c) By means of chromosomal translocation, keeping the growth-regulatory gene in the control of diverse promoter causes inappropriate expression of the gene

There is slight change in the protein product of an oncogene if it is formed through the first mechanism. An oncogene encodes for an oncoprotein, whereas the proto-oncogene encodes for a normal protein. In contrast, the protein products formed through the other two mechanisms are alike with the regular proteins; their oncogenic
persuade is occurred because of their superior expressions rather than the usual or their lack of expression in cells. The gain-of-function mutations govern here that alters proto-oncogenes to oncogenes; to be precise, mutation in one of the two alleles is adequate for initiation of cancer (Lodish et al, 2000).

2.2.2 VIRUS INDUCED CANCER

The first evidence for virus induced cancer came from the observations by Peyton Rous (1911). He expurgated fibrosarcomas or connective tissue tumors from chickens, continued by grouping them, removed cells and debris by centrifugation. The supernatant was passed through strains with tiny pores, and they were designed in such a manner that, it could retain even the negligible bacteria. Later on it was infused to the chickens. To be noticed, most of them were affected with sarcomas. The transmuting factor of filtrate was identified as Rous sarcoma virus (RSV). Later on, it has been studied that, RSV is incorporated into the host-cell genome and it is a retrovirus whose RNA genome is reverse transcribed into DNA. The retroviruses that are not transformed enclose three genes viz., *gag*, *pol*, and *env* and their corresponding products are virus structural proteins and the reverse transcriptase. The oncogenic viruses like RSV contain the *v-src* gene in addition to the normal retroviral genes viz., *gag*, *pol*, and *env*. Consequent experiments done on mutated RSV revealed that the *v-src* gene alone was required for cancer induction. Thus, this meticulous gene was recognized as an oncogene. Subsequent development reported in 1977, from the studies of Michael Bishop and Harold Varmus; they demonstrated that typical cells of chicks and various other species enclose one gene which is intimately correlated to the RSV *v-src* gene. Moreover, this proto-oncogene was generally identified through the prefix “c” (*c-src*). The correlation of viral oncogene and cellular proto-oncogene disclosed that cancer may be persuaded by the action of normal, or nearly normal, genes. By means of incorporating, or transducing, a normal cellular proto-oncogene into the genome, RSV and other oncogenic viruses could have been arisen. Later on, consequent mutation in the transduced gene would have been transformed it into an oncogene (Lodish et al, 2000).

The cells comprising an incorporated RSV genome, *v-src* (active mutant form of *c-src*) transcribed at improperly elevated rate, and unfettered action of *v-src* protein lead
to constant and inapt phosphorylation of target proteins. The protein v-src is the foremost gain-of-function mutant of c-src, because it can persuade transformation of cell in the existence of the usual c-src proto-oncogene. The oncogene ras is also present in Harvey sarcoma virus, which is a retrovirus. Analogous studies with tumor – cell DNA from numerous other tumors have led to the cloning of numerous oncogenes, suggesting that cancerous genes are also found in many animal retroviruses (Lodish et al, 2000).

2.2.3 TUMOR SUPPRESSOR GENES TO ONCOGENES

The products of tumor suppressor genes normally encode proteins that have specific roles in inhibition of cell division and growth. Functional modification of these proteins accelerates the cancer progression. The TGFβ family of signaling proteins is an example for these types of genes. Mislaying the growth suppression attained by TGFβ-mediated pathways confer to the dawn for various cancers. Presence of mutated TGFβ-RII is identified in colon carcinoma and another protein Smad4 is deactivated in many human cancers (Alberts et al, 2002). The following categories of proteins are commonly identified as products of tumor suppressor genes:

a) Intracellular proteins - eg., p16 cyclin-kinase inhibitor, that regulate/inhibit development through a specific stage of the cell cycle
b) Receptors for secreted hormones - e.g., tumor derived growth factor β that function to reduce cell proliferation
c) Checkpoint-control proteins- e.g., Mad1, that arrest the cell cycle if there is abnormalities in the number of chromosomes or there is a DNA damage
d) Proteins that uphold the programmed cell death
e) Enzymes that contribute in DNA repair

2.3 CANCER GENES EXAMINED IN THIS STUDY

2.3.1 TUMOR SUPPRESSOR GENE MAD1

The presence of abnormal number of chromosomes in cells or aneuploidy is a characteristic of cancer. The abnormality in the spindle check point of cell cycle can lead to aneuploidy (Jallepalli and Lengauer, 2001). Spindle check point is the governing factor
of mitosis which makes a delay in chromosome segregation until the entire sister chromatids connect to the bipolar mitotic spindle (Chao et al, 2012; Fava et al, 2011). This checkpoint mechanism depends on some proteins. Mad 2 (mitotic arrest deficient), one among the spindle check point proteins (Hoyt et al, 1991; Li and Murray, 1991), accumulates specifically on kinetochores that are not stably connected with spindle microtubules (Chen et al, 1996). Mad1 (encoded by MAD1), another spindle check point protein, is the factor helping Mad2 to bind with the kinetochores. Once Mad2 is attached to the kinetochore, it moves to an active form by binding with Cdc20, an activator of the anaphase promoting complex or cyclosome (APC/C) and stops the chromosome segregation (Kim et al, 1998; Peters, 2006; Brady and Hardwick, 2000). Thus Mad2 ensures that chromosome segregation is initiated only when all chromosomes have been attached to both poles of the mitotic spindle (Bharadwaj and Yu, 2004; Peters, 2008). Mad2-Mad1 complex is the only way of Mad2 to get attached with kinetochores. Mainly three functions are carried out by the Mad1 protein in spindle check point. It forms a complex with Mad2 through its C terminus domain and uses its N terminus domain for the localization of kinetochore-spindle attachment structure on chromosomes (Hwang et al, 1998; Chen et al, 1998). There is an essential role for Mad1 in the formation of Mad2-Cdc20 complex, as it helps the Mad2 for the conformational change for the binding of Cdc20 (Yu, 2002). Mad1 also forms a complex with Bub1-Bub3 (budding uninhibited by benzimidazole) in mitosis (Brady and Hardwick, 2000; Hardwick et al, 1996). Hence the mutation in genes related to spindle check point can lead to the condition of aneuploidy thus to cancer (Cahill et al, 1998).

2.3.2 ONCOGENE PIK3CA

The gene PIK3CA, encodes for the catalytic subunit p110α of phosphatidylinositol 3-kinase (PI3K) is located on chromosome 3q26.3. Gene amplifications, deletions and somatic missense mutations in the PIK3CA gene are found with various cancer types (Shayesteh et al, 1999; Ma, 2000). The mutations of this gene are supposed to elevate the kinase activity which is involved in cellular transformation (Katso et al, 2001; Migozuchi et al, 2004). The phosphatidylinositol 3-kinases (PI3Ks) are heterodimeric lipid kinases, consisting of a catalytic and an adaptor/ regulatory
subunit, also they are significant controllers of cell proliferation (Li and Peter, 2008; Volinia et al., 1994). The PI3K family consists of three main classes (class I, II and III). The catalytic subunit of class I PI3K has four isoforms namely p110α, p110β, p110δ and p110γ. The isoforms p110α, β and δ join with the standard regulatory subunit p85, and p110γ joins with the regulatory subunit p101 (Li and Peter, 2008; Karakas et al., 2006). All the four isoforms of PI3K share same domain composition: an amino-terminal adaptor-binding domain (ABD), Ras-binding domain (RBD), C2 domain, helical domain and carboxyl-terminal kinase domain. The platform for the regulatory subunit to interact with catalytic unit is considered to be the domain ABD and is also mediate the binding of p110 and Ras-GTP. The domain C2 is the region where lipid membranes are connected with, and the helical domain support other domains of the catalytic unit p110 (Walker et al., 1999). The inter SH2 (iSH2) domain of the regulatory subunit p85α is the binding domain for p110 and it also contains multi domains (Dhand et al., 1994; Klippel et al., 1994). Hence the missense mutations could interfere with the p85- p110α interaction which could lead to the relieving of the inhibition of p110α.

2.3.3 TUMOR SUPPRESSOR GENE STK11

The gene STK11 is a tumor suppressor gene and encodes for serine-threonine kinase which has a critical role in regulating cell growth and apoptosis (Alessi et al., 2006; Karuman et al., 2001). Inactivation of this gene leads to the development of cancer. Many of the mutations in STK11 are small deletions or point/missense mutations that are present in the STK11 catalytic kinase domain and minority occurs within its COOH-terminal non-catalytic region, thereby resulting in STK11 protein reduction, loss or inactivation (Schumacher et al., 2005; Boudeau et al., 2003). The germline mutations in STK11 gene also cause a rare dominantly inherited disease, Peutz–Jeghers syndrome (PJS) (Hemminki et al., 1998; Jenne et al., 1998) characterized by the presence of gastrointestinal (GI) hamartomatous polyps and mucocutaneous melanin pigmentation which is frequently expressed in lips and oral area (Calva et al., 2008; Schreibman et al., 2005). It is found that the hamartomatous polyps are commonly developed in small bowel, colon and in stomach, causing abdominal pain, bowel obstruction and severe gastrointestinal bleeding (Resta et al., 2013). The risk factor for cancer in PJS patients
has been estimated to be 15 - fold higher than the general population (Brosens et al, 2007). An increased risk for gastrointestinal (GI) and non - gastrointestinal (non-GI) cancer is observed in PJS patients. Frequently observed GI targeted cancers include stomach, small intestine, colon and pancrease cancer, whereas cancers in breast, endometrial, ovary and lung tumours are frequent in non-GI targeted cancer (Resta et al, 2013).

Mutations in the tumor suppressor gene STK11 on chromosome 19p13.3 is the only known cause for PJS (Amos et al, 2004). STK11 associates with the pseudokinase STRAD (STe20-Related ADaptor) and the scaffolding MO25 (MOuse protein 25) in a 1:1:1 heterotrimeric complex in cell (Zeqiraj et al, 2009). The catalytic kinase domain of STK11 is activated when pseudokinase domain of STRAD binds to it, and this causes the transport of STK11 to the cytoplasm (Baas et al, 2003). Contrasting to the usual activation of protein kinases by phosphorilation, STK11 is triggered by binding to STRAD and MO25 through an unrecognized, phosphorylation-independent mechanism. Although STK11 doesn’t have phosphorilated activation loop, it gets an active structural variation. The αC- helix of STK11 is rotated into the canonical closed conformation, by forming the conserved salt bridge between Lys (78) and Glu (98). This active conformation of STK11 appears to be achieved through contributions of both STRAD and MO25. The C-terminal lobe of STRAD interacts with both N and C terminal lobes of STK11 kinase domain (Boudeau et al, 2004; Hawley et al, 2003). Mutations in STK11 can lead to its inactivation without affecting this complex assembly.

2.3.4 PROTO-ONCOGENE BRAF

The proto-oncogene BRAF encodes for the protein B-Raf (serine/threonine-protein kinase B-Raf) and it is a member of the Raf kinase family which involves in cell signaling and cell growth. This protein plays a vital role in regulating the MAPK/ RAS-RAF-MEK-ERK signaling pathway, which mediates cellular growth, differentiation and survival. This pathway is recurrently deregulated in various cancers with RAS or RAF mutations (Haling et al, 2014; Davies et al, 2002). B-Raf is a 766-amino acid, and it is a combination of 3 conserved domains which are the characteristics of Raf kinase family, conserved regions 1, 2 and 3, a Ras - GTP binding self regulatory domain, a serine rich
hinge region and a catalytic protein kinase domain. B-Raf gets its dimerized form through hydrogen-bonding and by means of electrostatic interactions of its kinase domain in its active conformation (Daum et al, 1994; Cutler et al, 1998; Bollag et al, 2012). Out of the three isoforms of RAF (ARAF, BRAF and CRAF), BRAF is less regulated (Haling et al, 2014). RAF activates the dual-specificity protein kinases MEK1 and MEK2. Both MEK1 and MEK2 function specifically in the MAPK cascade (Hatzivassiliou et al, 2010; Heidorn et al, 2010).

2.4 PREDICTION OF THE POTENTIAL MISSENSE MUTATIONS BY VARIOUS TOLERANCE PREDICTORS

Even though the outcomes of missense mutations are difficult to foresee compared to the nonsense mutations, a number of tools are invented today to classify the missense mutations as neutral and passenger. They are commonly known as tolerance predictors, and they assess the effect of missense mutations on the protein. To find the impact of mutation, these predictors analyze the following criteria: evolutionary conservation, variations occurred in the physico-chemical properties of the amino acid, the surrounding sequence of the mutated amino acid and the structural variation in the protein. Based on the method used for the prediction, tolerance predictors can be divided in to three; and the methods are,

a) Machine learning methods, which uses trained classifier algorithms to identify the less stable mutants

b) Evolutionary based methods analyze a deleterious mutation by comparing the related sequences using a multiple sequence alignment

c) Bayesian methods, which apply the Bayesian statistics from a set of known data for damaging and neutral (Thusberg et al, 2011)
2.4.1 MACHINE LEARNING BASED TOLERANCE PREDICTORS

2.4.1.1 RANDOM FOREST CLASSIFIER

A random forest is a classifier consisting of a collection of tree structured classifiers \( \{h(x, \Theta_k), k=1, \ldots\} \) where the \( \{\Theta_k\} \) are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input \( x \). It is based on classification and regression trees (CART). CARTs are decision trees that assign vector data into classes. The vector data implies the attributes which are used by the trees to classify the data. The random forest algorithm can have vast number of classification trees recursively. A new data is assigned to the class when it gets maximum support from the majority of trees. The trees are grown based on the principle of bagging, more precisely they are grown in such a manner that, for each tree, \( N \) number of samples from the training set is chosen with replacement. The samples which are not selected will be used for the calculation of error in the classification (Breiman, 2001).

The random forest algorithm follows the technique of bootstrap aggregating or bagging. For a training set \( X = x_1, \ldots, x_n \) with responses \( Y = y_1, \ldots, y_n \), bagging repeatedly, or say \( B \) times selects a random sample with replacement of the training set and fits trees to these samples. The method of de-correlating the trees can be done by bootstrap sampling by showing them different training sets (Gareth et al, 2013).

2.4.1.2 SUPPORT VECTOR MACHINE CLASSIFIER

It is based on machine learning methods and it also exploit in classification of data. Data classification by support vector machines is based on a hyperplane or a set of hyperplanes in high dimensional space. It allows splitting data into classes and they are represented as points in space. The margins are defined by the separation of hyperplane and the nearest data point on each side of hyperplane. SVM exploit the margin around the extrication hyperplane. The verdict function is fully determined by a subset of training samples or the support vectors. Solving SVM is a quadratic programming problem (Hsu et al, 2003).
Fig. 2.2 Principle of Support Vector Machines (The margins are maximized between the planes, the nearest data points in both classes (H-H) and non-optimal separators A-A and B-B are also shown in the figure) (Hepworth et al, 2012).

2.4.1.3 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are widely used in various applications including data classification. They mimic the activity of biological neural networks such as the brain. An ANN is encompassed of a network of artificial neurons/nodes and which are connected to each other and the connections are assigned by weights. The strength of the weight defines inhibition (maximum being -1.0) or excitation (maximum being +1.0). The nodes are designed in such a way that, a transfer function is built in each of the design. Simple neural network architecture is called perceptron and it consists of two types of nodes: input nodes, which are used to represent the input attributes, and an output node, which is used to represent the model output. The figure below demonstrates an example of perceptron architecture. It contains 3 input nodes (x1, x2, and x3) and an output node y. Each input node has an identical weight of 0.3 to the output node and a bias factor of t = 0.4. The output node computes the weighted sum of the input value, subtracts the bias factor, and then produces an output that depends on the sign of the resulting sum (Pang et al, 2005; Toby, 2007).
A multilayer artificial neural network contains several intermediary layers or hidden layers between its input and output layers. The nodes embedded in these layers are called hidden nodes. An example of feed-forward neural network architecture is shown as below. The multilayer artificial neural network may use activation functions such as sign function, linear function, sigmoid function, hyperbolic tangent functions and so on. These various type activation functions allow the hidden and output nodes to model non-linear relationship between the input parameters and output values. The goal of the multilayer artificial neural network learning algorithm is to determine a set of weights $w$ that minimize the total sum of squared errors (Pang et al, 2005; Toby, 2007).
Various tools based on machine learning approach includes 1) I Mutant (Capriotti et al, 2005) is a neural-network-based web server for the automatic prediction of protein stability changes upon single-site mutations, 2) PhD – SNP (Capriotti et al, 2006) is a SVM based method which has been trained using human variant data from Swiss-Prot, 3) SNPs&GO (Calabrese et al, 2009) is a tool from the developers of PhD-SNP and this predictor also uses SVM classifier, 4) SNAP (Bromberg and Rost, 2007) is a machine learning method which makes its predictions based on a trained neural network, 5) CanPredict (Kaminker et al, 2007) is based on machine learning approach that predict whether the mutation is driver or passenger 6) CHASM (Carter et al, 2009) or Cancer specific high- throughput annotation of somatic mutations detects the driver mutations by using random forest classifier.

2.4.2 EVOLUTIONARY CONSERVATION BASED TOLERANCE PREDICTORS

The fundamental postulation of evolutionary conservation based tolerance predictors is that, conserved regions are essential for the function of the protein (Miller and Kumar, 2001) and the disease causing mutations are often found to influence conserved positions and covarying sites which in turn leads to structural instability and
change in function. For a single mutation, three different types of sequence conservation can be identified and they are invariant positions, physico-chemically conserved sites, and the network formed by covariant residues. Invariants in Type I conservation includes residues in a particular sequence position and if the residues are replaced but, the physicochemical properties such as hydropathy, size, and electrostatics etc., are still conserved, it is known as Type II conservation whereas, if the residues that evolve in a coevolving manner such that their interaction and function are retained, it is referred as Type III conservation. In order to gain knowledge on structurally and/or functionally important regions in proteins, the evolutionary conservation of the sequences needs to be examined. Residues with type I or II conservation can be identified by entropy calculations (Shen and Vihinen, 2004) and the low entropy of residue positions displays the highly conserved regions that are taking part in the main activity of the protein. Type II conservation can be identified by calculating information and entropies with the normal alphabet of 20 amino acids (Shenkin et al, 1991). The sequence conservation can be analyzed by using multiple sequence alignment (MSA) and there are several approaches to conduct MSA and they are designed to quantify the degree of conservation at each aligned position. Various methods can be used to enumerate positional covariation, including chi-squared statistical analysis and identification of reciprocal data between any base-pair positions which reflects covariation. Following are some of the tolerance predictors which follow evolutionary conservation, 1) SIFT (Ng and Henikoff, 2001), 2) Panther (Thomas et al, 2003).

2.4.3 BAYESIAN METHODS BASED TOLERANCE PREDICTORS

Naive Bayesian classifiers assign class labels to problem instances, denoted as vectors of feature values, where the class labels are drawn from some finite set. The classifier has to be trained with a training set for assigning data to a class. The training set comprised of a finite set of feature vectors for which the class is known. A statistical model will be constructed based on the training set to describe the data. This classifier is based on a conditional probability model explained by Bayes’ theorem and the theorem states that the probability of a feature vector V belonging to a particular class C can be determined by first calculating the product of prior probability that an arbitrary feature
vector belongs to class C and the likelihood of observing a particular feature vector V
given that this feature vector belongs to class C. This product is then divided by the
probability of observing this particular feature vector from any class (Ioan, 2006). There
are various tools using naive Bayesian classifier including, PolyPhen-2 (Ramensky et al,
2002), Mutation Taster (Schwarz et al, 2010).

2.5 MUTATIONS AND PROTEIN INTEGRITY

Mutations have importance in the structural and functional properties of a protein.
A mutation is considered as functional, if it is positioned at the ligand-binding surfaces,
catalytic site and regulatory sites of a protein and they destruct specific interactions
between the ligand and binding domain, which in turn affect the specificity and activity,
without triggering large agitations in the structure. In contrast, mutations can cause
changes in amino acid size, charge, hydrogen bonds, salt bridges, S-S bridges, if it affects
the buried residues in the protein core. As a result of this, the protein may undergo
anomalous folding and leads to aggregation and also it may loss its thermodynamic
stability. Various traits of the structural change up on mutation has to be considered in
order to comprehend the effect of mutation on the stability of protein; and the protein
stability relies on various strong and weak interactions (Vogt et al, 1997). The programs
PIC (Tina et al, 2007) and PSAIA (Mihel et al, 2008) are structure based algorithms and
are used to compute the interactions stabilizing native and mutant structures. The
importance of hydrogen bond for the stabilization of secondary structures is also
accountable in the compactness of the protein structure (Bikadi et al, 2007). The
programs HB plot (Bikadi et al, 2007) and HBAT (Tiwari and Panigrahi, 2007) belong to
the category of hydrogen bond analysis tools. Moreover there are many programs
available for the secondary structure prediction and validation which includes Stride
(Heinig and Frishman, 2004).

2.6 MOLECULAR MODELING APPROACH FOR THE SIGNIFICANT
MISSENSE MUTATIONS

The significant missense mutations can be predicted by adopting any of the
methods listed above. In addition to this, various molecular modeling approaches viz.,
Potential Energy Calculations, RMSD (Root Mean Square Deviation), Energy Minimization, Docking, Normal Mode Analysis, Molecular Dynamics and Conformational Sampling, could help for further evaluation of these missense mutations.

2.6.1 ENERGY MINIMIZATION

After creating the 3D structure of a protein by using any software which are available today, the structure should undergo for the process of energy minimization as during the construction process of the structure, it may have possessed unfavourable bond lengths, bond angles/torsion angles and unfavourable non-bonded interactions (i.e., atoms from diverse regions of the molecule inhabiting the equal area of space). The process of energy minimization is in such a way that, it computes the energy of the starting molecule, then the modified parameters such as bond length and bond angles will be given to generate a new structure in order to see whether it is energetically more stable or not. Minor alteration in bond angle or bond length will have a large impact on the overall energy of the molecule, if the preliminary structure is inherently unstable, which in turn results in a large energy difference as depicted in the figure below.

![Energy Minimization Diagram](Image)

**Fig. 2.5 Energy Minimization**
The energy minimization program will rectify this by distinguishing those which lead to stabilization, from those do not. As a result of this, a structure will be generated with minimum energy and has only slight changes up on structural variation. This structure will be considered as the stable structure and the program will stop at this stage (Stich et al, 1989; Brooks et al, 2009; Mackerell, 2004).

2.6.2 ROOT MEAN SQUARE DEVIATION

As the structure is much more conserved than sequence, the comparison of structures will give information on the various characteristics of proteins such as origin, function etc., and also the biological prehistory. Examining the structurally correlated proteins may disclose the impacts of the amino acid sequence modifications that have risen during evolution (Perutz, 1965). Studies done in individual protein families indicated that mutations, insertions and deletions create changes in 3D structure (Chothia and Lesk, 1986). Later on, evidence for increase of mutations at different rates in the buried, intermediate, and exposed regions of protein structures, with increasing divergence was evolved. It explained the exponential relationship between the divergence of structure and sequence (Sasidharan and Chothia, 2007). By optimally superposing, the structural divergence of two homologous proteins can be measured and it is the most common method used for structure comparison. This works in such a manner that, until the structure is draped on the top of the similar structure, it will undergo for rotation. RMSD or the Root-Mean-Square Distance/deviation is defined by,

\[ \text{rmsd} = \sqrt{\frac{\Sigma d^2}{n}} \]

Where \( d \) is the distance calculated between each of the \( n \) pairs of corresponding atoms in two optimally superposed structures. The value of RMSD is 0 for identical structures, and its value increases according to the variation that comes in the structure. The values of RMSD are considered as reliable indicators of variability when applied to very similar protein structures, such as various conformations of the same protein. RMSD is used to measure the structure of a partially folded protein and the structure of the native state. Many studies on protein folding simulations used RMSD to find where the target protein
stands between the folded and unfolded state (Carugo and Pongor, 2001; Mizuguchi and Go; 1995; Carugo, 2007).

2.6.3 NORMAL MODE ANALYSIS

The structure – function relationship of proteins is considered as one of the important exemplars in biology. Structural change occurred by mutation may cause to change the overall macroscopic properties too. The function of the protein frequently related to the movement of the protein domains or a subsection of protein. It is also to be considered that the great movements normally occurring at hinges, flaps, gates and floppy regions of the protein. The traditional method to determine the motion of protein is molecular dynamics (MD) simulations. These simulations shows, as the timescales of the dynamics are small enough, that MD is a tool that cannot be replaced with any other tool. There are other tools apart from MD that gives insight into dynamical fluctuations in a different aspect. Normal mode analysis is one among them (Eric and Otto, 2010). For the prediction of all possible movements of a given macromolecule, Normal mode analysis (NMA) is considered as a potent tool (Tama, 2003). Normal mode analysis finds the normal intensive movement of the protein by examining the vibrational normal modes resolute by its structure and the interactions within its diverse regions. A number of methods are used to compute the normal modes of proteins, viruses, or large protein assemblies to foresee their functional movements. Studies have explained that, the elastic network and other coarse grained models are dominant tools that can abridge a complex protein into a plain network of connections which can give a sensible picture of its dynamical properties. The current growth of tractable all-atom techniques offers quantitative approximates of frequencies and more accuracy in unfolding displacement patterns (Eric and Otto, 2010). Relevance of NMA in structural biology is highly accepted by the research community and they widely used in the study of protein conformational changes after a ligand binding, membrane channel opening and closure and so on.
2.6.4 CONFORMATIONAL SAMPLING

By means of searching the conformational space of proteins, conformational sampling technique finds out the rigid connection between accessible conformations and function of protein. First aim in conformational sampling is to obtain stable states of protein. Procedures to sample conformational space aim to find the conformations with least energy and the computation of canonical ensembles (Liwo et al, 2008). Conformational ensembles portray the structure of intrinsically unstructured proteins with a flexible nature and without a stable tertiary structure. Therefore, they cannot be explained with a lone structural representation (Fisher and Stultz, 2011; Varadi et al, 2014; Dyson and Wright, 2005). An ensemble consists of a set of conformations that collectively depicts the structure of a flexible protein. Although the measure of conformational freedom is tremendously high, flexible/disordered protein normally fluctuates from completely random coil structures (Communie et al, 2013; Kurzbach et al, 2103). The genetic algorithms such as the conformational space annealing (CSA) are presently the highly capable method to uncover the global minimum of the potential-energy function. Further acceleration in computations can be attained by introducing the hierarchical/multiscale approach, in which the huge computations are performed at the coarse-grained plane (Liwo et al, 2008; Lee et al, 1997; Hansmann, 1997; Sugita and Okamoto, 1999).

2.6.5 DOCKING

Computational approach to predict the binding modes and affinities of different compounds is referred as docking and it carries huge importance in the structure based drug design. The capability to replicate the experimental binding modes of ligands is deliberated as the major characteristic of a good docking program. A docking program operates in the following steps, 1) Collection of conformations will be made for a small molecule during the docking or before the docking, 2) By using techniques such as Monte Carlo simulation every single conformation will be traced in the active site in diverse orientations and it is known as “pose”, 3) Finally among the poses, best overall pose is selected by means of a scoring function. (Kuntz, 1992; Verdonk et al, 2003; Vigers and Rizzi, 2004).