ANALYSIS OF VARIATION SITE IN POST-TRANSLATIONAL MODIFICATION

Master's Thesis

Masters degree program in Bioinformatics

Institution of Biomedical Technology (IBT)

University of Tampere, Finland

Etsehiwot Girum Girma

August 2012

Acknowledgment

First of all I would like to thank God who has helped me during my studies and through my life.

I am grateful for my thesis advisor professor Mauno Vihinen and the Bioinformatics research group members who has helped and supported me during my thesis work.

I would like to thank my wonderful parents Girum Girma, Yemisrach Debebe and also my brother Michael Girum whom I love and adore for having them in my life.

I would also like to thank my beloved husband Gossa Garedew. I don't have any words for his support and love which has enable me to complete my studies.

August 2012

Etsehiwot Girum

UNIVERSITY OF TAMPERE

Masters degree program in Bioinformatics

Etsehiwot Girum Girma: Analysis of variation sites in post-translational modification

Master of Science Thesis

August 2012

Major subject: Bioinformatics

Supervisor: Prof. Mauno Vihinen

Reviewers: Prof. Mauno Vihinen, Docent Csaba Ortutay

Abstract

Post-translational modification (PTM) is a modification of a protein after its translation. This modification can occur either by the covalent addition of particular chemical groups or by enzymatic cleavage of peptide bond. Post-translational modifications play an important role in signaling pathways, protein stability, oxidative regulation of proteins and cellular localization. Variations of the modification sites can highly affect or disrupt these important biological processes and can lead to disease. The aim of this study was to analyze the variation sites of post-translational modification sites and their relation to disease.

Experimentally verified post-translational modifications were downloaded from Human Protein Reference Database (HPRD) web site. The Single-nucleotide Polymorphism (SNP) data, which contains both pathogenic and neutral missense variations, were matched against the post-translational data to filter out Post-translational Modifications (PTMs) in variation sites. DRUMs, WAVe and Locus specific databases were used to separate disease-causing variations, which occur at the PTM sites. To study the conservation of the variation sites ConSurf was used and a statistical analysis was done by using hyper geometric distribution and t- test.

Disease causing variations were found in both pathogenic and neutral datasets. The conservation score of disease causing variation has indicated that they are more conserved than the benign variations. To further study the variation sites of PTM one can investigate the gene function of PTMs in order to understand which molecular and cellular functions are disrupted by disease causing variations.

Abbreviations

AT	Acetyltransferases
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
CID	Collision-induced Dissociation
dbSNP	SNP Database
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
HPRD	Human Protein Reference Database
IDBASE	Immunodeficiency Database
KMTs	Methyltransferases
MS	Mass Spectrometry
NAT	N-acetyltransferase
PrP	Prion Protein
PTM	Post-translational Modification
RNA	Ribonucleic Acid
SNP	Single-nucleotide Polymorphism

Table of Contents

1 INTRODUCTION

The interest in analyzing protein modifications and their effect in cell biology and pathogenesis has made human proteome and system biology the most important field of study (Tejaswita and Amrita, 2011). The structure of a protein plays a big role in its functionality and chemical alternations of proteins after translation, which is known as post-translational modifications, have a great significance to the structural and functional diversity of the proteins (Li et al., 2010).

These modifications can be caused by covalent addition of chemical groups to the amino acid side chain or by cleavage of the peptide bond. And all this changes have consequences in the cell function, in the activity state, localization and turnover of the protein itself.

Over the years many studies related to the importance of post-translational modifications have been done. For example, in one study it mentions that the importance of glutathionylation (oxidative PTM) in directly regulating human glyoxalase 1 (Glo1) which is cystolic enzyme that catalyze the conversion of toxic α-oxo-aldehydes into the corresponding α-hydroxy acids using L-glutathione (GSH) as a cofactor (Birkenmeier et al., 2010). In another study post-translational modification of lysine residues at the tumor suppressor P53 C terminus plays an important role in regulating the stability and activity of P53 (Olsson et al., 2007).

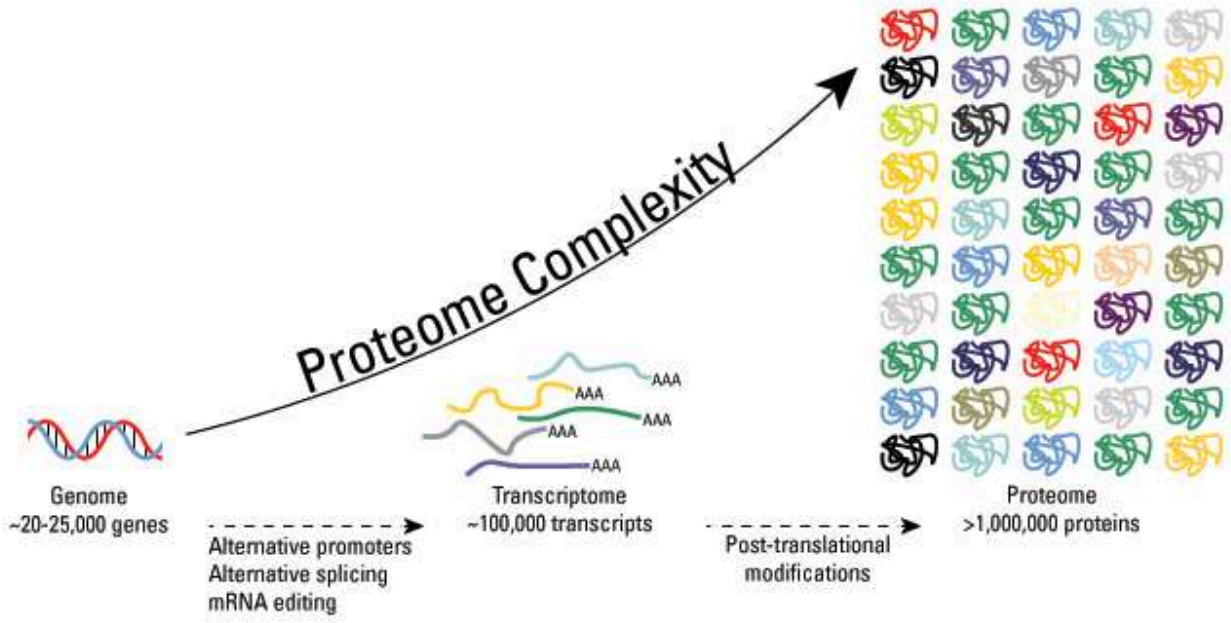
When a post-translational modification site is affected by variation the role may change and become a threat to health. The variations could be loss or gain of the modifications site due to substitution of amino acid residues. Systematic studies have been done by the help of databases containing disease-causing variations by relating post-translational modification to disease. It was stated in one research that protein phosphorylation predominantly occurs within intrinsically disordered protein regions (Lilia et al., 2004). Vogt et al. looked into the gain of N-linked glycosylation sites and their involvement in disease predicting that a number of disease associated mutations introduce changes in glycosylation patterns by creating NX[ST] motifs (Vogt et al., 2005; Vogt et al., 2007). Also in another research gain and loss of phosphorylation target sites may be an active mechanism in human cancer (Radivojac et al., 2008).

This study was done to analyze the distribution of post-translational modification sites in disease causing variations and to see the conservation of amino acid residues in both disease causing variations and post-translational modification sites. In general the study analyzes the influence of disease causing variations on post-translational modification sites.

2 LITERATURE REVIEW

2.1 Post-translational modification

The human genome is estimated to encode 20,000 to 25,000 protein-coding genes. This number has been revised many times through the years. And the total number of proteins in the human proteome is around 1 million (Jensen et al., 2004). As it can be seen from this number one gene can encode multiple proteins. This makes the human proteome complex and the complexity is further facilitated by post-translational modifications (Li et al., 2010).



Source: http://www.piercenet.com/browse.cfm?fldID=7CE3FCF5-0DA0-4378-A513-2E35E5E3B49B

FIGURE 1 Post-translational modification and proteomic diversity.

Knowing post-translational modification is important as it is a well-known fact but they are not discovered as often as possible because of the lack of suitable methods. In previous years PTMs have been identified using various methods such as Edman degradation, amino acid analysis, isotopic labeling or immunochemistry (Mann and Jensen, 2003).

2.2 Detection of post-translational modifications by tandem mass spectrometry

The identification of the protein is the first step in PTM analysis. After identifying the protein it will be compared to a known amino acid sequence

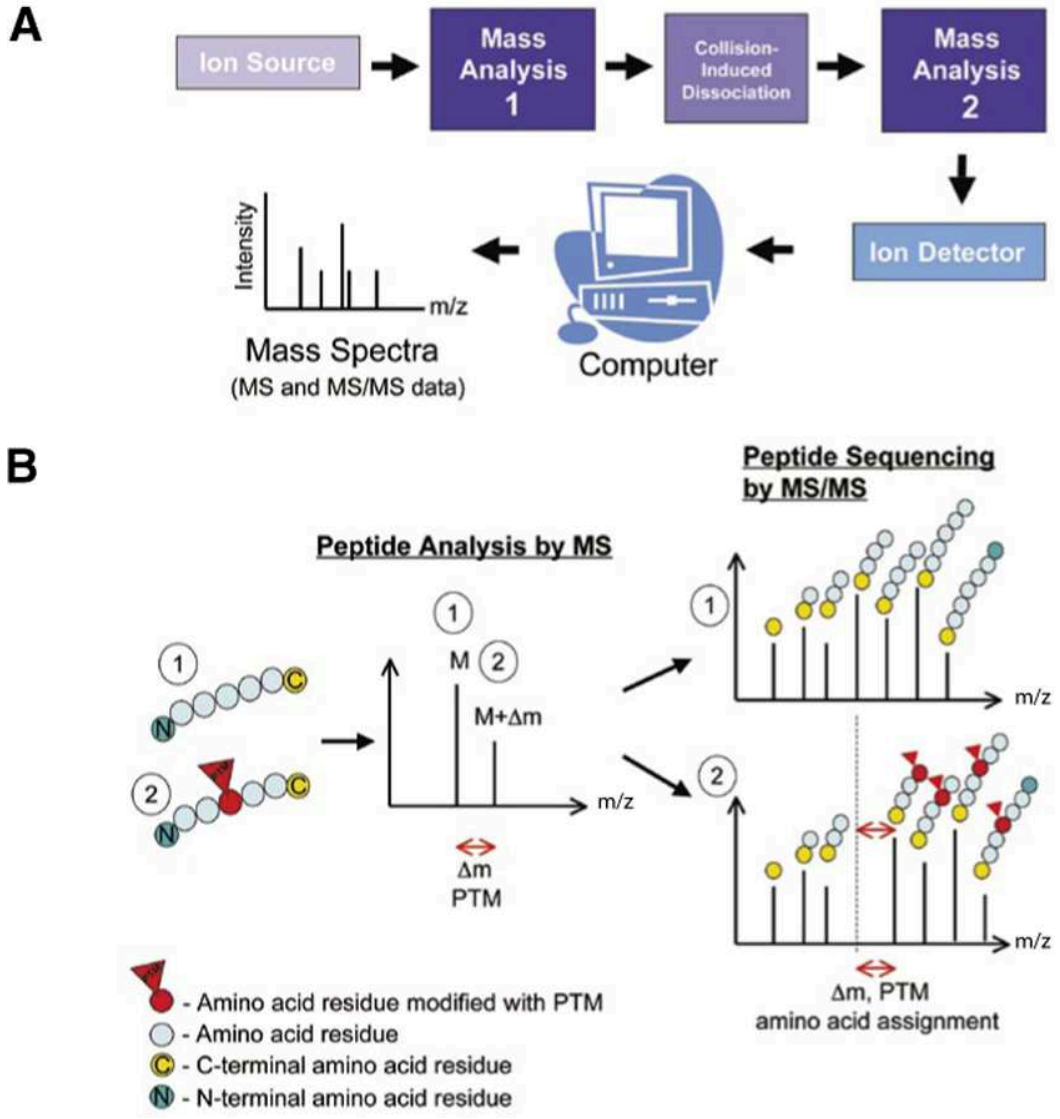
When a protein undergoes post-translational modification the presence of covalent modification changes the molecular weight of the modified amino acid and this can be detected by tandem mass spectrometry (MS/MS) which involves multiple steps of mass spectrometry. MS has plenty of advantages over the other methods such as it is very sensitive, it has a great ability to identify PTM sites and it identifies PTM in complex mixture of proteins (Larsen et al., 2006).

By using chemical reagents or proteases the protein can be converted to peptides because peptides are more amenable to MS and MS/MS. Ionization of the peptide will help to determine the exact weight of the peptide.

As shown in figure 2 to detect PTMs by using tandem mass spectrometry first the ion source which contain the ionized peptide will go under MS survey scan to analyze the mass (the first MS) then by using the mass-to-charge ratio (m/z) value the peptide ion of interest can be separated. In order to activate the separated peptide ion species collision-induced dissociation (CID) is used this will pass on internal energy to the ions and activate their fragmentation. Then the m/z values of the fragments are determined by mass spectrometry (the second MS) (Larsen et al., 2006). This collection of fragment ion reveals the sequences of the amino acids (Steen and Mann, 2004).

The fragment ion signals reflect the amino acid sequence as read from either the N-terminal (b-ion series) or the C-terminal (y-ion series) direction. By determining the mass difference between b-ion series or y-ion series it is possible to identify the individual amino acids (Roepstorff and Fohlman, 1984).

Knowing the exact mass difference is significant in order to describe what types of modifications are present. By comparing the experimentally obtained molecular mass with the calculated amino acid sequence of the protein mass will determine any mass increment. The mass increment is caused by the covalently attachment of chemical groups for example phosphorylation (+80Da) and nitration (+45 Da). And the other scenario that causes mass difference is the hydrolytic cleavage of the peptide bond which leads to mass deficit (Larsen et al., 2006).



Source: http://www.biotechniques.com/multimedia/archive/00002/BTN_A_000112201_O_2231a.pdf

FIGURE 2. Tandem mass spectrometry (MS/MS) for mapping post-translational modifications

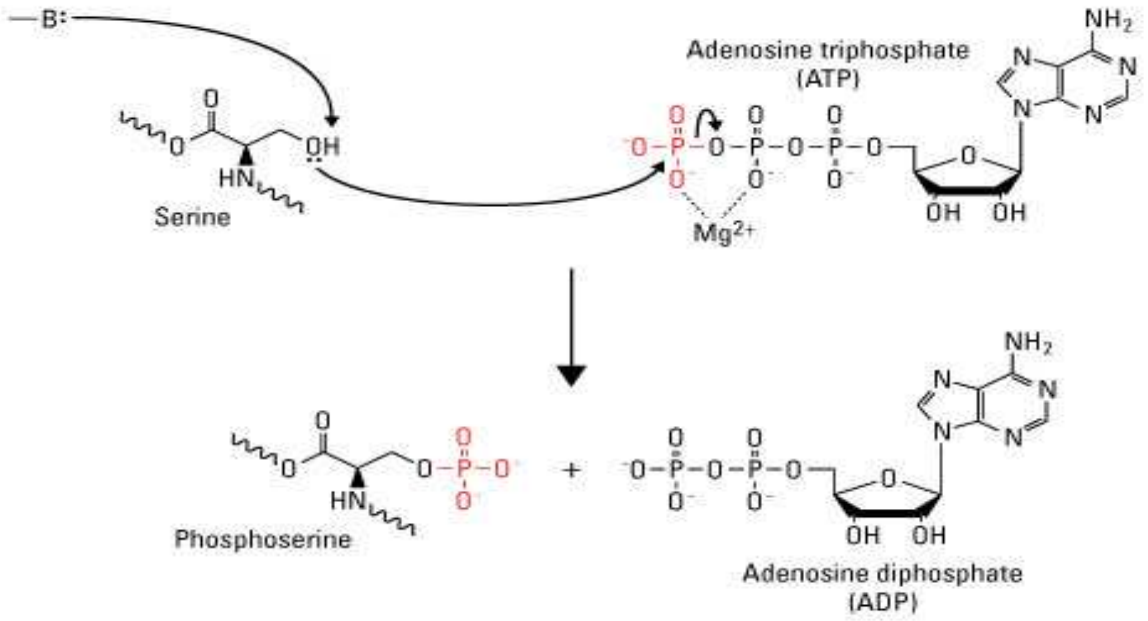
2.3 Types of PTM

2.3.1 Phosphorylation

Phosphorylation is one of the most common and well-studied types of modification. It plays a vital role in regulating protein function and transmitting signals throughout the cell.

The enzymes that catalyze the protein phosphorylation are the largest class of post-translational modification enzymes and they are called kinases. It is estimated that there are more than 500 kinases in human genome (Walsh et al., 2005).

Phosphorylation happens when a phosphate group is added to serine, threonine or tyrosine. When the amino acids attack the terminal phosphate group ($Y\text{-}PO_3^{2-}$) on ATP (adenosine triphosphate) with their nucleophilic (-OH) group the magnesium (Mg^{2+}) will facilitate the phosphate group to transfer to the amino acid side chain. Figure 2.3 below shows serine phosphorylation. The (-OH) group of serine facilitates nucleophilic attack of γ-phospehate group on ATP which results the transfer of phosphate group to serine forming phosphoserin and ADP.



Source: http://www.piercenet.com/browse.cfm?fldID=4E12BA00-5056-8A76-4E40-CD0254A2E35

FIGURE 3. The diagram of serine phosphorylation

2.3.1.1 Phosphorylation in bacteria

Bacteria proteins are involved in serine/threonine specific phosphorylation. These modifications are associated with secondary metabolism, oxidative stress response and sporulation (Cozzone et al., 2005). Phosphates and serine/threonine kinases are also involved in bacterial virulence.

The first tyrosine phosphorylation was discovered in Escherichia coli (E.coli) (Manai and Cozzone, 1982). Capsule production, growth, proliferation and migration are some of the essential cellular process that are directed by protein tyrosine phosphorylation (Zhang et al., 2005). The enzyme that catalyzes bacterial tyrosine phosphorylation is bacterial tyrosine (BY) kinases. BY kinases can function in two ways, as anchor and intercellular catalytic domain (Grangeasse et al., 2007). In its structure it contains walker A (P- loop) and B motif. BY kinases regulate the synthesis and secretion of polysaccharides through posphorylation and activation of UDP sugar dehydrogenases and glucosyltransferases (Stulke et al., 2010). The dephosphorylation of tyrosyl-phosphorylated protein is catalyzed by bacterial tyrosine phosphatases.

2.3.1.2 Phosphorylation in plant

Proteins in plant undergoes through reversible phosphorylation. Protein phosphorylation occurs as a response to many signals including pathogen invasion, temperature stress and nutrient deprivation.

In mid 1998 around 500 plant protein kinases have been discovered. In Arabidopsis thaliana alone there are 175 protein kinases. Based on the article published by Hanks and Hunter in 1995 the four major families of plant protein kinases are ACG group, CaMK group, CMGC group and conventional PTK group. Cell growth, gene expression and sensing environment conditions are controlled by network of protein serine/threonine kinases. (Hardie et al.,1999).

The dephosphorylation of proteins is catalyzed by protein phosphatase. In plant protein phosphatase activity has been reported in sub cellular compartment including mitochondria, chloroplast, nuclei and cytosol (Huber et al., 1994 and Mackintosh et al., 1991).

2.3.2 Glycosylation

Glycosylation is a process when a protein is attached to carbohydrate group (sugar moieties) by glycosidic bonds. Based on there glycosidic linkages there are five types of glycosylation.

1. **N-linked glycosylation:** as the name implies N- glycosylation occurs when glycans are covalently bound to the carboxamido nitrogen on asparagines (Asn or N) residues (ionsource.com). Even if N-glycosylation is grouped as type of post-translational modification it often happens co-translationally when the protein is

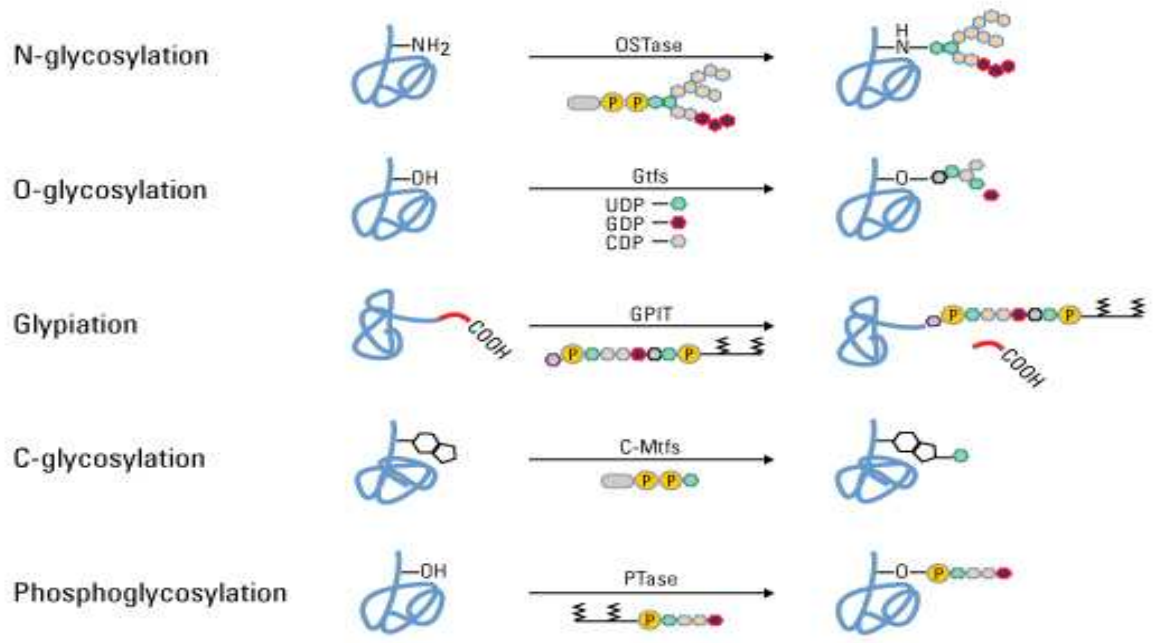
being translated not after the translation. N-linked glycosylation happens in the ER. (Thermo scientific)

2. **O-linked glycosylation**: it occurs between monosaccharide N- acetylgalactosamine and the hydroxyl group of amino acids serine or threonine (ionsource.com). O-linked glycosylation occur in ER, Golgi, cystosol and nucleus (Thermo scientific).

3. **Glypiation (GPI anchors):** it occurs when a protein is linked to a phospholipid by glycan core (Thermo scientific).

4. **C-linked glycosylation**: when mannose residue covalently attached to tryptophan residue C-linked glycosylation occurs (Uniport, 2011). It differs from other types of glycosylation because the reaction forms carbon-carbon bond not carbon-nitrogen or carbon-oxygen bond like the others do (Thermo scientific).

5. **Phosphoglycosylation**: occurs when phosphodiester bond binds glycan to serine. It is common in parasites and slim molds.



Source: http://www.piercenet.com/browse.cfm?fldID=4E12331D-5056-8A76-4E72-1C5A427505F1

FIGURE 4 Types of glycosylation

2.3.2.1 Glycosylation in bacteria

Before it was believed prokaryotes are not able to synthesize glycoprotein, many studies have shown proof to the contrary (Benz and Schmidth, 2002). The first protein glycosylation discovered in prokaryotes is in archaea which has a glycosylated surface layer (S-layer) protein (Hitchen et al., 2006).

Bacteria proteins can undergo both N-linked and O-linked glycosylation (Harald et al., 2010). In 2003, more than 70 bacterial glycoproteins were reported (Szymanski et al., 2003). Most of the glycoproteins that are present in bacteria are surface or secreted proteins this means that they affect how the bacteria interact with the environment (Schmidt et al., 2003). For example glycoproteins in Escherichia coli bacteria are secreted as autotransporters. These proteins are family of outer membrane proteins which are involved in toxication, invasion and aggregation. Glycoproteins are glycoslated by the addition of heptoses (Klemm et al., 2006).

2.3.2.2 Glycosylation in yeast

Both N-linked and O-linked glycosylation occur in yeast. Animal cell and yeast N-linked glycosylation are identical in their initial stage. However, O-linked glycosylation in yeast is different from higher eukaryotes.

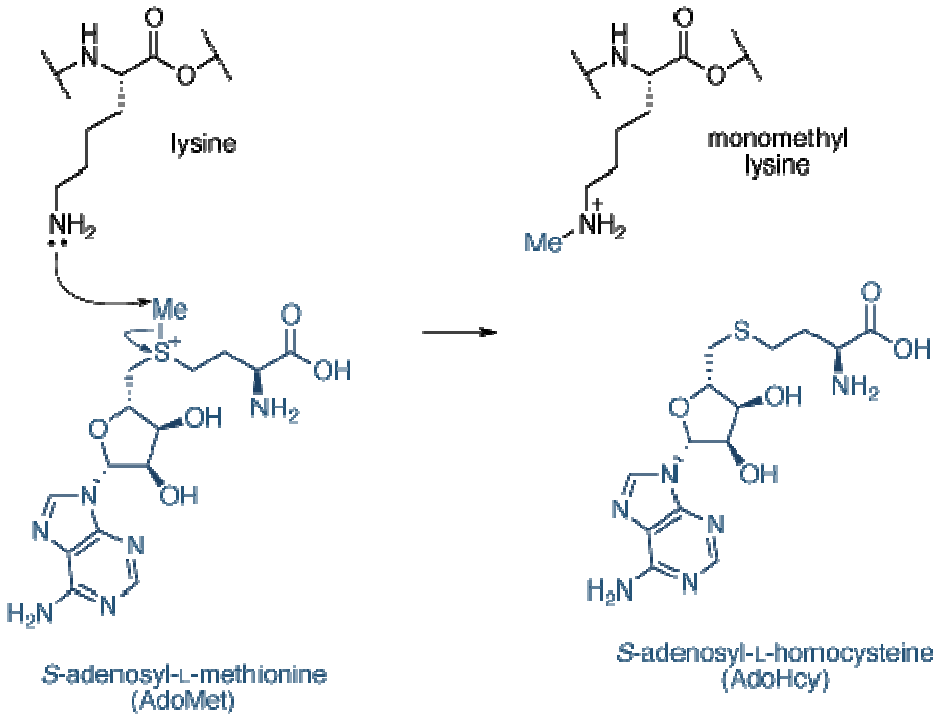
N-linked glycosylation starts in endoplasmic reticulum (ER). The first step is the transfer of dolichol-bound precursor oligosaccharide $Glc_3Man_9GlcNAc_2$ to nascent polypeptide by the help of enzyme called oligosaccharyltransferase. The glucose sugar that is found on oligosaccharide branch is removed by glucosidase I and glucosidase II. The removal of the glucose sugar will initiate a process called glycan-mediated chaperoning. If there are glycoproteins leaving the ER with different structure due to misfolding they will be reglycoslated and transported in to cytosol. Therefore this process is useful for quality control. Mannose sugars and mannosylphosphate transferases are added on the resulting Man8GlcNAc2-containing glycoprotein (Mochizuku et al., 2001 and Lehle et al., 1992)

2.3.3 Methylation

Methylation occurs in two ways, the first one is when one carbon methyl groups transfer to nitrogen (N-methylation) and the second one is when it is transferred to oxygen (O-methylation). The residues that are methylated on nitrogen include ϵ-amine of lysine, the imidazole ring of histidine, the guanidine moiety of arginine and the side chain of amide nitrogens of glutamine and asparagines. (Lee et al., 2005)

Histone lysine methylation occurs on histone H3 and histone H4. Lysine 4, 8, 14, 27, 36 and 79 are methylated in histone H3 and lysine 20 and 59 in histone H4 (Strahl and Allis, 2000). Lysine methyltransferases (KMTs) and lysine demethylases are the two enzymes which add or remove

methylation mark on lysine residues respectively (Zhang et al., 2012). The enzymes involved in lysine methylation were believed to be only histone specific but with enough evidence it has been found that they are not histone specific. For example the first non-histone protein methylation that was reported was methylation of p53 by KMT7. Most histone lysine modifications are involved in activation or repression of transcription (Lee et al., 2005).



Source: http://www.atdbio.com/content/56/Epigenetics#Histone-methylation

FIGURE 5. Lysine methylation mechanism, Mechanism of methylation of lysine by histone lysine methyltransferases (KMTs)

The above figure shows the conversion of S-adenosyl-L-methioine (AdoMet) which is the source of the methyl group in to S-adenosyl-L-homocysteine (AdoHcy)

Arginine methylation is common in eukaryotes (Bedford et al., 2007). It is found on both nuclear and cytoplasmic proteins. Protein arginine N-methyltransferase (PRMT) is the enzyme that catalyzes methylation of arginine (Bedford and Richard, 2005). Methylation of arginine plays important role in regulating protein-protein interaction, transcriptional regulation and signal transduction. Both histone and arginine methylations are irreversible (Lee et al., 2005).

There are indication that methylation is useful in protecting proteins in two ways. By blocking sites of ubiquitination it prevents from protein degradation and methylation reaction repair damaged protein molecules Mathews and Christopher, 2000).

2.3.4 N-Acetylation

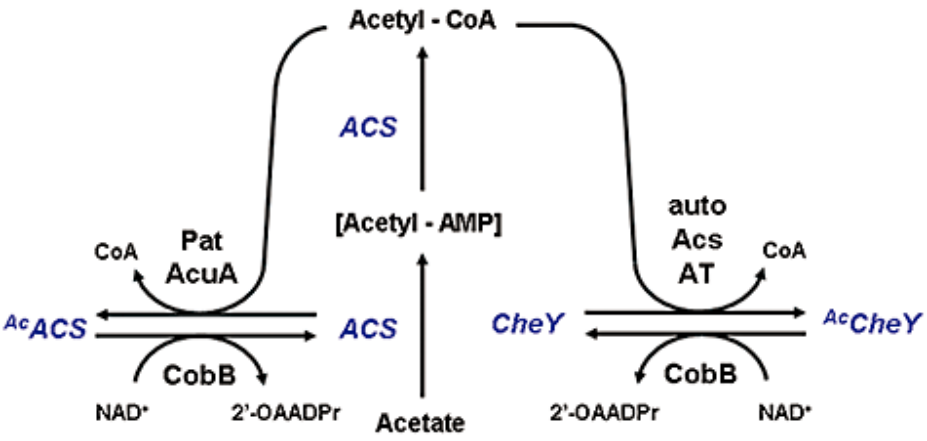
Acetylations occur for two specific biological purposes. The first one is in eukaryotic proteins acetylation occurs at N- terminal co transitionally and the other is acetylation of histones and transcriptional factors which affects chromatin structure and selective gene transcription (Walsh et al., 2006).

N-terminal acetylation occur after the cleavage of the N-terminal methionine by methionine aminopeptidase (MAP) the amino acid is replaced by acetyl group which is acetyl-CoA by using the enzyme called N-acetyltransferase (NAT) and the histone acetylation occur at ϵ-NH_2 of lysine on histone N-termini. Generally acetylation plays a great role in cell biology (Walsh et al., 2006).

2.3.4.1 Acetylation in bacteria

Acetylation is catalyzed by the enzyme N-acetyltransferase (NAT). There are three types of NAT termed NatA, NatB and NatC.

Bacterial acetylation occurs on N_C group of bacterial protein. Protein acetylation has been considered as eukaryotic phenomenon. Everything that is known about bacterial acetylation came from two proteins: the central metabolic enzyme ACS and the signaling protein CheY (Hu et al., 2010). ACS is controlled by reversible acetylation of single lysine residue. Reversible phosphorylation of aspartate and reversible acetylation of multiple lysine residues controls the ability of signaling protein CheY.



Source: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2010.07204.x/full

FIGURE 6: Reversible acetylation of ACS and CheY

In figure 6 reversible acetylation of ACS acetyl –CoA is used as acetyl donor. NAD^+ is used for deacetylation of ACS also CheY can be acetylated by ACS or some other acetyltransferases (AT) and the deacetylation is catalyzed by CobB.

2.3.5 Proteolytic cleavage

When the peptide bond between amino acids breaks proteolytic cleavage occur. The enzymes that carry out this process are called peptidases or proteases (CLC Bio et al., 2005). These enzymes can be classified in two groups the first one is based on their site of action when a single amino acid is removed from the termini (exopeptidases) and when internal peptide bond is cleaved (endopeptidases). The second is based on the nature of active site residues involved in mechanism. These are serine proteases, cysteine proteases, aspartyl proteases and zinc (metallo) proteases (Walsh et al., 2006).

Proteases act on protein substrate due to various reasons some of them are mentioned below

- After translation when N-terminal methionine residues are removed.
- Cleavage of proteins or peptides in order to be used as nutrients.
- During translocation when signal peptides are removed through membrane.

Proteolytic cleavage plays a diverse biological role such as signal transduction, proliferation, homeostasis, blood coagulation and fibrinolysis (Walsh et al., 2006).

2.4 Biological significance of PTMs

Rapid growth in understanding proteome has increased the knowledge of proteins. Also the focus on post-translational modifications and their effect on protein function have given a new insight in changes that are caused by post-translational modifications. For example signaling pathway from membrane to nucleus involves a series of protein modifications in response to external stimuli (Seo and Lee, 2012).

PTMs have greater importance because of their involvement in supervising gene expression, activation/ deactivation of enzymatic activity, protein stability or distraction and mediation of protein-protein interaction. Below Table 1 shows the function of post-translational modifications, the modification site and the amino acid residue change (Walsh et al., 2006).

TABLE 1 Biological function of post-translational modifications

PTM Type	Modified amino acid residues	Position	Effects
Acetylation	S K	N-term Anywhere	Protein stability, regulation of protein functions
Phosphorylation	Y,S,T,H,D	Anywhere	Regulation of protein activity, signaling
Cys oxidation disulfide bond glutathionylation sulfenic acid sulfinic acid	 C C C C	 Anywhere	Cell proliferation, differentiation
Acylation farnesylation myristoylation palmitoylation	 C G K C(S,T,K)	 Anywhere N-term Anywhere Anywhere	 Cellular localization to membrane
Glycosylation O-linked (O-Glc-NAc) N-linked	 S,T N	Anywhere	Cell-cell interaction and regulation of protein
Methylation monomethylation dimethylation trimethylation	 K K K	Anywhere	Regulation of gene expression, protein stability
Nitration S-Nitrosylation	Y C		Regulation of gene expression, protein stability
Ubiquitination Sumoylation	K K	Anywhere [ILFV]K.D	Signal transduction, DNA repair
Hydroxyproline Pyroglutamic acid	P Q	 N-term	Protein stability

Source: http://bmbreports.org/jbmb/jbmb_files/%5B37-1%5D0401271834_035-044.pdf

2.5 Variation of post-translational modification sites and disease

Variations are changes in DNA and RNA. They play a vital role in human proteome. Changes in protein conformation that leads to enzymes that are non-functioning or differently functioning and unusual protein structure can be a result of variation. For example Duchenne muscular

dystrophy is an inability to produce dystrophin protein. The lack of this protein cause abnormal or no cell structure organization and Huntington's disease is a progressively deterioration of the nervous system caused by abnormal proteins. So we can conclude that variations are a cause of many human diseases. Also variation affects post-translational modification sites (Li et al., 2010).

There have been studies about the variations of post translational modification and their contributions to human disease. Some of these studies are mentioned below.

- In prion protein (PrP) gene a heterozygous T183A has undergone mutation which has resulted the removal of N-linked glycoslation of PrP. This variation was detected in a patient with spongiform encephalopathy. Some of the symptoms are early-onset dementia as the predominant sign, along with global cerebral atrophy and hypometabolism there are also neurological signs which occur at late stage of the disease including cerebellar ataxia and EEG abnormalities (Grasbon et al., 2004).

- On androgen receptor acetylation occur on lysine residue the loss of this acetylation site has been linked to Kennedy's disease which is inherited neurodegenerative disorder. The variation of lysine residues at 630, 632 and 633 to alanine markedly delays ligand-dependent nuclear translocation in androgen receptor (Thomas et al., 2004).

- Familial advanced sleep phase syndrome (FASPS) is caused by variation in binding region of hPER2 casein kinase Iepsilon (CKIepsilon) from serine to glycine at a phosphorylation site (Toh et al., 2001).

3 OBJECTIVES

The main objective of this study was to analyze the variations at post-translational modification and their relevance to diseases. There are two groups in the PTM sites with variations disease causing variations and not disease causing variations. By considering this

- Analyze if certain types of PTMs have been enriched or depleted in the two groups
- Study how well the variations at PTMs are conserved

4 MATERIALS AND METHOD

4.1 Materials

In this study two data sets were used the single nucleotide polymorphism data set and the data downloaded from human protein reference database (HPRD) which contains 93,710 experimentally verified PTM sites.

The SNP data set contain a total of 32,003 variations of which 14,610 were pathogenic variations and 17,393 were neutral missense variations. The missense pathogenic variations were built from PhenCode database (Giardine et al., 2007) (downloaded in June 2009), registries in IDbases (Piirilä et al., 2006) and from 18 individual LSDBs. The neutral missense variations are obtained from dbSNP database (Sherry et al., 2001]) build 131.

4.1.1 Databases

The following tools were used to analyze the data.

- DRUMS: is a search engine for human disease related genetic variations. It collects the genotype – phenotype data from LSDBs and makes it available for users.
- WAVe: is a web-based application that integrates locus specific databases and gathers available genomic variations in a single working environment.
- Locus specific mutation database: this database is available on HGVS website.
- ConSurf: a bioinformatics tool used for estimating conservation of amino acid in proteins.
- UniProtKB: it contains a collection of proteins with their functional information.

4.2 Methods

4.2.1 Filtering the SNP data set

The SNP data were matched with the experimentally verified post-translational modification sites in order to separate the variants which have undergone post-translational modification. python script was made to do the matching.

4.2.2 Matching post translational modifications with amino acid substitutions

By filtering the SNP data a set of human post-translational modifications have been created. With the purpose of investigating the relationship between post-translational modification and amino acid substitution exact matching has been done in the sites were the substitution occurred at a modification sites.

4.2.3 Disease related and not disease related variations

As mentioned above the SNP data set contains two kinds of variation the pathogenic and neutral missense variations. In each of these variations the position which the amino acid substitution occur have been analyzed to see if this variations are disease related at this exact position. In order to accomplish this, three databases such as Locus Specific Mutation Database, DRUMS and WAVe were used. The databases were searched manually for each variation.

4.2.4 Conservation score analysis

The position-specific conservation score for the variants were calculated using the ConSurf server. First by using CSI-BLAST close homologous sequences were obtained then a multiple alignment was constructed using MAFFT. Bayesian method was used in calculating the position specific conservation score. The conservation score values have nine scales from one to nine one being the least conserved (Ashkenazy et al., 2010).

4.2.5 Statistical analysis

In the statistical analysis Excel and R programs were used. The statistical methods that were applied are hyper geometric distribution and T-test.

4.2.5.1 Hypergeometric distribution

Hypergeometric distribution is a probability distribution of the number of successes in a hyper geometric experiment. In hyper geometric experiment the researcher select randomly without replacement from a finite population and every item in the population can be categorized as success or failure.

The Hypergeometric formula is:

$$h(x;N,n,k) = [{}_kC_x]\,[{}_{N-K}C_{n-x}] / [{}_NC_n] \qquad /1/$$

Notations

- N: The number of items in the neutral and pathogenic variation dataset.
- k: The number of items in the neutral and pathogenic variation dataset that are classified as successes.
- n: The number of items in the sample .
- x: The number of items in the sample that are classified as successes.
- ${}_kC_x$: The number of combinations of k things, taken x at a time.
- h(x; N, n, k): hypergeometric probability - the probability that an n-trial hypergeometric experiment results in exactly x successes, when the neutral and pathogenic variation dataset consists of N items, k of which are classified as successes.

Suppose let's say 84 variants were selected randomly without replacement from disease causing data set. What is the probability of getting exactly 37 variants which has phosphorylation site? By using R the hypergeometric distribution was calculated. This will show if certain types of PTMs are depleted or enriched in any of the two datasets.

The hypergeometric tests were done by comparing the 32,003 variations (the primary variation data) with both the disease related and not disease related variations.

4.2.5.2 T-test

To compare the means of disease related variations and not-disease related variations the T-test was used

Assumptions

- the distribution of the variants are normal
- samples are independent
- they have equal variance

The sample size of the two data sets were different by calculating the ratio of two data set and multiplying one data set with ratio normalization was done to make the total number equal.

Null Hypothesis:

The type of variation (disease or not disease related) has no effect on the type of post-translational modifications.

Alternative Hypothesis:

The type of variation (disease or not disease related) has effect on the type of post-translational modifications

5 RESULTS

5.1 Variations with post-translational modification sites

There were a total of 32,003 variations which include pathogenic and neutral variations. By matching the post-translational dataset and variation data set 35,103 variations with both PTM sites and variation site were found. From the 28 types of post-translational modifications found more than half were a phosphorylation modification.

TABLE 2 Types and total number of post-translational modification with variation sites

Post-translational modification	**Total number of sites**	**Post-translational modification**	**Total number of sites**
Phosphorylation	28,191	Carboxylation	25
Acetylation	1951	Prenylation	14
Glycosylation	1852	Myristoylation	22
Proteolytic cleavage	1053	Transgutamination	8
Disulfide Bridge	1214	Glycosyl	11
Dephosphorylation	206	Nitration	20
Farnesylation	2	Amidation	8
Methylation	94	Glutathionylation	6
SUMOylation	145	Neddylation	9
Palmitoylation	78	Hydroxylation	10
S- Nitrosylation	46	Alkylation	5
Sulfation	24	ADP Ribosylation	7
Ubiquitination	58	Deacetylation	4
Glycation	39	Deacylation	1

5.2 Amino acid substitutions

From 35,103 variations with post-translational modification the number of amino acid substitution that lie directly on the modification site are 242 and from this 96 were pathogenic variations and 146 were neutral variations.

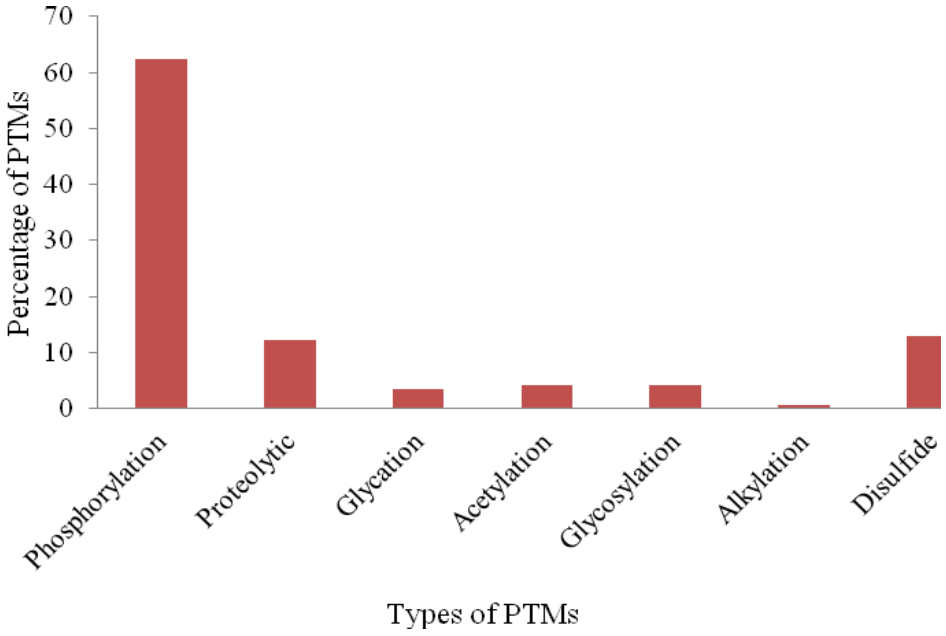


FIGURE 7. The distribution of PTMs in the neutral variation data set

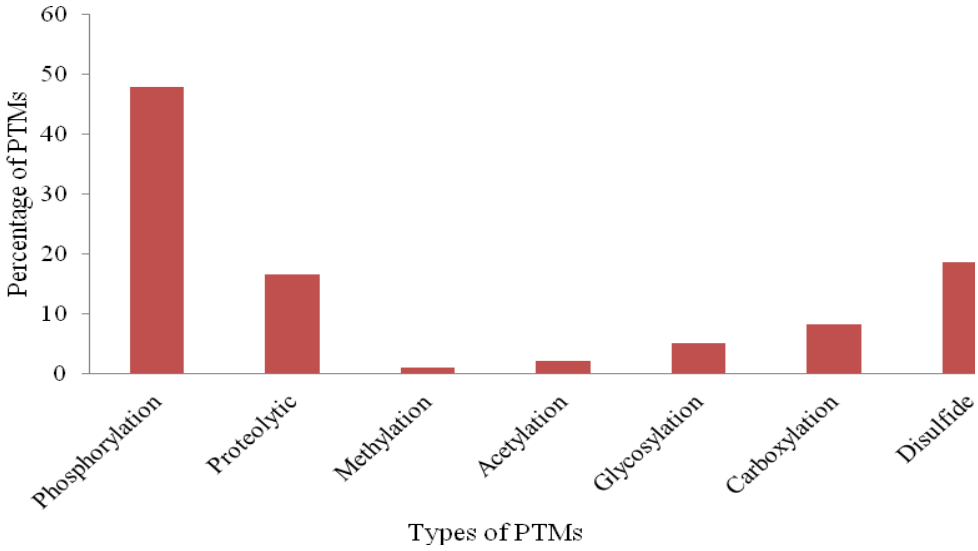


FIGURE 8. The distribution of PTMs in pathogenic variations

5.3 Mutations of post-translational modification sites

The 242 proteins which have post-translational modification sites have also undergone through mutation which is single nucleotide polymorphism. From this variants it was found that in the

neutral variation data set categories which include 146 variation 58 of them were found to be disease causing, 19 of them are not related to any disease and for the 69 variations no information were obtained in any of the database used on their relation to disease. In the case of pathogenic variations from 96 variations 84 were found to be disease related, 1 variation was not related to any disease and on 11 variations there were no information obtained about their relation to any disease.

TABLE 3 Total summaries of neutral and pathogenic variations and their relation to disease

	Disease causing	Not disease causing	No information
Neutral Variations	58	19	69
Pathogenic Variations	84	1	11

For conservation score and statistical analysis two categories of variations were chosen

- Disease causing variations from pathogenic data set, and
- From neutral variation not disease causing variation plus the variation which no information has been obtained are added together and considered as benign variations.

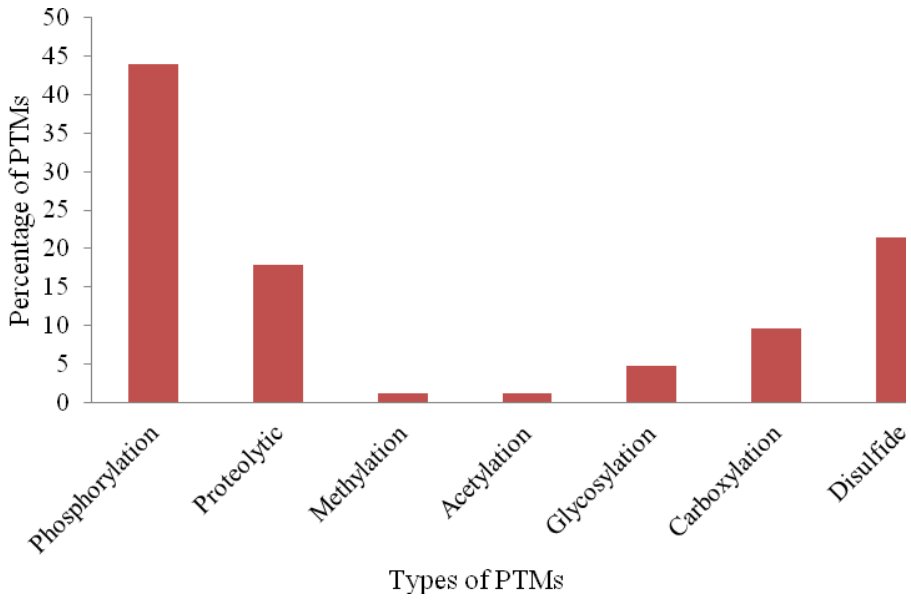


FIGURE 9. The distribution of PTMs in disease causing variations

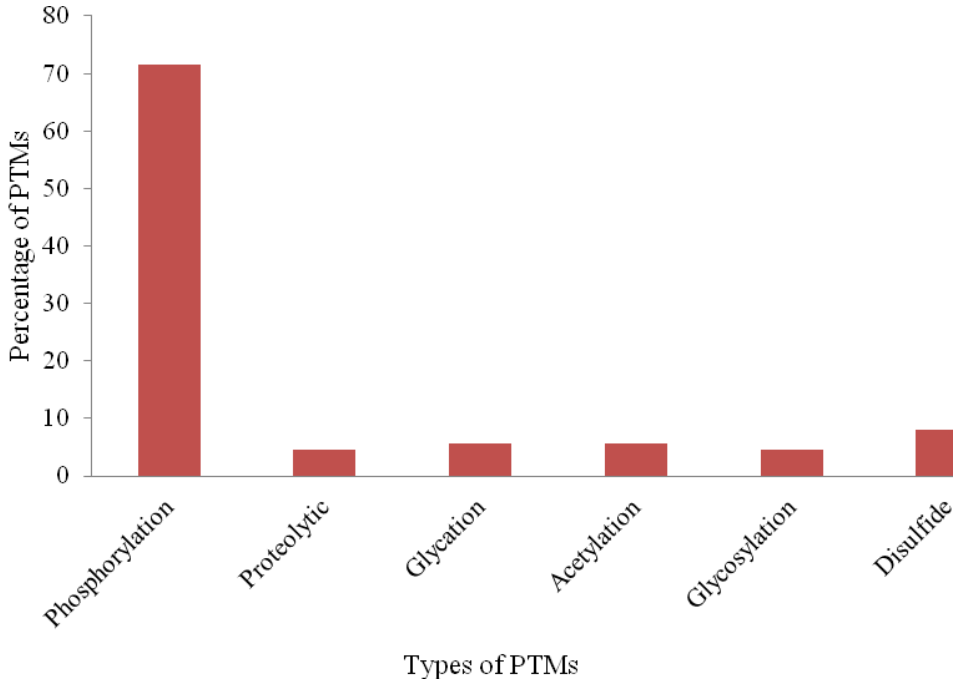


FIGURE 10. The distribution of PTMs in benign variations

5.4 Conservation of post-translational modification sites

From ConSurf server analysis the conservation score of both disease and benign variation has been obtained. In disease causing variation 69 variations (82.15%) are highly conserved, 10 variations (11.90%) are average and 5 variations (5.95%) are not conserved or they are variable.

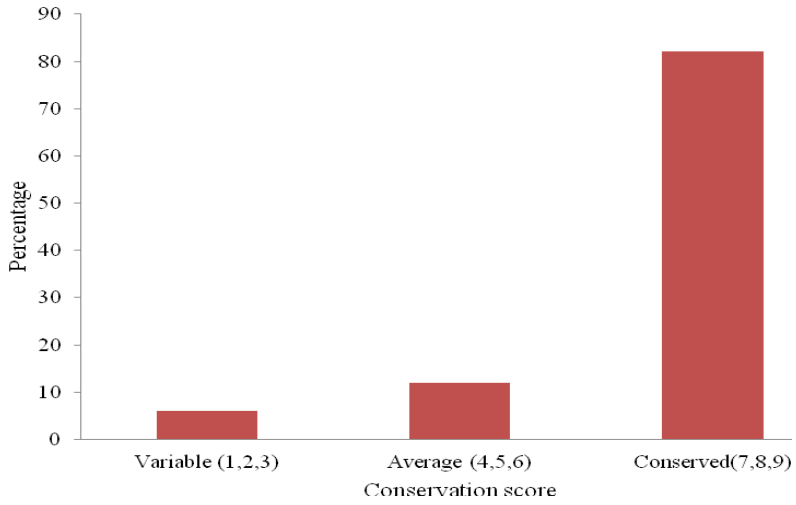


FIGURE 11. The distribution of conservation score in disease causing variations

For benign varations 35 varations (39.8%) are found to be not conserved, 21 varations (23.9%) are average and 32 varations (36.4%) are highly conserved as shown below in table

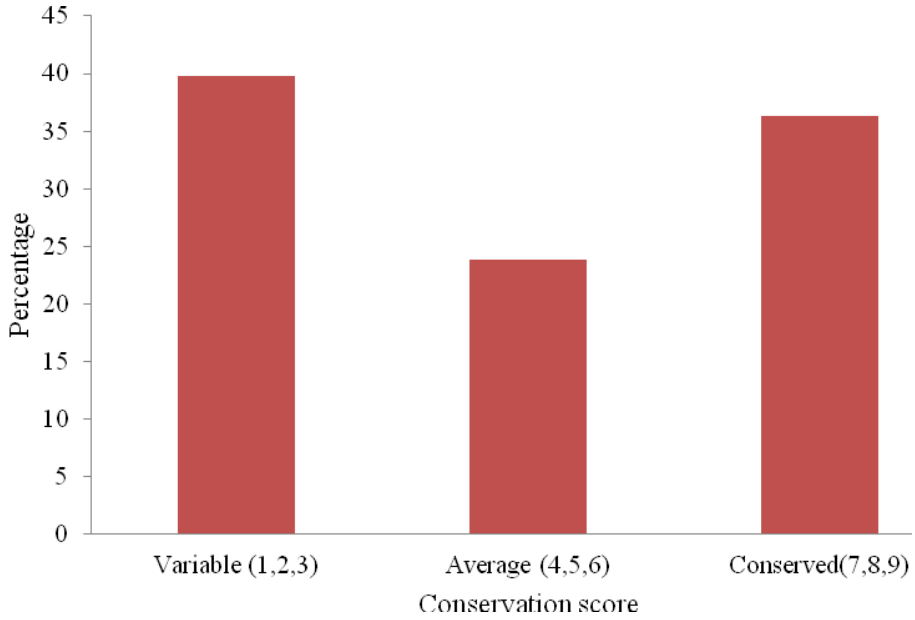


FIGURE 12. The distribution of conservation score in benign variations

5.5 Hypergeometric test result

As it can be seen from the table below the hypergeometric distribution was calculated for disease causing variations and benign variations. In the table the hyper geometric probability P(X=x) was calculated which shows if certain types of PTMs are enriched or depleted in any of the categories.

TABLE 4. Hypergeometric distributions of disease causing and benign variations

Diseases causing variations						**Benign variations**				
Types of PTM	N	K	n	X	P(X=x)	N	K	n	X	P(X=x)
Phosphorylation	35103	28191	84	37	1.74e-13	35103	28191	88	63	0.014
Proteolytic	35103	1053	84	15	2.43e-08	35103	1053	88	4	0.147
Methylation	35103	94	84	1	0.181	35103	94	88	0	0.789
Acetylation	35103	1951	84	1	0.041	35103	1951	88	5	0.181
Glycosylation	35103	1852	84	4	0.196	35103	1852	88	4	0.191

Carboxylation	35103	25	84	8	7.95e-16	35103	25	88	0	0.94
Disulfide	35103	1214	84	18	4.26e-10	35103	1214	88	7	0.022
Glycation	35103	39	84	0	0.911	35103	39	88	5	4.68e-08

The above table shows the hypergeometric probability P(X=x) of disease causing variations and benign variations. Form the result shown in Table 4 most of the PTMs in disease causing variations are enriched except glycation. This can be viewed on figure 13 the red bars indicates the hyper geometric probability P(X=x) of disease causing variation and the blue ones indicate the hyper geometric probability P(X=x) of benign variations. For example the value of P(X=x) for glycation in benign variations is so small that it cannot be seen on the diagram.

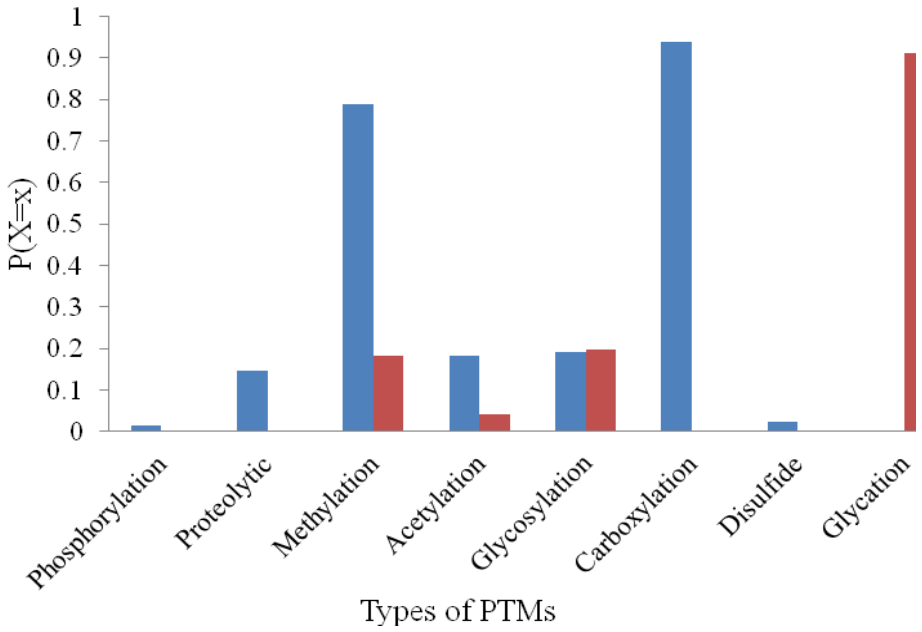


FIGURE 13. Hypergeometric distribution of disease causing variations and benign variations data

5.6 T-test result

Table 6 shows the data used to calculate the t-test in benign variations column the values stated are the normalized value. The t-test has given a p-value of 0.9997 which is greater than the significance level this will indicate the data in this study may not be sufficient enough to reject the null hypothesis.

TABLE 6. Total numbers of PTMs in both diseases related and not disease related variations

Post-translational modifications	**Disease related**	**Benign variations**
Phosphorylation	37	60.14
Proteolytic	15	3.82
Glycation	0	4.78
Acetylation	1	4.78
Glycosylation	4	3.82
Disulfide	18	6.69
Methylation	1	0
Carboxylation	8	0

Two Sample t-test
data: a and b
t = -4e-04, df = 14, p-value = 0.9997
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-18.06656 18.05906
Sample estimates:
Mean of x mean of y
10.50000 10.50375

6 DISCUSSIONS

Post-translational modifications are crucial in changing physiochemical properties of proteins. These alternations often may result initiation of important process such as gene expression, oxidative regulation of protein and cell-to-cell interaction. The function of such biological process becomes interrupted when the modification sites become mutated. This study has analyzed the post-translational modification sites and disease associated variations that occur on the modification sites.

In this study, neutral and pathogenic variations which have undergone post-translational modification have been analyzed. It was found that in both variation groups phosphorylation has the highest percentage as seen in Figure 7 and 8.

There have been incidents in which variation of post-translational modification sites are involved in disease. Both neutral and pathogenic variations were evaluated and it was found that from the pathogenic variations 87.5% of the variations are diseases causing and from the neutral variation only 39.8% were disease related.

When analyzing the conservation score of disease causing and benign variations first the protein sequence in fasta format is submitted to ConSurf server. The result is a text document which includes the conservation score of each amino acid. As it can be seen in Figures 11 and 12 the disease causing variations were more conserved than the benign variations.

The hyper geometric distribution test has revealed that methylation, glycosylation and acetylation are enriched in disease causing variations. This may indicate that disease-causing variations are more likely to affect post-translational modifications than benign variations.

The p-value obtained from the t-test were 0.9997 which is greater than the significance level (0.05) this indicates that the data may not be sufficiently persuasive to reject the null hypothesis which states the type of variation (disease or not disease related) has no effect on the type of post translational modifications.

There have been studies which have analyzed variations at PTM sites. A study carried out by Radivojac and his colleague's who analyzed gain and loss of phosphorylation sites have concluded that the variations at the phosphorylation sites are more likely a mechanism in cancer. In another study Li et al. looked in to the loss of PTM sites in disease and have found that disease causing variation were highly conserved (Li et al., 2010).

Because of the data used in this study there may be problems that lead to ascertainment bias. The first one is the variation data which contains post-translational sites is heavily skewed towards phosphorylation which is more than 50%. The reason behind this could be phosphorylation has

been discovered more frequently than the other post-translational modifications because of discovery techniques like mass spectrometry has been effective and its high influence on biological processes has also played a big role. The second problem can arise from the variations which are not disease causing. This data may probably contain undiscovered disease mutations.

7 CONCLUSIONS

The main objective of this study was to investigate post-translational modification sites and their effect on disease. Based on this from the 242 modification sites which contain 96 pathogenic variations and 146 neutral variations the disease causing variation has been filtered out by the help of biological data bases like DRUMs, WAVe, locus specific database and locus specific mutation database. 84 variations form pathogenic variation and 58 variations form neutral variations were found to be disease-causing variants. The other objective of the study was to see the distribution of post-translational modification sites in both disease causing and benign variations. Histograms, hyper geometric test and t-test were used to do the statistical analysis. Finally to see how these modification sites were conserved ConSurf were used to calculate the conservation score.

The result of this study indicated that the disease associated mutations that occur in modification sites change the protein functions by disrupting the post-translational modification sites. It was found that 87.5% of the pathogenic variants and 39.8% of neutral variations are disease causing. From this one can conclude that post-translational modifications are not the major cause of disease. It was also found that disease causing variations are highly conserved. The hyper geometric test has shown which type of post-translational modifications are enriched or depleted in disease and benign variations.

A further study could assess the following points for example this study has shown that in neutral variations category from 146 variations 58 variations were disease causing and for the 69 variations in this category there were no information found on any of the databases used about their relation to disease this indicates that further study can be done to find more about this specific variations. And another point to see further will be what kind of molecular and cellular functions were disrupted because of the disease causing variations.

8 REFERENCES

Aletta JM, Cimato TR and Ettinger MJ .1998 .Protein methylation: a signal event in post-translational modification. Trends Biochem Sci 23:89–91

Ashkenazy H., Erez E., Martz E., Pupko T. and Ben-Tal N. 2010. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucl.Acids Res.

Bedford M. 2007. Arginine metylation at glance. J Cell Sci 120, 4243-4246.

Bedford M and Richard S. 2005. Arginine methylation an emerging regulator of protein function. *Mol. Cell* 18, 263-272.

Benz, I. and Schmidt, M. A. 2002. Never say never again: protein glycosylation in pathogenic bacteria. Molecular Microbiology, 45: 267–276

Birkenmeier G, Stegemann C, Hoffmann R, Günther R and Huse K. 2010. Posttranslational Modification of Human Glyoxalase 1 Indicates Redox-Dependent Regulation. PLoS ONE 5(4): e10399. doi:10.1371/journal.pone.0010399

Chen L, Li Z, Zwolinska AK, Smith MA, Cross B, Koomen J and Yuan ZM.2010.MDM2 recruitment of lysine methyltransferases regulates p53 transcriptional output. EMBO J; 29:2538-2552

Chuikov S, Kurash JK,Wilson JR, Xiao B, Justin N, Ivanov GS and McKinney K. 2004.Regulation of p53 activity through lysine methylation. Nature; 432:353-360.

Cozzone J.2005. Role of protein phosphorylation on serine/threonine and tyrosine in the virulence of bacterial pathogens. J Mol Microbiol Biotechnol 2005; 9(3/4): 198–213

GEN2PHEN. Locus specific data bases (LSDBs). Retrieved March, 2012 from gen2phen.org: http://www.gen2phen.org/data/lsdbs

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, and others. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat* 28:554–562.

Grasbon-Frodl E, Lorenz H, Mann U, Nitsch RM, Windl O and Kretzschmar HA. 2004. Loss of glycosylation associated with the T183A mutation in human prion disease. Acta Neuropathol ;108(6):476-84

Grangeasse C, Cozzone J and Deutscher J. 2007.Tyrosine phosphorylation: an emerging regulatory device of bacterial physiology. Trends Biochem Sci2007; 32(2): 86–94

Harald N. & Christine M. 2010.Protein glycosylation in bacteria: sweeter than ever. Nature Reviews Microbiology 8, 765-778

Hardie D. 1999. Plant protein serine/threonine kinases: classification and functions. Annu. Rev. Plant Physiol. Plant Mol. Biol. 50:97–131

Hitchen G. and Anne Dell. 2006. Bacterial glycoproteomics. Microbilogy vol. 152 no. 6 **1575-1580**

Horaitis R., Talbot C., Martin T., and Mao J. 2011. Locus Specific Mutation Databases. Human genome variation society.

Huang J, Perez-Burgos L, Placek BJ, Sengupta R, Richter M, Dorsey JA, and KubicekS,.2006 Repression of p53 activity by Smyd2-mediated methylation. Nature;444:629-632.

Huber SC, Huber JL, McMichael RW.1994. Control of plant enzyme activity by reversible protein Phosphorylation. Int.Rev. Cytol. 149:47–98

Human protein reference database.HPRD_FLAT_FILES_041310.tar.gz created 04-13-10 Apr 13, 2010. Johns Hopkins University and the Institute of Bioinformatics

Hu, L. I., Lima, B. P. and Wolfe, A. J. (2010), Bacterial protein acetylation: the dawning of a new age. Molecular Microbiology, 77: 15–21. doi: 10.1111/j.1365-2958.2010.07204.x

International Human Genome sequence Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature.431,931-45

Jensen O.N. (2004) Modifcation-specific proteomics: Characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Bio1.8, 33-41

Jenuwein T and Allis CD .2001. Translating the histone code. *Science* 293:1074–1080

Lilia M., Predrag Radivojac,Celeste J., Timothy R., Jason G., Zoran Obradovic,and Keith A.2004. The importance of intrinsic disorder for protein Phosphorylation. Nucl. Acids Res. 32(3):1037-1049

Lopes, P., Dalgleish, R. and Oliveira, J. L. (2011), WAVe: web analysis of the variome. Human Mutation, 32.

MacKintosh C, Coggins J, Cohen P. 1991. Plant protein phosphatase: sub cellular distribution, detection of protein phosphatase 2C and identification of protein phosphatase 2A as the major quinate dehydrogenase phosphatase. Biochem. J. 273:733–38

Manai M, Cozzone J.1982. Endogenous protein phosphorylation in Escherichia coli extracts. Biochem Biophys Res Commun ; 107(3): 981–988

Mann M. and Jensen ON. 2003. Proteomic analysis of post-translational modifications. Nature Biotechnology 21, 255 – 261

Mathews, Christopher K.2000.Biochemistry.3rd edition. New York: Addison Wesley Longman, Inc.

Mescher, M.F., Strominger, J.L., and Watson, S.W. (1974) Protein and carbohydrate composition of the cell envelope ofHalobacterium salinarum. *J* Bacteriol 120: 945–954

Mochizuku S., Hamato N., Hirose M., Miyano K., Ohatani W., Kameyama S., Kuwae S., Tokuyama T., Ohi H., (2001)

Larsen MR, Trelle MB, Thingholm TE and Jensen ON. 2006. Analysis of posttranslational modifications of proteins by tandem mass spectrometry. BioTechniques, Vol. 40, No. 6, pp. 790–798

Lee D.,Teyssier C., Strahl B and Stallcup M. 2005.Role of Protein Methylation in Regulation of Transcription. Endocrine Reviews April 1, vol. 26 no. 2 147-170

Lehle L.1992. Protein glycosylation in yeast. Biomedical and Life Sciences 133-134 vol61

Li S, Iakoucheva LM, Mooney SD and Radivojac P. 2010. Loss of Post-translational modification sites in disease. Pac Symp Biocomput. : 337-47.

Magrane M. and Consorthium U. 2011. Uniprot Knowledgebase: A hub of integrated protein data. Database, 2011

N- And O- linked protein Glycosylation. Retrieved June, 2012 from Ionsource.com: http://www.piercenet.com/browse.cfm?fldID=7CE3FCF5-0DA0-4378-A513-2E35E5E3B49B

Olsson A., Manzl C., Strasser A. and Villunger A. 2007. How important are post-translational modifications in p53 for selectivity in target-gene transcription and tumour suppression? Nature Cell Death and Differentiation 14, 1561–1575

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). *Hum Mutat* 27:1200–1208.

Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW and Mooney SD.2008.Gain and loss of Phosphorylation sites in human cancer. Bioinformatics 24 (16):i241-i247.

Roepstorff P. and Fohlman J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed. Mass Spectrom. 11:601.

Seo J and Lee KJ. 2004. Post-translational modifcatio and their biological functions: proteomic analysis and systematic approach. J Biochem Mol Biol. Jan 31; 37(1):35-44

Shen EC, Henry MF, Weiss VH, Valentini SR, Silver PA and Lee MS .1998. Arginine methylation facilitates the nuclear export of hnRNP proteins. Genes Dev 12:679–691

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation.*Nucleic Acids Res* 29:308–311.

Shi X, Kachirskaia I, Yamaguchi H, West LE, Wen H, Wang EW and Dutta S. 2007. Modulation of p53 functions by SET8-mediated methylation at lysine 382. Mol Cell; 27:636-646.

Stat Trek: Teach yourself statistics .2012.
http://stattrek.com/onlinecalculator/hypergeometric.aspx

Steen H. and Mann M. 2004. The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5:699-711.

Strahl B and Allis C. 2000 .The language of covalent histone modifications. *Nature*403:41–45

Stulke J.2010. More than just activity control: Phosphorylation may control all aspects of a protein's properties. Mol Microbiol 2010; 77(2): 273–275.

Szymanski M . Logan M, Linton D, Wren W. 2003. Campylobacter--a tale of two protein glycosylation systems Trends Microbiol. 11: 233 [PMID: 12781527]

Tejaswita M. and Amrita K. 2011. Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease .Journal of Amino Acids, vol. 2011, Article ID 207691, 13 pages.

Thermo scientific. Overview of post-translational modifications (PTMs). Retrieved June 10, 2012 from piercenet.com: http://www.piercenet.com/browse.cfm?fldID=7CE3FCF5-0DA0-4378-A513-2E35E5E3B49B

Thomas M, Dadgar N, Aphale A, Harrell JM, Kunkel R, Pratt WB and Lieberman AP. 2004. Androgen receptor acetylation site mutations cause trafficking defects, misfolding, and aggregation similar to expanded glutamine tracts. J Biol Chem.; 279(9):8389-95.

Toh KL, Jones CR, He Y, Eide EJ, Hinz WA, Virshup DM, Ptácek LJ and Fu YH. 2001. An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. Science. ; 291(5506):1040-3

Van Holde KE .1989. Chromatin. New York: Springer-Verlag; 111–148

Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, Fieschi C, Boisson-Dupuis S, Alcais A, Filipe-Santos O, Bustamante J, de Beaucoudrey L, Al-Mohsen I, Al-Hajjar S, Al-Ghonaium A, Adimi P, Mirsaeidi M, Khalilzadeh S, Rosenzweig S, de la Calle Martin O, Bauer TR, Puck JM, Ochs HD, Furthner D, Engelhorn C,Belohradsky B, Mansouri D, Holland SM, Schreiber RD, Abel L, Cooper DN, Soudais C and Casanova JL.2005.

Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. Nature Genetics 37, 692 - 700

Universal Protein Resource (UniProt) manual.2011.Glycoslation.web http://www.uniprot.org/manual/carbohyd

Vogt G, Vogt B, Chuzhanova N, Julenius K, Cooper DN, and Casanova JL. 2007. Gain-of-glycosylation mutations. Curr Opin Genet Dev.245-51.

Walsh C. 2006.Posttranslational Modification of Proteins.Roberts and Co; 7-24

Walsh CT, Garneau-Tsodikova S and Gatto GJ Jr. 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl.; 44(45):7342-72.

Zhang X., Wen H. and Shi X.2012. Lysine metylation: beyond histones. Acta Biochim Biophys Sin (2012)44 (1): 14-27.

Zhang ZY.2005. Functional studies of protein tyrosine phosphates with chemical approaches. Biochim Biophys Acta 2005; 1754(1/2): 100–107

Zuofeng Li, 2011. DRUM. Retrived March, 2012 from scbit.org: http://www.scbit.org/dbmi/drums/about.html

9 APPENDICES

TABLE 7. Disease causing variations

Gene symbol	Residue change	Modifications	Conservation Score
APP	L705V	Proteolytic	9
APP	A713T	Proteolytic	9
APP	T714A	Proteolytic	9
APP	T714I	Proteolytic	9
APP	D694N	Proteolytic	7
BRCA1	R1204I	Phosphorylation	7
BRCA1	S1217Y	Phosphorylation	9
BRCA1	S1187I	Phosphorylation	2
BRCA1	P1150S	Phosphorylation	8
BRCA1	R866Q	Glycosylation	9
BRCA1	G960D	Phosphorylation	4
CTNNB1	S45F	Phosphorylation	9
CTNNB1	S45P	Phosphorylation	9
CTNNB1	S33F	Phosphorylation	9
CTNNB1	S33L	Phosphorylation	9
CTNNB1	S33Y	Phosphorylation	9
CTNNB1	S37C	Phosphorylation	9
CTNNB1	S37F	Phosphorylation	9
CTNNB1	S37Y	Phosphorylation	9
CTNNB1	S37A	Phosphorylation	9
CTNNB1	T41A	Phosphorylation	9
CTNNB1	T41I	Phosphorylation	9
COL2A1	G909C	Proteolytic	9
COL2A1	G981S	Proteolytic	9
EDN3	Y127C	Proteolytic	9
FGFR1	C277Y	Proteolytic	9
FUS	R244C	Methylation	8
GH1	Q117L	Phosphorylation	7
MAPT	S622N	Phosphorylation	9
MPZ	K236E	Phosphorylation	4
NFKBIA	S32I	Phosphorylation	9
RAF1	T491I	Phosphorylation	9
RAF1	T491R	Phosphorylation	9
RAF1	S259A	Phosphorylation	9
RAF1	S259F	Phosphorylation	9
RAF1	S257L	Phosphorylation	9

RAF1	T260R	Phosphorylation	8
PRNP	E196K	Glycosylation	5
PROC	E62A	Carboxylation	9
PTPN11	Y62D	Phosphorylation	2
PTPN11	Y63C	Phosphorylation	6
PTPN11	T2I	Acetylation	5
MECP2	Y120D	Phosphorylation	6
BTK	Y361C	Phosphorylation	7
EMD	S54F	Phosphorylation	7
GLA	N215S	Glycosylation	7
FGD1	S205I	Phosphorylation	1
F9	E79D	Carboxylation	6
F9	E53A	Carboxylation	9
F9	E54G	Carboxylation	9
F9	E66V	Carboxylation	9
F9	E67)	Carboxylation	9
F9	E73K	Carboxylation	9
F9	E73V	Carboxylation	9
F9	R191C	Proteolytic	2
F9	R191H	Proteolytic	2
F9	R226G	Proteolytic	9
F9	R226Q	Proteolytic	9
F9	R226W	Proteolytic	9
F9	C134Y	Disulfide	9
F9	C155F	Disulfide	9
F9	C170F	Disulfide	9
IDS	N115Y	Glycosylation	9
DNM2	S619L	Phosphorylation	7
DNM2	S619W	Phosphorylation	7
FGF23	R179Q	Proteolytic	6
EDNRB	S305N	Phosphorylation	4
KCNJ1	S219R	Phosphorylation	9
HSPB1	S135F	Phosphorylation	9
RHO	C110F	Disulfide	9
RHO	C110Y	Disulfide	9
AGA	C163S	Disulfide	6
NDP	C39R	Disulfide	8
NDP	C65W	Disulfide	8
NDP	C69S	Disulfide	8
NDP	C65Y	Disulfide	8

HAMP	C70R	Disulfide	8
FSHB	C69G	Disulfide	9
AGA	C306R	Disulfide	8
NDP	C96W	Disulfide	8
NDP	C96Y	Disulfide	8
NDP	C110G	Disulfide	8
NDP	C110R	Disulfide	8
HAMP	C78Y	Disulfide	8

TABLE 8. Benign variations

Gene symbol	Position	aa	Modifications	Conservation score
BGLAP	94	R	Proteolytic	7
DSG2	903	T	Phosphorylation	5
AHSG	256	T	Glycosylation	1
HBB	2	V	Glycation	8
HBB	9	K	Glycation	7
HBB	9	K	Glycation	7
IGF2R	1107	T	Phosphorylation	6
IGF1	115	A	Proteolytic	1
LMNA	10	T	Phosphorylation	1
PTPRC	191	N	Glycosylation	4
SPP1	224	S	Phosphorylation	2
DDX5	480	S	Phosphorylation	1
SEMG1	372	R	Proteolytic	4
CD180	430	T	Phosphorylation	9
AKR7A2	255	S	Phosphorylation	1
PDLIM5	136	S	Phosphorylation	3
EML4	978	S	Phosphorylation	1
PDCD6IP	730	S	Phosphorylation	1
CCDC99	508	Y	Phosphorylation	1
HBB	73	S	Phosphorylation	3
HBB	5	T	Phosphorylation	5
ITIH2	60	S	Phosphorylation	3
KRT8	26	T	Phosphorylation	1
LMNA	22	S	Phosphorylation	9
LMNA	588	S	Phosphorylation	5
LMNA	583	S	Phosphorylation	8
MDH2	301	K	Acetylation	3
MBD1	505	T	Phosphorylation	8

NF1	2502	S	Phosphorylation	1
CTTN	87	K	Acetylation	4
SPP1	210	S	Phosphorylation	9
PAM	860	S	Phosphorylation	7
ARID4A	864	S	Phosphorylation	9
RPLP0	285	T	Phosphorylation	1
RANBP2	1118	S	Phosphorylation	1
DOCK1	1857	T	Phosphorylation	2
ADD3	651	S	Phosphorylation	5
ADD3	650	S	Phosphorylation	4
HIP1	35	S	Phosphorylation	7
CTAGE5	647	S	Phosphorylation	2
ABLIM1	578	Y	Phosphorylation	1
FOXM1	658	T	Phosphorylation	9
TRA2B	95	S	Phosphorylation	7
RAD51AP1	326	K	Acetylation	7
ATRN	1185	T	Phosphorylation	4
EIF4B	462	S	Phosphorylation	9
LMO7	1182	S	Phosphorylation	8
LMO7	663	S	Phosphorylation	7
AKAP13	808	S	Phosphorylation	9
PAK2	20	S	Phosphorylation	9
PAK2	2	S	Phosphorylation	5
TACC2	151	T	Phosphorylation	7
HSD17B7	321	K	Acetylation	3
TJP2	1012	S	Phosphorylation	4
TJP2	1010	Y	Phosphorylation	3
GNL3	478	S	Phosphorylation	3
NCAPD2	1325	Y	Phosphorylation	1
PHLDB2	252	Y	Phosphorylation	4
CTR9	943	S	Phosphorylation	4
ZNF267	424	T	Phosphorylation	5
ZWILCH	85	T	Phosphorylation	2
L1TD1	548	T	Phosphorylation	4
PHACTR4	131	S	Phosphorylation	1
TTC19	6	T	Phosphorylation	6
ORAI1	223	N	Glycosylation	1
CC2D1A	92	T	Phosphorylation	3
ZNF768	83	S	Phosphorylation	6
MKI67	2720	T	Phosphorylation	4

SRGAP1	1001	T	Phosphorylation	2
PLEKHA5	382	S	Glycosylation	6
VWF	2754	C	Disulfide	9
SERPING1	205	C	Disulfide	9
IFNGR1	85	C	Disulfide	8
VWF	1130	C	Disulfide	9
VWF	1458	C	Disulfide	9
VWF	2804	C	Disulfide	9
VWF	804	C	Disulfide	9
HBB	9	K	Glycation	7
HBB	131	Y	Phosphorylation	1
HBB	131	Y	Phosphorylation	1
HBB	145	K	Acetylation	3
HBB	67	K	Glycation	8
AHSG	340	R	Proteolytic	2
ACE	716	T	Phosphorylation	1
CAPN3	417	T	Phosphorylation	8
CAPN3	345	S	Phosphorylation	1
CSF1	461	S	Phosphorylation	4
EPB41	410	S	Phosphorylation	9