Pacific Symposium on Biocomputing 15:337-347(2010)

LOSS OF POST-TRANSLATIONAL MODIFICATION SITES IN DISEASE

SHUYAN LI
School of Informatics and Computing, Indiana University
Bloomington, IN 47408, U.S.A.
College of Chemistry and Chemical Engineering, Lanzhou University
Lanzhou, Gansu 730000, China

LILIA M. IAKOUCHEVA
Laboratory of Statistical Genetics, The Rockefeller University
New York, NY 10065, U.S.A.

SEAN D. MOONEY
Buck Institute for Age Research
Novato, CA 94945, U.S.A.

PREDRAG RADIVOJAC
School of Informatics and Computing, Indiana University
Bloomington, IN 47408, U.S.A.

Understanding and predicting molecular cause of disease is one of the major challenges for biology and medicine. One particular area of interest continues to be computational analyses of disease-associated amino acid substitutions. To this end, various studies have been performed to identify molecular functions disrupted by disease-causing mutations. Here, we investigate the influence of disease-associated mutations on post-translational modifications. In particular, we study the loss of modification target sites as a consequence of disease mutation. We find that about 5% of disease-associated mutations may affect known modification sites, either partially (4%) of fully (1%), compared to about 2% of putatively neutral polymorphisms. Most of the fifteen post-translational modification types analyzed were found to be disrupted at levels higher than expected by chance. Molecular functions and physiochemical properties at sites of disease mutation were also compared to those of neutral polymorphisms involved in the process of post-translational modification site disruption. Disease-associated mutations in the neighborhood of post-translationally modified sites were found to be enriched in mutations that change polarity, charge, and hydrophobicity of the wild-type amino acids. Overall, these results further suggest that disruption of modification sites is an important but not the major cause of human genetic disease.

1. Introduction

1.1. *Function of post-translational modifications*

Protein post-translational modifications are reversible or irreversible chemical alterations of a protein after its translation. They include covalent additions of particular chemical groups (e.g. phosphoryl), lipids (e.g. palmitic acid), carbohydrates (e.g. glucose) or even entire proteins (e.g. ubiquitin) to amino acid side chains, as well as the enzymatic cleavage of peptide bonds [1]. With some exceptions (e.g. hydroxylation), protein post-translational modifications occur at side chains that can act as either strong (C, M, S, T, Y, K, H, R, D, E) or weak (N, Q) nucleophiles, while the remaining residues (P, G, L, I, V, A, W, F) are rarely involved in covalent modifications of their side chains. Post-translational modifications frequently affect protein function via changes in the protein structure and dynamics. Alternatively, modified residue may be a part of a binding region directly recognized by a partner. For example, phosphotyrosines are known to be directly targeted by the SH2 domains [2] and acetyllysines are targeted by bromodomains [3]. Similarly, proteolitic cleavage is typically a part of degradation pathways. Biologically, post-translational modifications are involved in a number of activities such as regulation of gene expression, activation/deactivation of enzymatic activity, protein stability or destruction, mediation of protein-protein interactions etc. [1]. Whatever the molecular context, the major role of post-translational modifications is to enable signaling and regulatory mechanisms that modulate protein's cellular function.

Correspondence: Predrag Radivojac, 901 E. 10th St., Bloomington IN, 47408, U.S.A.; Email: predrag@indiana.edu. Availability: http://mutdb.org

There are more than 200 documented types of post-translational modifications, many of which were discovered only recently [4]. More interestingly, a large fraction of them are catalyzed by modifying enzymes. It is estimated that about 5% of the genes in *Homo sapiens* are modifying enzymes [1]. There are 518 kinases in the human genome and more than 150 phosphatases [5]. Similarly, the human genome also codes for around 600 E3 ubiquitinating ligases and 80 deubiquitinases [6]. These modifying enzymes are ubiquitous in all kingdoms of life, especially in eukayotes. For example, there are 1019 kinase- and 300 phosphatase-coding genes in *Arabidopsis thaliana* and even the yeast genome codes for 119 kinases [7]. However, despite the increasing recognition of their importance, the commonness and full functional repertoire of post-translational modifications are still unknown. The focus of this study is on the phenotypic effects of the disruption of post-translationally modified sites by the single amino acid substitution events.

1.2. *Mutation of post-translational modification sites may lead to disease*

There are a number of cases in which mutations of the post-translational target sites were found to be directly involved in disease. One example is a loss of N-linked glycosylation in the prion protein (PRNP), where amino acid substitution T183A was shown to be involved in autosomal dominant spongiform encephalopathy [8]. This particular variant causes numerous clinical symptoms such as early-onset dementia, cerebral atrophy, and hypometabolism. Interestingly, a wild-type form of PRNP was also found to be protease-resistant in the presence of the mutant. N-linked glycosylation occurs on asparagine residues in NX[ST] motifs, thus the loss of the threonine in the consensus sequence prevents the attachment of a carbohydrate. Modifications of the NX[ST] motif have previously been implicated in intracellular accumulation of PRNP *in vitro* [9]. Another example is a loss of acetylation sites in androgen receptor (AR). Loss of AR acetylation has been implicated in Kennedy's disease, an inherited neurodegenerative disorder. Here, amino acid substitution K630A or both K632A and K633A have been shown to cause a significant slowdown of ligand-dependent nuclear translocation [10]. Furthermore, the non-acetylated mutants misfold and form aggregates with several other proteins, including ubiquitin ligase E3, thus affecting proteosomal degradation. And yet another example involves serine phosphorylation in the period circadian protein homolog 2 protein (PER2). Mutation of S662 is associated with the familial advanced sleep phase syndrome, an autosomal dominant disorder with early sleep onset (around 7:30pm) and early awakening (around 4:30am), but normal sleep duration [11]. Biochemical studies have shown that phosphorylation of S662 affects phosphorylation (by casein kinase CKIε) of several other residues in PER2, resulting in an overall hypophosphorylation of PER2. Interestingly, creation of a negative charge by S662D or an excess of CKIε restores the phosphorylation patterns of PER2. The current working hypothesis regarding PER2 is that phosphorylation of S662 likely creates a recognition site for CKIε and triggers a cascade of downstream effects. However, functional roles of phosphorylated PER2 are still largely unknown [11].

In addition to the individual examples, systematic studies implicating post-translational modifications in disease are now facilitated by the rapid growth of databases containing disease-associated mutations, human polymorphisms, and also post-translational modifications. One of the first such studies was carried out by Wang and Moult who analyzed protein structures and concluded that a large majority of human inherited disease mutations affect protein stability [12]. Only a small percentage of amino acid substitutions were estimated to affect post-translational modifications and binding sites in general; however, only N-linked glycosylation was investigated. In addition, Wang and Moult studied only protein structures, whereas several types of post-translational modifications were shown to be preferentially occurring in the disordered protein regions [13-15]. 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 [16, 17]. Lee et al. [18] and Yang et al. [19] matched experimentally determined modification sites with amino acid substitutions from different databases and found 47 and 64 substitutions to affect post-translational modifications. In our previous work, we studied modification of confidently predicted phosphorylation sites affected by the somatic mutations and found that both gain and loss of phosphorylation target sites may be an active mechanism in cancer [20]. This study was

recently extended to include confident predictions of methylation, ubiquitination, and O-linked glycosylation, implicating all three modifications in disease [15, 21, 22].

1.3. *Outline of the study*

In this study, we adopt a simple strategy and analyze a larger number of post-translational modifications in the context of disease-associated and putatively neutral amino acid substitutions. Experimentally verified sites of post-translational modifications were searched against amino acid substitution databases with the goal of investigating whether (and in what ways) changes of post-translational modifications are affected by inherited and somatic disease mutations. We found that disease-associated mutations are enriched in the fraction of directly disrupted modification sites, but also those found in their close proximity. In contrast, the putatively neutral polymorphisms occur less frequently in the neighborhoods of the modification sites. Furthermore, we found that the sites of post-translational modifications were enriched in amino acid substitutions that change physicochemical properties of the wild-type amino acids.

2. Materials and Methods

2.1. *Data sets*

The data sets of post-translational modifications were collected from several public databases and the literature. We mined Swiss-Prot [23], Human Protein References Database (HPRD) [24], Phospho.ELM [25], Protein Data Bank (PDB) [26], O-GlycBase [27], PhosphoSite [28], and PhosphoPOINT [19]. Only modification types containing 50 or more instances were of interest, resulting in 15 different post-translational modifications from a number of different species. In total, these data sets contained 78,975 unique sites (Table 1).

Table 1. Summary of the data sets of post-translational modifications. All modifications were extracted from Swiss-Prot and HPRD. Glycosylation sites were also extracted from O-GlycBase and PDB. Phosphorylation sites were additionally extracted from PDB, Phospho.ELM, PhosphoSite, and PhosphoPOINT.

Post-translational modification	Total sites	Total proteins	Human sites	Human proteins
Phosphorylation	62,269	17,116	30,838	8,428
N-linked glycosylation	4,971	2,257	2,181	906
O-linked glycosylation	2,853	367	295	93
Acetylation	2,600	1,896	1,024	677
Amidation	2,163	1,339	44	30
Hydroxylation	1,301	251	211	29
Proteolytic cleavage	1,285	531	1,285	531
Methylation	911	407	430	143
Pyrrolidone carboxylic acid	728	590	78	74
Ubiquitination	516	353	266	196
Carboxylation	447	122	88	15
SUMOylation	381	201	319	160
Palmitoylation	328	200	163	88
Sulfation	229	145	80	38
Myristoylation	156	153	61	58

The data set of the inherited amino acid substitutions in humans (Disease-I) was assembled from the Human Gene Mutation Database (HGMD) [29] and Swiss-Prot. The data set of somatic mutations in cancer (Disease-S) was also collected from Swiss-Prot and several recent cancer gene resequencing projects reviewed by Lee et al. [30]. The sites already present in the Disease-I data set were removed from Disease-S. Finally, the putatively neutral polymorphisms (Neutral) were downloaded from the Swiss-Prot database. All polymorphisms found in the disease sets were removed. We assumed that only a small fraction of neutral polymorphisms may be involved in disease, that is, that the large majority of them are either neutral or have minor phenotypic effects. The data sets of amino acid substitutions are summarized in Table 2. In total, the set contained 73,463 amino acid substitutions from 12,987 proteins.

Table 2. Summary of the data sets of amino acid substitutions. The number of sites includes the set of unique positions of amino acid substitutions in a particular set.

Data set	Number of substitutions	Number of sites	Number of proteins	Source
Putatively neutral (Neutral)	29,190	28,864	10,416	Swiss-Prot
Disease, inherited (Disease-I)	40,512	33,416	2,605	HGMD, Swiss-Prot
Disease, somatic (Disease-S)	3,761	3,191	2,336	Swiss-Prot, Literature

2.2. *Matching post-translational modifications with amino acid substitutions*

In order to investigate the relationships between post-translational modifications and amino acid substitutions, different scenarios were considered. First, a set of human post-translational modifications was created by: (1) including only those sites that were experimentally identified in human proteins, and (2) mapping of 25-residue long fragments from any other species (modification site ±12 amino acids around it) to the human proteins such that all 25 residues were identical to the corresponding residues in the human protein. Clearly, in the latter case, the correctness of such modification sites is not guaranteed; however, an exact 25-residue fragment match is expected to be a strong indication of functional similarity. This is often true for the modifications where only local interaction exists with the modifying enzyme (e.g. kinases), however, in some other cases with long-range interactions (e.g. E3 ligase binding in ubiquitination) the assumption may be less likely to hold. The fragment length of 25 was chosen based on the phosphorylation data for which there is evidence of physical kinase-substrate binding within about 7-12 residues of the modification site [31]. We refer to the experimentally verified human modification sites as *true sites*, while the ones obtained by the exact fragment matches are referred to as the *homology sites*.

Two types of matching between amino acid substitutions and post-translational modifications were considered: (1) matches where the substitutions occurred at a modification site and (2) matches where the substitution site was in the neighborhood of the modification site (i.e. between residues −3 and +3). This matching was based on an assumption that a mutation can affect the post-translational modification if it is in the vicinity of the target residue. One such situation occurs with mutation R16C, which diminishes phosphorylation of S19, in human PTP synthase and causes hyperphenylalaninemia [32-34]. The situation where a substitution site and the modification site are at the same position is referred to as the *direct match*. A substitution site that is no more than 3 residues away from the modification site is referred to as the *neighborhood match*. An example of a neighborhood match to a homology site is shown in Figure 1.

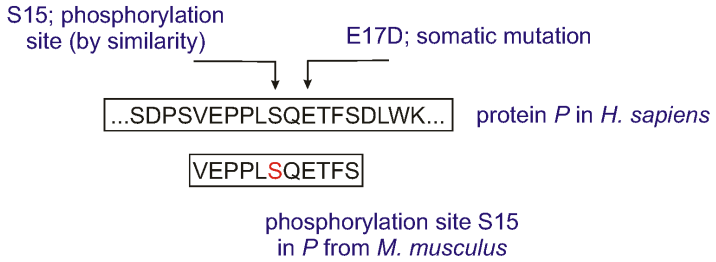

Figure 1. Stylized depiction of the matching between post-translational modification sites and amino acid substitutions. In this example, an 11-residue fragment corresponding to an experimentally verified phosphorylation site (serine, colored in red) from *M. musculus* is first matched to a homologous protein from *H. sapiens*. The corresponding residue S15 in the human protein is subsequently designated as phosphorylated (homology site). Finally, an observed amino acid substitution E17D, located within 3 residues from S15, is considered to impact the phosphorylation site (neighborhood match).

2.3. *Statistical analysis*

With all the matched sites of post-translational modifications and amino acid substitutions, two strategies were adopted to estimate statistical confidence of the observed trends. First, we used the t-test and the binomial test to estimate whether a certain group of amino acid substitutions (Neutral, Disease-I, Disease-S) is enriched or depleted in a particular modification. The hypergeometric test was used to estimate enrichment and depletion in functional terms for each of the categories. The amino acid property changes were studied for the mutations in the

neighborhood of modification sites (±3 residues). Three properties were investigated: side chain polarity (polar, non-polar), charge (positive, negative, neutral), and hydrophobicity (hydrophobic, hydrophilic), and the significance of those results was estimated using the t-test.

2.4. *Conservation index analysis*

Positional conservation was calculated using a commonly used conservation index AL2CO [35]. First, all 12,987 human proteins with mutations were searched against GenBank for the 500 best hits. These sequences were subsequently aligned using ClustalW program [36]. Then, the positional entropy by Henikoff and Henikoff [37] was calculated as the conservation index for all modification sites. The conservation index value was normalized to the 0-1 interval; the higher the value a position gets, the more conserved the position is.

3. Results

Using the two scenarios for obtaining post-translational modification sites (true and homologous sites) and using two strategies of matching them to the amino acid substitutions (direct and neighborhood matches), we analyzed amino acid substitutions in inherited disease, somatic disease, and neutral polymorphisms with respect to post-translational modifications.

3.1. *Disease-associated and neutral mutations affecting post-translational modifications*

The percentage of all amino acid substitutions that lie directly on or in the neighborhoods of modification sites in Disease-I, Disease-S, and Neutral data sets was investigated first. We found that direct and neighborhood mutations were in the vicinity of true and homology modification sites in 4.5% of cases in Disease-I, 3.1% of cases in Disease-S, and 2.1% of cases in Neutral data set (Figure 2). When only unique substitution sites were considered, these frequencies were 3.9%, 3.3%, and 2.1%, respectively (Figure 2). The most significant differences between the sets of inherited disease and neutral substitutions were detected in the cases of N-linked glycosylation (233 out of 306 in Disease-I; P = 1.6e–19), carboxylation (62/63; P = 2.4e–17), hydroxylation (68/72; P = 8.6e–16), acetylation (75/91; P = 4.9e–10), proteolytic cleavage (84/120; P = 1.9e–5), and O-linked glycosylation (32/41; P = 3.5e–4). Thus, the disease-associated mutations are more likely to affect post-translational modifications than the neutral substitutions (P = 5.6e–4).

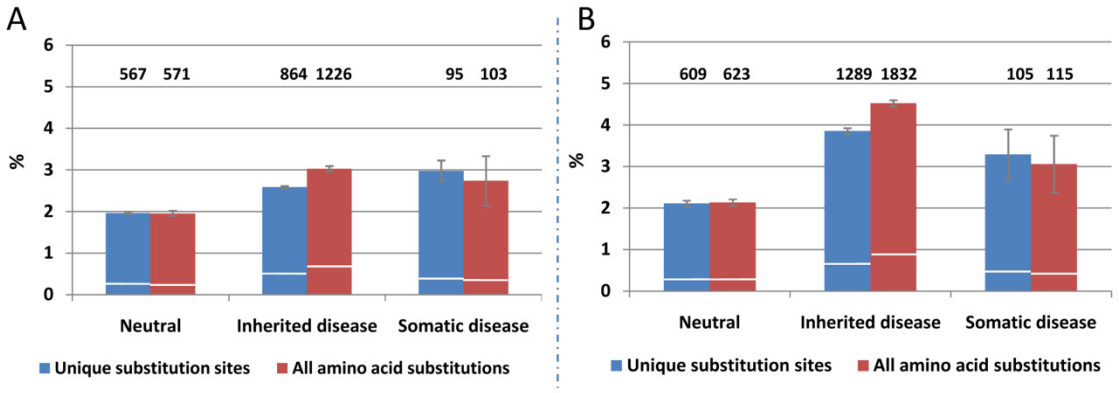

Figure 2. The fraction of disease-associated amino acid substitutions (inherited and somatic) that are assumed to affect post-translational modifications, compared to the fraction of neutral polymorphisms: (A) true modification sites and (B) homology modification sites. The white line in each bar separates direct matches from neighborhood matches. The error bars indicate 68% confidence intervals estimated by bootstrapping with 10,000 iterations.

Figure 3 shows the trends of enrichment and depletion of unique amino acid substitution sites that directly match experimentally verified (i.e. true) modification sites. The trend T was calculated as

$$T = \frac{f_{obs} - f_{exp}}{f_{obs} + f_{exp}}$$

where f_{obs} and $f_{exp} \neq 0$ are the observed and expected rates (relative frequencies) of substitutions that match modification sites, respectively. The trend is positive if $f_{obs} > f_{exp}$ and negative if $f_{exp} > f_{obs}$. $T = 1$ is the maximum value and involves a hypothetical situation with $f_{exp} = 0$ and $f_{obs} \neq 0$; $T = -1$ is the minimum value and indicates that $f_{obs} = 0$. Since inherited disease, somatic disease, and neutral polymorphisms contain 51%, 5%, and 44% of all substitution sites used in this study, f_{exp} was set to 0.51, 0.05, and 0.44 for the three data sets, respectively. The ratio of the three groups of amino acid substitutions also determines the null hypothesis that was used to calculate statistical significance of the observed enrichment or depletion. Trends similar to those observed in Figure 3 were present when the matching process was extended to homologous modification sites and neighborhood matches (Figure 4).

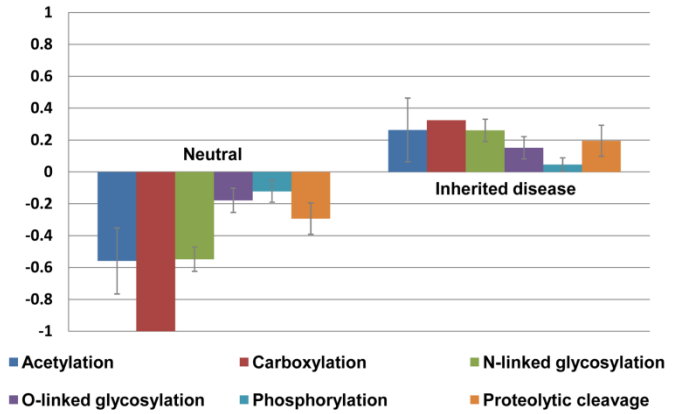

Figure 3. Trends that indicate whether a data set is enriched ($T > 0$) or depleted ($T < 0$) in direct matches between amino acid substitutions and true post-translational modifications. Only modifications with 5 matches or more are shown. The error bars indicate 68% confidence intervals estimated by bootstrapping with 10,000 iterations.

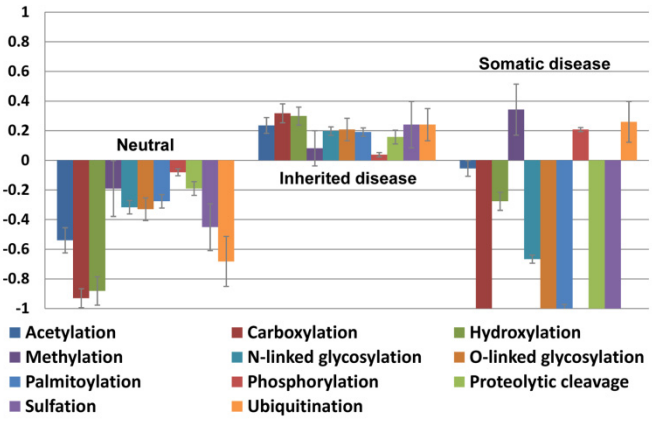

Figure 4. Trends that indicate whether a data set is enriched ($T > 0$) or depleted ($T < 0$) in direct and neighborhood matches between amino acid substitutions and true and homology post-translational modifications. Only modifications with 10 matches or more are shown. The error bars indicate 68% confidence intervals estimated by bootstrapping with 10,000 iterations.

Table 3 shows the observed numbers of amino acid substitutions matching post-translationally modified sites. The P-value for the positive trends T was calculated as

$$P = \sum_{i=k}^{K} \binom{K}{i} f_{exp}^{i} (1 - f_{exp})^{K-i}$$

where K is the total number of matches to any of the three substitution sets and k is the observed number of matches in a particular data set. The P-value for the negative trends was calculated by replacing k by 0 and K by k in the limits of the summation operator.

Table 3. The observed rates of matches between amino acid substitutions and true and homology post-translational modification sites. Example: for N-linked glycosylation, out of 31 mutations that directly matched N-linked glycosylation sites, 4 were observed in the neutral set (Neutral) and 27 in the inherited disease set (Disease-I). The expected number of matches were 13.6 and 15.8, respectively. Blue color indicates lower-than-expected observed frequencies; red indicates higher-than-expected observed frequencies. P-values were calculated using the binomial distribution. The Bonferroni-corrected value of 5.6e–4 (n = 90) was used to assign statistical significance for each group.

Post-translational modification	Direct match; true sites						Direct and neighborhood match; homology sites					
	Neutral		Disease-I		Disease-S		Neutral		Disease-I		Disease-S	
	f_{obs}	P	f_{obs}	P	f_{obs}	P	f_{obs}	P	f_{obs}	P	f_{obs}	P
Phosphorylation	47/136		76/136		13/136		457/1218	2.0e–6	670/1218		91/1218	5.3e–5
N-linked glycosylation	3/28	2.2e–4	27/31	2.7e–5	0/31		70/306	8.3e–15	233/306	1.6e–19	3/306	1.8e–4
O-linked glycosylation	4/13		9/13		0/13		10/45		35/45	2.1e–4	0/45	
Acetylation	1/8		7/8		0/8		12/91	2.4e–10	75/91	4.5e–10	4/91	
Amidation	1/1		0/1		0/1		2/4		2/4		0/4	
Hydroxylation	0/2		1/2		1/2		2/72	1.1e–15	68/72	6.6e–16	2/72	
Proteolytic cleavage	7/29		22/29		0/29		36/120		84/120	1.9e–5	0/120	
Methylation	0/4		4/4		0/4		6/20		12/20		2/20	
Pyrrolidone carboxylic acid	1/1		0/1		0/1		2/2		0/2		0/2	
Ubiquitination	0/0		0/0		0/0		2/24	1.7e–4	20/24		2/24	
Carboxylation	0/29	4.8e–8	29/29	3.4e–9	0/29		1/63	6.3e–15	62/63	2.4e–17	0/63	
SUMOylation	0/0		0/0		0/0		2/7		4/7		1/7	
Palmitoylation	0/0		0/0		0/0		4/16		12/16		0/16	
Sulfation	0/3		3/3		0/3		2/12		10/12		0/12	
Myristoylation	0/1		1/1		0/1		1/3		2/3		0/3	

Figure 3, Figure 4, and Table 3 indicate that the substitutions associated with inherited disease affect the sites of post-translational modifications with frequencies higher than expected by chance. In contrast, putatively neutral polymorphisms affect modification sites with lower-than-expected frequencies.

3.2. *Conservation index analysis*

It has been widely studied that disease-associated mutation sites are more evolutionarily conserved than human polymorphic sites [38]. Here, we analyzed the conservation of the post-translationally modified sites for which there are known amino acid substitutions either at the modification site itself or in its neighborhood. Not so surprisingly, we find that the modification sites directly matching disease-associated substitution sites are more conserved than those matching neutral polymorphic sites (Figure 5A). However, post-translational modifications lying in the neighborhood (±3 residues) of inherited disease mutations are also more conserved than the modification sites corresponding to the neutral polymorphisms, with a similar margin. On the contrary, the conservation of somatic mutations is significantly lower when the neighborhoods of modification sites were considered.

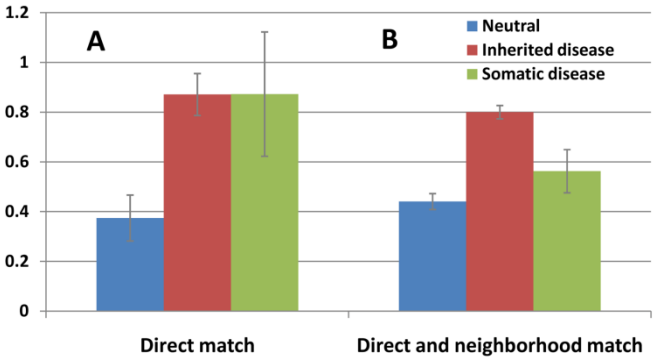

Figure 5. Sequence conservation index for post-translational modification sites matching disease-associated and neutral substitution sites: (A) direct match and (B) direct and neighborhood match. The conservation index was calculated as described in Section 2.4, with higher numbers indicating higher conservation. The error bars indicate 68% confidence intervals estimated by bootstrapping with 10,000 iterations.

3.3. *Gene function analysis*

To further study the impact of amino acid substitutions that occur in the vicinity of the modification sites, we analyzed gene functions of all types of post-translational modifications. The genes were first separated into the gene set containing disease mutations (inherited and somatic) and the gene set containing neutral polymorphisms (the two sets of genes were overlapping). Each of the sets was further split into the set of genes where amino acid substitutions impact modifications sites and the remaining genes. Then, the gene enrichment analysis of the two pairs of data sets was performed using the GOstat software [39]. It is important to understand that the set of disease-associated mutations impacting modification sites was compared to the remaining set of genes containing disease mutations in order to avoid biases that correspond to the disease genes [40]. In this way, we assume that it may be possible to identify molecular and cellular functions that were disrupted by the disease mutations or those that are regulated by post-translational modifications and related to minor phenotypic variations. The results of this analysis, using the Gene Ontology [41] category molecular function, are shown for the phosphorylation data set (Figure 6).

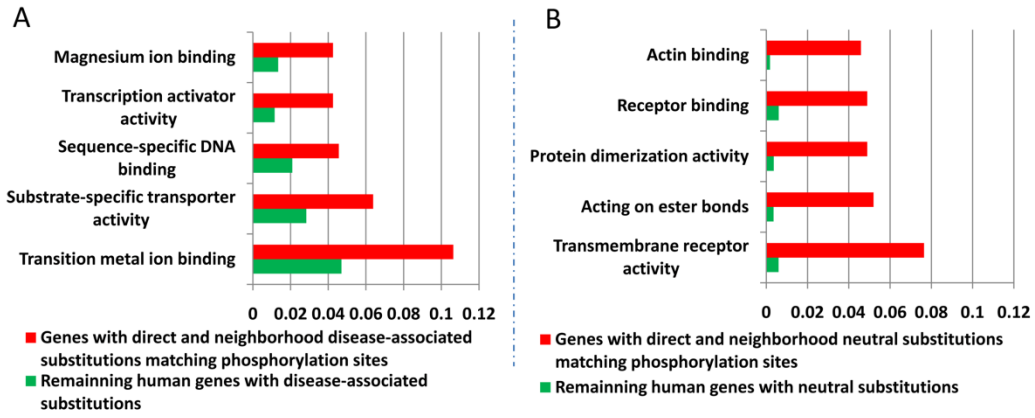

Figure 6. GO enrichment analysis between genes containing mutations in the vicinity of the predicted phosphorylation sites vs. the remaining genes from a selected data set. The analysis was done separately for genes containing disease-associated mutations and genes containing neutral polymorphisms. Top 5 molecular functions are shown, where the upper (red) bars indicate functions enriched in the set with phosphorylation site disruption, while the lower (green) bars indicate functions for the remaining set of genes from the group.

Interestingly, both sets of genes (with disease related mutations and with neutral polymorphisms) in the vicinity of phosphorylation sites have significant enrichment in different molecular functions. For example, kinase, transferase, and signal transduction activities are significantly enriched in the disease-associated set, whereas RNA binding, transcription factor and receptor activities are significantly enriched in the neutral substitutions set. However, both

sets are enriched in important molecular functions, thereby suggesting that both disease and neutral substitutions in the vicinity of phosphorylation sites, may have an impact on protein function.

3.4. *Amino acid properties of substitutions in the vicinity of modification sites*

Next, we analyzed physicochemical properties of the amino acid substitutions affecting post-translational modifications. Polarity, charge, and hydrophobicity were chosen for this analysis. These properties were studied for the amino acid substitutions occurring directly and in the neighborhood (±3 residues) of all types of modified sites. Figure 6 shows the enrichment and depletion of the observed changes in all three data sets. We observed that the inherited disease mutations are enriched in the change of all three properties for several modification types. On the other hand, neutral mutations are depleted in such changes. Somatic mutations do not show significant signals potentially due to the small size of the data set.

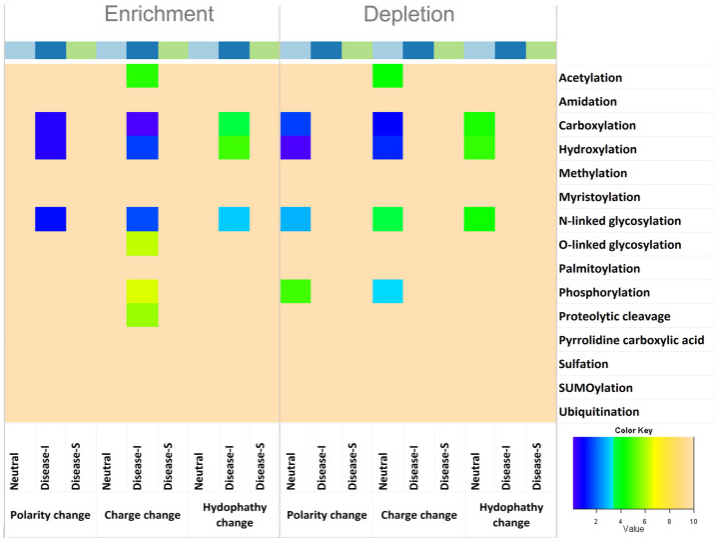

Figure 7. Enrichment and depletion of types of amino acid substitutions observed in the sets of disease-associated substitutions and putatively neutral polymorphisms. P-values were calculated using the binomial distribution. The plotted value is $10 + \log_{10}(P + 1 \cdot 10^{-10})$, where P is the P-value. The darker colors indicate values with lower P-values. All values below the Bonferroni-corrected value 3.7e–4 were set to 1.

4. Discussion

When personalized medicine is the next frontier for scientists, industry, and the general population, it is important to develop computational approaches that can lead to a better understanding of the etiology of disease. Integration of genetic and molecular information is a sensible step in this direction because it provides a structural and functional perspective to the human variation data.

In this study, we analyzed disease-associated and putatively neutral amino acid substitution data and found that about 4.5% of amino acid substitutions (3.9% of unique sites) may affect protein function through disruption of post-translational modifications. On the other hand, about 2.1% of neutral polymorphisms may be affecting post-translational modifications. These numbers further indicate that post-translational modifications are not the majority cause of human genetic disease. However, we have still found 238 post-translationally modified sites in human proteins whose mutation was causative of disease. In total, 1,289 modification sites were found to be in the close proximity to the inherited disease mutations and represent candidates for further experimental verification.

Given our data, there are several problems that could have lead to the ascertainment bias. For example, our data set of post-translational modifications was heavily skewed towards phosphorylation (79%), where mass spectrometry techniques have lead to a recent explosion in the number of identified sites. On the other hand, it may be argued that the modifications not identified using high-throughput methods may be more likely to be disease-

relevant. It is also unclear whether the sets of inherited disease data are representative since it may be expected that genetic-association studies are more successful in identifying markers of monogenic diseases or familiar forms of complex diseases. Finally, the set of neutral polymorphisms is probably contaminated with yet undiscovered disease mutations and has not been controlled for population biases.

We also analyzed the enrichment and depletion of amino acid substitutions for each post-translational modification and found that most follow similar trends when inherited disease is compared to the neutral polymorphisms. These trends held for both experimentally verified modification sites and those transferred by homology. In the case of somatic mutations, we observed some interesting cases as well. For most examples, we have not found matches between post-translational modifications and observed somatic mutations. However, in the cases of methylation, phosphorylation and ubiquitination, there was an increased trend of disruption of post-translational modifications. Previous work has already addressed disruption of confidently predicted phosphorylation sites in cancer [20]. Thus, the correspondence between actual sites and somatic mutations found in this study further supports this hypothesis.

While direct disruption of post-translational modifications is likely to have functional implications, the partial disruption of modified sites has a potential to lead to subtle phenotypic effects that may be more dependent on the variation present in other genes before causing organism-wide dysregulation. We believe that such changes are particularly fitting to the framework of complex disease and interaction between genetic and environmental factors.

Acknowledgments

This work was supported by the NIH grant R01LM009722-01 to SDM, NIH grant R21CA113711 to LMI, and NSF grant DBI-0644017 to PR.

References

1. Walsh, C.T., Posttranslational modification of proteins: expanding nature's inventory. 2006, Englewood, CO: Roberts and Company Publishers.
2. Felder, S., et al., SH2 domains exhibit high-affinity binding to tyrosine-phosphorylated peptides yet also exhibit rapid dissociation and exchange. *Mol Cell Biol*, 1993. **13**(3): p. 1449-55.
3. Yang, X.J., Lysine acetylation and the bromodomain: a new partnership for signaling. *Bioessays*, 2004. **26**(10): p. 1076-87.
4. Mann, M., O.N. Jensen, Proteomic analysis of post-translational modifications. *Nat Biotechnol*, 2003. **21**(3): p. 255-61.
5. Manning, G., et al., The protein kinase complement of the human genome. *Science*, 2002. **298**(5600): p. 1912-34.
6. Komander, D., et al., Breaking the chains: structure and function of the deubiquitinases. *Nat Rev Mol Cell Biol*, 2009. **10**(8): p. 550-63.
7. Wang, D., et al., Systematic trans-genomic comparison of protein kinases between Arabidopsis and Saccharomyces cerevisiae. *Plant Physiol*, 2003. **132**(4): p. 2152-65.
8. Grasbon-Frodl, E., et al., Loss of glycosylation associated with the T183A mutation in human prion disease. *Acta Neuropathol*, 2004. **108**(6): p. 476-84.
9. Rogers, M., et al., Intracellular accumulation of the cellular prion protein after mutagenesis of its Asn-linked glycosylation sites. *Glycobiology*, 1990. **1**(1): p. 101-9.
10. Thomas, M., et al., Androgen receptor acetylation site mutations cause trafficking defects, misfolding, and aggregation similar to expanded glutamine tracts. *J Biol Chem*, 2004. **279**(9): p. 8389-95.
11. Toh, K.L., et al., An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science*, 2001. **291**(5506): p. 1040-3.
12. Wang, Z., J. Moult, SNPs, protein structure, and disease. *Hum Mutat*, 2001. **17**(4): p. 263-70.
13. Iakoucheva, L.M., et al., The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 2004. **32**(3): p. 1037-1049.

14. Daily, K.M., et al.. Intrinsic disorder and protein modifications: building an SVM predictor for methylation. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), p. 475-481, 2005. San Diego, California, U.S.A.
15. Radivojac, P., et al., Identification, analysis and prediction of protein ubiquitination sites. *Proteins*, 2009.
16. Vogt, G., et al., Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. *Nat Genet*, 2005. **37**(7): p. 692-700.
17. Vogt, G., et al., Gain-of-glycosylation mutations. *Curr Opin Genet Dev*, 2007. **17**(3): p. 245-51.
18. Lee, T.Y., et al., dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*, 2006. **34**(Database issue): p. D622-7.
19. Yang, C.Y., et al., PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, 2008. **24**(16): p. i14-20.
20. Radivojac, P., et al., Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, 2008. **24**(16): p. i241-7.
21. Mort, M.E., et al., In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. In review.
22. Li, B., et al., Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 2009.
23. Bairoch, A., et al., The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 2005. **33 Database Issue**: p. D154-9.
24. Keshava Prasad, T.S., et al., Human Protein Reference Database--2009 update. *Nucleic Acids Res*, 2009. **37**(Database issue): p. D767-72.
25. Diella, F., et al., Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 2004. **5**(1): p. 79.
26. Berman, H.M., et al., The protein data bank. *Nucleic Acids Res*, 2000. **28**(1): p. 235-242.
27. Gupta, R., et al., O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res*, 1999. **27**(1): p. 370-2.
28. Biondi, R.M., Phosphoinositide-dependent protein kinase 1, a sensor of protein conformation. *Trends Biochem Sci*, 2004. **29**(3): p. 136-42.
29. Stenson, P.D., et al., The Human Gene Mutation Database: 2008 update. *Genome Med*, 2009. **1**(1): p. 13.
30. Lee, W., et al., Analytical methods for inferring functional effects of single base pair substitutions in human cancers. *Hum Genet*, 2009.
31. Songyang, Z., et al., Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol*, 1994. **4**(11): p. 973-982.
32. Oppliger, T., et al., Structural and functional consequences of mutations in 6-pyruvoyltetrahydropterin synthase causing hyperphenylalaninemia in humans. Phosphorylation is a requirement for in vivo activity. *J Biol Chem*, 1995. **270**(49): p. 29498-506.
33. Scherer-Oppliger, T., et al., Serine 19 of human 6-pyruvoyltetrahydropterin synthase is phosphorylated by cGMP protein kinase II. *J Biol Chem*, 1999. **274**(44): p. 31341-8.
34. Thony, B., et al., Hyperphenylalaninemia due to defects in tetrahydrobiopterin metabolism: molecular characterization of mutations in 6-pyruvoyl-tetrahydropterin synthase. *Am J Hum Genet*, 1994. **54**(5): p. 782-92.
35. Pei, J. and N.V. Grishin, AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 2001. **17**(8): p. 700-12.
36. Thompson, J.D., et al., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 1994. **22**(22): p. 4673-4680.
37. Henikoff, S., J.G. Henikoff, Position-based sequence weights. *J Mol Biol*, 1994. **243**(4): p. 574-8.
38. Ng, P.C., S. Henikoff, Predicting deleterious amino acid substitutions. *Genome Res*, 2001. **11**(5): p. 863-74.
39. Beissbarth, T. and T.P. Speed, GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 2004. **20**(9): p. 1464-5.
40. Dalkilic, M.M., et al., From protein-disease associations to disease informatics. *Front Biosci*, 2008. **13**: p. 3391-407.
41. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000. **25**(1): p. 25-29.