Human Proteins with Target Sites of Multiple Post-Translational Modification Types Are More Prone to Be Involved in Disease

Qianli Huang,†‡ Jinhui Chang,‡ Man Kit Cheung,† Wenyan Nong,‡ Lei Li,† Ming-tsung Lee,† and Hoi Shan Kwan†§‡∗

†School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR 852000, China
‡Shenzhen Research Institute, The Chinese University of Hong Kong, 2nd Yuexing Road, Nanshan District, Shenzhen, China

Supporting Information

ABSTRACT: Many proteins can be modified by multiple types of post-translational modifications (Mtp-proteins). Although some post-translational modifications (PTMs) have recently been found to be associated with life-threatening diseases like cancers and neurodegenerative disorders, the underlying mechanisms remain enigmatic to date. In this study, we examined the relationship of human Mtp-proteins and disease and systematically characterized features of these proteins. Our results indicated that Mtp-proteins are significantly more inclined to participate in disease than proteins carrying no known PTM sites. Mtp-proteins were found significantly enriched in protein complexes, having more protein partners and preferred to act as hubs/superhubs in protein–protein interaction (PPI) networks. They possess a distinct functional focus, such as chromatin assembly or disassembly, and reside in biased, multiple subcellular localizations. Moreover, most Mtp-proteins harbor more intrinsically disordered regions than the others. Mtp-proteins carrying PTM types biased toward locating in the ordered regions were mainly related to protein–DNA complex assembly. Examination of the energetic effects of PTMs on the stability of PPI revealed that only a small fraction of single PTM events influence the binding energy of >2 kcal/mol, whereas the binding energy can change dramatically by combinations of multiple PTM types. Our work not only expands the understanding of Mtp-proteins but also discloses the potential ability of Mtp-proteins to act as key elements in disease development.

KEYWORDS: post-translational modification, protein–protein interaction, disease development, intrinsically disordered region, topological coefficient

INTRODUCTION

Proteins are usually chemically modified after translation to extend the scope of functions. Resulting from the improvement of detection technologies, the number of post-translational modification (PTM) types reported has risen to over 200, and more than tens of thousands of modified sites have been detected.1−3 With the sharp increase in identified PTM types and modified sites, the interplay among multiple types of PTMs on individual proteins has attracted increasing interest.4 A complex interplay known as the “histone code” describes that cross-regulations of PTMs in histone can form a code to be read by different effectors and control gene expressions through regulating the chromatin states.5−9 Recently, a similar code for transcription factors (TFs) has been proposed, describing specific combinations of PTMs with specific outcomes that can be induced via TFs, such as FOXO factors, C/EBPβ, and NFκB.10−12 For example, while growth-factor signaling-induced phosphorylation of FOXO factors stimulated cell growth by transcriptional inactivation, cellular stress induced phosphorylation and monoubiquitylation arrested cell cycle by increasing transcriptional activation.13,14 Besides, interactions between PTM types have also been observed in many other proteins. For instance, adjacent phosphorylation and methylation sites demonstrated cross-inhibition in DNMT1;15 ubiquitylation and acetylation were interconnected in the p53 regulation;16 and positive and negative roles of lysine acetylation on phosphorylation were reported in Cdc6 and RIP140, respectively.17,18 These suggested that combinatorial interactions between multiple PTM types are a universal mode of protein regulation. Combination multiple types of PTMs can influence their target proteins from several different aspects. Santonico et al.19 reported that multiple modifications of RNF11 drove its correct localization to the endosomal compartment. Protein–protein interactions (PPIs) are also regulated by combinational multiple types of PTMs. The tumor suppressor protein p53, which contains multiple overlapping PTM motifs, was found to facilitate binding to different interaction partners in a mutually exclusive manner.20,21 These studies showed that many key developmental regulators such as histone, p53, and Cdc6 are
regulated by combinatorial interactions of multiple PTMs, and this combinatorial mode confers many special and crucial roles in protein regulation.

Apart from normal cellular dynamics, various PTMs are also found to impact disease pathology. Ser-129 phosphorylation of α-synuclein, for instance, is responsible for the formation of Lewy bodies in Parkinson’s disease and dementia. Threonine phosphorylation (T163) on protein α-tyrosine of cognitive protein can also cause many impacts on human diseases ranging from cancers, chronic lung disease, to cardiovascular disease. Although more and more efforts have been attracted to understand the relevance of PTMs in the cellular context, revealing biological significance of all the PTM sites on normal development or disease is still a huge challenge owing to the constant increase in the number of new PTM sites and types. Here, to further disclose the correlation of PTM in disease, we examined the relationship of human Mtp-proteins and disease and systematically characterized features of these proteins.

## MATERIALS AND METHODS

### Data Sets of Experimentally Determined PTM Events and Disease-Related Genes in Human

PTM events in the proteomes of HUMAN (Homo sapiens) were retrieved from the UniProt database (http://www.uniprot.org/). Annotations carrying nonexperimental qualifiers (“By similarity”, “Probable”, and “Possible”) were excluded to ensure a high accuracy of the annotations, using the workflow, as described by Khoury et al. Thus, 39,230 experimentally determined PTM events in human were deposited in data set one (Supplementary File 1a in the Supporting Information).

To examine whether Mtp proteins are more inclined to participate in disease, we then compiled a disease-related gene data set according to the OMIM database (Online Mendelian Inheritance in Man: http://www.omim.org/) and HGMD (Human Gene Mutation Database). A data set composed of 9029 human disease-related genes was constructed (Supplementary File 1b in the Supporting Information).

### Classification of Human Proteins

To examine potential features and functions of Mtp-proteins, defined here as having target residues of at least two PTM types, proteins of the human proteome were classified into different groups based on the number of PTM types present: group “0” represents that proteins in this group do not carry target residue of any known PTM type, group “1” denotes that the proteins in the group only contain target residues of one PTM type, and so on. The classification information on protein groups was deposited in Supplementary File 2 in the Supporting Information.

### Construction of Human PPI Data Set and PPI Networks

A data set of human PPIs was constructed by combining APID, HIPPIE, and HitPredict to reduce possible bias of using a single database. To avoid redundant records, repeated records among the databases were removed, and only the unique records were kept. PPI information was then assigned to different protein groups. The plug-in BisoGenet in platform Cytoscape, which is based on the databases BIOGRID, INTACT, MINT, DIP, BIND, and HPRD, was used to construct the PPI networks for the whole human proteome. After isolated nodes, small network components, and self-loops were removed, the NetworkAnalyzer (a versatile and highly customizable Cytoscape plugin that can compute and display a comprehensive set of topological parameters and centrality measures for undirected and directed networks) was used to calculate the degree of each node (a protein in the network).

Here, degree represents the number of partner proteins to which a given protein can bind. According to the degree number of each node, all nodes were classified into four categories (Peripheral-A: degree = 1; Peripheral-B: 20 > degree > 1; Hub: 100 > degree ≥ 20; and Superhub: degree ≥ 100) following the scheme by Lu et al. The proportions of each node category among different protein groups were then compared. Corresponding subnetworks constituted by different protein groups were also constructed using the same method. Topological coefficient (Tc) values were calculated using NetworkAnalyzer. Tc is a measure of the extent to which a protein in a PPI network shares the same neighbors with other proteins. A large Tc value means a higher tendency of the protein to have shared partners.

### Functional Characterization and Subcellular Localization Analysis

To detect potential functional bias of Mtp-proteins, functional annotations of different protein groups were first assigned using DAVID version 6.7 (Database for Annotation, Visualization, and Integrated Discovery, which provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes). Enrichment of GO terms (Gene Ontology terms) among the groups was then calculated to detect over-represented functional annotations in each group using a value of 1.0 × 10^-5 as the cutoff for functional clustering. On the basis of the clustering, similar terms were replaced by representative substitutes. Yloc (an interpretable web server for predicting subcellular localization) was used to predict subcellular localizations for the human proteome using the “HighRes” prediction model. Annotations of subcellular localizations from LocDB, an expert-curated database that collects experimental annotations for the subcellular localization of proteins, were also retrieved for analysis to improve precision. Distributions of subcellular localizations among different protein groups were then compared.

### Annotation of IDRs

SPINE-D was used to detect IDRs of the human proteome. The program can accurately predict not only long but also short disordered regions using the single neural-network-based method. Before running SPINE-D, PSIBLAST was first performed against the NR database for each protein. SPINE-X (http://sparks.informatics.iupui.edu/SPINE-X/) and IUpred (http://iupred.enzim.hu/) were then used to generate input files, such as dphi and dpsi, for SPINE-D. Before comparison of the IDR content among different protein groups, IDRs detected were divided into long IDRs and short IDRs: an IDR was regarded as “long” if it was longer than two times the average length of all IDRs to avoid potential negative effects caused by comparatively lower prediction accuracies for short IDRs.

### Binding Energy Analysis

Annotations of dimers were downloaded from PDBePISA (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html). On the
basis of the PDB IDs of the dimers, corresponding PDB structures were retrieved from the RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). The target sites of different PTM types were then mapped onto the PDB structures. The energetic contribution of each site was assessed by alanine scanning in FoldX (an empirical force field that was developed for the rapid evaluation of the effect of mutations on the stability, folding, and dynamics of proteins and nucleic acids). All residues were divided into three groups: residues on the target sites were presented as group "R1"; residues in "R2" are the ones identical to the target residues but not located on the target sites; and residues in complexes were considered as group "R3", no matter what the residue is and which site it is on. Energetic effects among different PTM types and different residue groups were then compared. Besides, on the basis of the data set of experimentally determined PTM events, possible combinations of PTM types were listed, and we calculated the number of proteins belonging to each combinatorial pattern. Among all combinatorial patterns composed of two PTM types, the one composed of phosphorylation and acetylation was the most common, and this pattern (pattern PA) was chosen to test whether combinatorial multiple types of PTMs have more serious influence on the stability of PPI. Thus, to evaluate the energetic contribution of potential combinatorial interplay between multiple PTMs, target sites of phosphorylation and acetylation were both replaced with alanine, and the energy differences were then identified with the “multiple mutations using mutant file” method in FoldX.

Statistical Analyses

Features of different groups were analyzed using SAS (Statistical Analysis System). Values in different groups were compared using the t test and ANOVA with the F test and ranked by the Student–Newman–Keuls test. Comparisons of proportions between two sample sets were performed with the chi-squared test, and corresponding trends among groups were examined with the Cochran–Armitage trend test.

RESULTS AND DISCUSSION

Mtp-Proteins Are More Inclined to Be Associated with Human Diseases

To examine whether Mtp proteins are more inclined to participate in disease, we first compiled a disease-related protein data set according to the OMIM database (Online Mendelian Inheritance in Man: http://www.omim.org/) and HGMD (Human Gene Mutation Database). All human proteins were divided into groups based on the number of PTM types in individual proteins. The proportion of disease-related proteins in the different protein groups was then calculated (See Materials and Methods, Supplementary Files 1 and 2 in the Supporting Information). Our results revealed that the proportion of disease-related proteins became higher with an increasing number of PTM types harbored in the proteins ($R^2 > 0.99$) (Figure 1). Specifically, for protein “group 0”, in which all proteins carry no known PTM event, the percentage of disease-related proteins involved was only 35%. The value increased to 53% in “group 1”, 68% in “group 2”, and up to 88% in “group 5”. Significant differences were detected not only between “group 0” and “group 1” ($X^2 = 530.32$, $p < 0.01$) but also between “group 1” and “group >1” ($X^2 = 963.72$, $p < 0.01$). Also, we tested the trend using the Cochran–Armitage trend test, and results indicated that this increasing trend was significant ($X^2 = 417.55$, $p < 2.2 \times 10^{-16}$). These suggest that compared with the proteins carrying no known PTM event, a significantly higher proportion of Mtp-proteins is disease-associated, implying a preference of these proteins to be involved in disease. However, why are Mtp proteins more prone to be associated with human diseases? Are there any special features and functions of these proteins that make them the key elements in disease development?

Exploratory Analysis on Mtp-Proteins: Mtp-Proteins Are Enriched in Protein Complexes

Recently, it has been shown that the “NLRP3 inflammasome” complex plays a critical role in the development of Alzheimer’s disease; the mutant forms of either PINK1 or Parkin in the Parkin/PINK1/DJ-1 complex are associated with Parkinson disease; and there are many potential cancer-associated protein complexes. To investigate whether Mtp-proteins prefer to form complexes or whether complex-forming proteins prefer to be post-translationally modified, we analyzed 1845 human complexes, comprising 8842 components, retrieved from the CORUM database (Comprehensive Resource of Mammalian Protein Complexes). Among all components, 7173 (92%) of them were proteins with PTM sites. This is 2.3 times higher than the frequency of proteins with PTM sites in the human proteome. Similarly, for the nonredundant list of 2556 complex-forming proteins, 1890 proteins carried PTM sites. Compared with the human proteome background, the enrichment was also significant ($X^2 = 1414.05$, $p < 0.01$). This suggests that complex-forming proteins have a stronger tendency to carry PTM sites. This finding can be partially explained by the fact that some PTMs are necessary for or could facilitate the formation of protein complex; it has been reported that phosphorylation and sumoylation are required for efficient spindle orientation complex formation by regulating Bim1p and Kar9p interaction.

To further test whether the Mtp-proteins have any special tendency in complex formation, we classified all human proteins into groups based on the number of PTM types embedded in individual proteins and detected the correspond-
increasing number of PTM types carried in the proteins, the complex-forming proteins in different protein groups. On the basis of the classification method mentioned in Figure 1, the human proteins were classified into different groups, as shown on the x axis. The y axis demonstrates corresponding percentage of complex-forming proteins in different protein groups. With an increasing number of PTM types carried in the proteins, the percentage of complex-forming protein also became higher.

Figure 2. Percentage of complex-forming proteins in different protein groups. On the basis of the classification method mentioned in Figure 1, the human proteins were classified into different groups, as shown on the x axis. The y axis demonstrates corresponding percentage of complex-forming proteins in different protein groups. With an increasing number of PTM types carried in the proteins, the percentage of complex-forming protein also became higher.

proteins carrying no known PTM sites ("group 0"), the percentage of complex-forming proteins was only 5%. The value increased to 18% in "group 1", 35% in "group 2", and 52% in "group 4". Significant differences were detected not only between "group 0" and "group 1" ($X^2 = 726.05, p < 0.01$) but also between "group 1" and "group >1" ($X^2 = 2067.13, p < 0.01$). It suggests that post-translationally modified proteins prefer to form complex more than those lacking PTM sites, and multiple types of PTMs confer Mtp-proteins with more capacity to form complex. Possibly, through harboring different types of PTMs, Mtp-proteins can bind to different partners and thereby improve the chance to form complexes. Integrating this to our aforementioned finding that Mtp-proteins are preferred to be involved in human disease, it is possible that the mutant forms of Mtp-proteins may dysfunction the corresponding complex and hence may be associated with disease. Accordingly, we propose a shift from gene/protein-centered disease research to Mtp-containing complex-centered disease research.

Exploratory Analysis on Mtp-Proteins: Mtp-Proteins Function As Hubs or Superhubs in PPI Networks

To examine the tendency of Mtp-proteins to develop more interactions, we collected the human PPI data and compared the percentage of proteins having partners and the number of partners among different protein groups. (See the Materials and Methods.) Our results showed that the percentage of proteins with known partners increased when the number of PTM types increased (Figure 3). For instance, only 54% proteins in "group 0" had partners, but the percentage increased to 90% in "group 2". Besides, with an increasing number of PTM types in proteins, the average number of partners also increased (Figure 3). Significantly more partner-bearing proteins in groups containing at least one type of PTM target were also observed than those did not contain any PTM targets ($X^2 = 1389.50, p < 0.01$). These suggest that compared with proteins containing no known PTM sites, Mtp-proteins seldom function individually but rather in a modular fashion. This is in concordance with the aforementioned finding that Mtp-proteins were more preferred to form complex. Apart from the single type of PTM event that affects the target protein’s interaction with its partners, a wide spectrum of PTMs acting as epigenetic-like codes for modulating specific functions of some proteins has also been reported. For instance, p53 harbors a multitude of PTM sites including phosphorylation, acetylation, methylation, ubiquitylation, and sumoylation, through binding to different partners. Hence, it is potentially a general code that PTM-site-bearing proteins have many more protein interactions and favor complex formation.

To further investigate the role of Mtp-proteins in protein interactions, we constructed PPI networks and compared the proportion of each node category, based on the number of partners to which a given protein can bind among different protein groups. The group with a single PTM type (group "1") occupied a significantly higher proportion of "Hub" than the group without any PTM (group "0") ($X^2 = 144.76, p < 0.01$) (Figure 4). However, it showed a significantly lower proportion in Peripheral-A ($X^2 = 36.84, p < 0.01$) and Peripheral-B ($X^2 = 4.5, 0.01 < p < 0.05$). No significant difference on Superhub was detected between groups "1" and "0" ($X^2 = 2.99, p > 0.05$). However, there was a significantly higher proportion of
Superhubs in group “2” than group “0” ($X^2 = 11.27, p < 0.01$). When the number of PTM types increased, an increasing proportion of Hub and Superhub was observed. The centrality-lethality rule states that deletion of a hub protein is more lethal than the deletion of a nonhub protein. Multiple modified proteins may play core roles in cellular biological processes, whereas those with no PTM sites, located at the periphery of the network, may be less essential. This further implies a critical role for Mtp-proteins in PPI and thereby in cell survival and may partially explain why Mtp-proteins are more prone to be involved in disease than other proteins.

Topological coefficient (Tc) is a measure of the extent to which a protein in a PPI network shares the same neighbors with other proteins. A large Tc value means a higher tendency of the protein to have shared neighbors. Our results showed that no matter whether a certain protein group carries PTM sites, the Tc value decreased with an increasing number of neighbors (Figure 5). This is consistent with the previous finding that hubs are more exclusive to interact with common partners than proteins with fewer neighbors. However, the Tc values decreased when the number of PTM types increased. The Tc values of proteins in group “0” were significantly higher than those in group “1” (average: 0.27 versus 0.21, ANOVA: $F = 241.08, p < 0.0001$). These suggest that proteins without PTM sites are preferred to interact with common partners, usually hubs, whereas proteins with multiple PTM sites are prone to interact with peripheral proteins. This, together with our aforementioned observation that Mtp-proteins prefer to act as hub, implies a concerted interplay between proteins with and

![Figure 4](image1.png)

Figure 4. Distribution of different types of nodes in different protein groups. On the basis of the classification method mentioned in Figure 1, the human proteins were classified into different groups, as shown in the x-axis. In the constructed PPI network, all nodes were classified into four categories (Peripheral-A: degree = 1; Peripheral-B: $20 >$ degree $> 1$; Hub: $100 >$ degree $\geq 20$; and Superhub: degree $\geq 100$). The y-axis shows the corresponding percentage of the different node types.

![Figure 5](image2.png)

Figure 5. Distribution of topological coefficient (Tc) in subnetworks constructed with different protein groups. On the basis of the classification method mentioned in Figure 1, the human proteins were classified into different groups, with which subnetworks were then constructed. Subnetwork “0” means that the network was constructed with protein group “0”, subnetwork “1” indicates that the network was constructed with protein group “1”, and so on. The x-axis shows the number of neighbors of a single node, and the y-axis demonstrates the topological coefficients.
The functional categories “regulation of phosphorylation” and “cytoskeleton” were significantly enriched in group “1”. Phosphorylation is the most dominant type of experimentally identified PTMs. Therefore, the enrichment of the function type “regulation of phosphorylation” in group “1” seems reasonable. Many types of PTMs, such as phosphorylation, acetylation, ubiquitination, sumoylation, and palmitoylation, are involved in the regulation of the microtubule cytoskeleton. The over-represented “cytoskeleton” in group “1” suggests that proteins related to cytoskeleton are frequently post-translationally modified with single types of PTM but not in a combinatorial manner. In groups with more than one type of PTM, the predominant functional categories included “cell cycle”, “RNA processing”, “translation”, “nucleotide binding”, and “chromatin assembly or disassembly”. This is consistent with some other recent studies that showed the important roles of interplay between PTMs of certain proteins in cell cycle progression, RNA processing, and translation regulation. Hence, apart from the consistence with the well-known histone code and TF code, our results suggest that the regulation of cell cycle, RNA processing, and translation are also major issues of coordinate interaction among multiple types of PTM and further demonstrate the potentially functional influences associated with Mtp-proteins in disease.

**Exploratory Analysis on Mtp-Proteins: Mtp-Proteins Demonstrate Distinct Functional Focus**

To comprehensively analyze potential functional biases of Mtp-proteins, we assigned functional annotations of different protein groups and computed the enrichment in Gene Ontology annotations assigned to different protein groups with respect to the total human proteome (See Materials and Methods). Different groups were shown to present distinct patterns of enriched functional categories (Table 1). Compared with the protein groups with PTM sites, the function category “intrinsically to membrane” was significantly enriched in the protein group without any PTM site \((p < 1 \times 10^{-9})\). This suggests that most membrane proteins do not undergo PTMs, although PTM of membrane proteins has been identified in some cases.67

The functional categories “regulation of phosphorylation” and “cytoskeleton” were significantly enriched in group “1”. Phosphorylation is the most dominant type of experimentally identified PTMs. Therefore, the enrichment of the function type “regulation of phosphorylation” in group “1” seems reasonable. Many types of PTMs, such as phosphorylation, acetylation, ubiquitination, sumoylation, and palmitoylation, are involved in the regulation of the microtubule cytoskeleton. The over-represented “cytoskeleton” in group “1” suggests that proteins related to cytoskeleton are frequently post-translationally modified with single types of PTM but not in a combinatorial manner. In groups with more than one type of PTM, the predominant functional categories included “cell cycle”, “RNA processing”, “translation”, “nucleotide binding”, and “chromatin assembly or disassembly”. This is consistent with some other recent studies that showed the important roles of interplay between PTMs of certain proteins in cell cycle progression, RNA processing, and translation regulation.60–72 Hence, apart from the consistence with the well-known histone code and TF code,71 our results suggest that the regulation of cell cycle, RNA processing, and translation are also major issues of coordinate interaction among multiple types of PTM and further demonstrate the potentially functional influences associated with Mtp-proteins in disease.

**Table 1. Enriched Functional Categories in Different Protein Groups**

<table>
<thead>
<tr>
<th>functional category</th>
<th>protein group*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007186−G-protein coupled receptor protein signaling pathway</td>
<td>√/√</td>
</tr>
<tr>
<td>GO:0031224−intrinsic to membrane</td>
<td>√</td>
</tr>
<tr>
<td>IPR013087:zinc finger, C2H2-type/integrase, DNA-binding</td>
<td>√/√</td>
</tr>
<tr>
<td>GO:0042325−regulation of phosphorylation</td>
<td>√</td>
</tr>
<tr>
<td>GO:0005856−cytoskeleton</td>
<td>√</td>
</tr>
<tr>
<td>GO:0051276−chromosome organization</td>
<td>×</td>
</tr>
<tr>
<td>GO:0007049−cell cycle</td>
<td></td>
</tr>
<tr>
<td>GO:0006396−RNA processing</td>
<td>√</td>
</tr>
<tr>
<td>GO:0006412−translation</td>
<td>√</td>
</tr>
<tr>
<td>GO:0005524−ATP binding</td>
<td>√/√</td>
</tr>
<tr>
<td>GO:0000166−nucleotide binding</td>
<td>√</td>
</tr>
<tr>
<td>GO:0031974−membrane-enclosed lumen</td>
<td>√/√</td>
</tr>
<tr>
<td>isopeptide bond</td>
<td></td>
</tr>
<tr>
<td>GO:0006333−chromatin assembly or disassembly</td>
<td></td>
</tr>
<tr>
<td>IPR007125:histone core</td>
<td></td>
</tr>
</tbody>
</table>

*Protein groups were classified according to the number of PTM types carried in the individual proteins. The “√” denotes that the particular functional category was significantly enriched in the protein group.

PTMs are required for subcellular localization of certain proteins. To examine whether proteins carrying PTM sites, in particular of multiple types, have preferential subcellular localizations, we predicted the subcellular localization of all proteins in the human proteome using Yloc (see Materials and Methods, Supplementary File 3a in the Supporting Information) and matched the results with the subcellular localizations of each protein group. The proportion of each category of subcellular localization among different protein groups was then compared. The results showed that compared with groups without PTM sites, protein groups carrying PTM sites had significantly different distribution of subcellular localizations \((X^2 = 180.53, p < 2.2 \times 10^{-16})\). Specifically, there were higher proportions of the categories cytoplasm \((X^2 = 298.34, p < 0.01)\) and nucleus \((X^2 = 205.41, p < 0.01)\) and lower proportions of extracellular space \((X^2 = 210.08, p < 0.01)\), plasma membrane \((X^2 = 305.42, p < 0.01)\), Golgi apparatus \((X^2 = 44.71, p < 0.01)\), peroxisome \((X^2 = 26.70, p < 0.01)\), and mitochondrion \((X^2 = 21, p < 0.01)\) (Figure 6). No significant differences were detected for the categories endoplasmic reticulum and lysosome. Similar results were obtained for the analysis using data from LocDB (Supplementary File 3b in the Supporting Information). In general, glycosylation plays important roles in the protein secretory pathway.77 Thus, secretory and membrane proteins should be enriched for glycosylation and possibly some additional PTMs. However, our data demonstrated that protein groups carrying PTM sites were associated significantly with lower proportions of extracellular space and plasma membrane. This may result from two aspects. First, because secreted proteins are usually at low concentrations in culture media and membrane-bound or membrane-associated proteins are associated with lipids, it is challenging to isolate or solubilize such proteins for current proteomics techniques.78,79 This may introduce bias against such proteins in our data set.
sorting or targeting of several proteins, such as the sodium-dependent purine-selective nucleoside transporter (SPNT) and the sodium-sulfate cotransporter Na/Si-1, is independent of glycosylation or lipid raft.

It suggests that PTM such as glycosylation is not necessary for protein sorting and partially accounts for the phenomena mentioned. Nonetheless, further studies are warranted to reveal why groups carrying PTM sites also had significantly lower proportions of Golgi apparatus, peroxisome, and mitochondrion. Indeed, nuclear and cytoplasmic localization of many proteins are mediated by PTM. It is also possible that some functions in nuclear and cytoplasm, such as microtubules formation in cell division and chromatin remodeling in transcriptional regulation, are PTM-dependent. The proportion of proteins located in the nuclear and cytoplasm categories were also observed to increase with an increasing number of PTM types. Ubiquitylated von Hippel–Lindau (VHL) tumor suppressor protein is located predominantly in the cytoplasm, but Sumoylation could result in increased VHL protein stability and nuclear redistribution.

Similar mechanisms are observed to regulate the CLOCK nuclear translocation in the circadian system. Hence, it seems that Mtp-proteins can be localized in double localizations (nuclear and cytoplasm) through PTM-dependent translocation.

Interestingly, with an increasing number of PTM types, the percentage of proteins with two and more than two localizations was also shown to increase significantly ($X^2 = 64.02, p < 0.01$ and $X^2 = 52.86, p < 0.01$, respectively), with a significant decrease in the percentage of proteins in single subcellular compartments ($X^2 = 130.18, p < 0.01$) and a significantly different whole distribution of location types ($X^2 = 235.30, p < 2.2 \times 10^{-16}$) (Figure 6). These, together with the remarkably higher number of partners observed, suggest that Mtp-proteins may play different roles in biological processes through not only binding to different partners but also locating in different subcellular localizations. It also implies that combinatorial multiple types of PTM may play a critical role in protein sorting and shuttling between nuclear-cytoplasm or even among multiple localizations. Notably, changes in the subcellular localizations of proteins such as CLN3 and TMEM165 have been found to be involved in disease, and protein localization is considered to be a principal feature of the etiology and comorbidity of genetic diseases. Thus, biased and multiple subcellular localizations of Mtp-proteins may increase the possibilities for introducing disease-caused mutations.

**Figure 6.** (a) Distribution of different types of subcellular localizations in different protein groups. On the basis of the classification method mentioned in Figure 1, the human proteins were classified into different groups, as shown in the $x$ axis. After annotating the subcellular locations with Yloc, the percentage of proteins with different subcellular localizations was calculated in the different protein groups, as shown in the $y$ axis. (b) Distribution of proteins with multiple subcellular localizations in different protein groups. Similarly, the $x$ axis indicates the different protein groups. On the basis of the annotation of proteins with multiple subcellular locations retrieved from the LocDB database, the percentage of proteins with multiple subcellular localizations was calculated in the different protein groups, as shown in the $y$ axis.

**Exploratory Analysis on Mtp-Proteins: Intrinsically Disordered Region Is Important for the Special Features of Mtp-Proteins**

Intrinsically disordered regions (IDRs), which do not have a stable structure under native conditions, have been shown to be associated with a variety of PTM sites. For instance, protein phosphorylation occurs predominantly within IDRs, disordered regions are important for substrate binding and efficient degradation in ubiquitylation, acetylated and methylated lysines are located in the disordered regions, and hydroxylation of proline residues is facilitated by IDR. To examine the frequency of PTM target sites located in IDRs, we detected IDRs of the human proteome, to which target sites of PTM were then mapped (Supplementary File 4a,b in the Supporting Information). The results showed that 30 511 out of 38 264 (80%) of the target sites were embedded in IDRs. Only 23% of them were located in the short IDRs, and the remaining were located in long IDRs. Our data showed that ~80% of the PTM events occurred in IDRs and hinted the importance of IDRs for the special features of Mtp-proteins.

The percentages of the IDR locating target residues of seven dominating PTM types were calculated and compared to determine what types of PTM targets are more prone to be located in IDRs. The results showed that the target sites of phosphorylation, hydroxylation, sulfation, and amidation were over-represented (>70%) in IDRs, whereas acetylation, N-linked glycosylation, and methylation were less represented (<60%) in IDRs (Figure 7). The long IDRs accounted for most of the target sites of phosphorylation (84%), hydroxylation (99%), and sulfation (57%) and accounted for fewer in acetylation (37%), N-linked glycosylation (23%), methylation (7%), and amidation (28%). These showed that although most PTM events occurring in IDRs have been found on the whole, different PTM types have different preferences. To act as a functional switch only by modifying IDRs is not enough, and modification of the ordered regions is sometimes needed. It is suggested that phosphorylation can control the functions of proteins through causing disorder-to-order transitions. Our results suggest that apart from the disorder-to-order transition, which might be a general mechanism of many PTM sites such as hydroxylation and sulfation, unfolding factors induced by PTM events biased in ordered regions such as acetylation, N-linked...
glycosylation, and methylation might also perturb the folded domains. Also, chaperone proteins and the protein structure might control the accessibility of target residues to the enzymes required for certain PTMs. The conformational constraints may partially explain the predominance of a few PTM types on the modification of target residues within the ordered regions. It is also worth mentioning that our observation here is not consistent with a previous report, which found that acetylation was found to dominate IDR from 77 N-acetyllysine residues. This incongruence may result from the limited amount of data used for analysis in that study.

To test whether PTM site-bearing proteins, particularly Mtp-proteins, contain more IDRs, the content of IDRs among different protein groups was compared. The results showed that the average numbers of IDRs were significantly higher in the protein group ≥1 (combining groups 1, 2, 3, and 4 into one group for comparison) than the group without PTM, and in an increasing trend (Figure 8) (t-test, $p = 5.80 \times 10^{-66}$). However, there was no significant difference between protein groups with more than four PTM types and the group without PTM ($t$ test, $p = 0.116$). To examine the observed turning of the trend, we further investigated proteins with more than four PTM types. The results indicated that 306 out of 423 (72%) of the target sites in these groups were located in IDRs. This value is significantly lower than the foregoing 80% detected on the whole ($X^2 = 14.13, p < 0.01$). When considering the target sites in only long IDRs, the results were even more significant: only 37% were located in the long IDRs ($X^2 = 472.79, p < 0.01$). We also found that these groups had a significantly higher density of target sites in individual IDR than on the whole (average density: 4.37 versus 2.72; $t$ test, $p < 0.001$). These suggest that proteins with few IDRs can also be modified by multiple PTM types through increasing the number of target sites in the ordered regions and the density of target sites in a single IDR. Strikingly, proteins related to "DNA packaging", "protein–DNA complex assembly", and "chromatin assembly or disassembly" were found to be significantly enriched in these special groups ($p < 1.0 \times 10^{-30}$). These suggest that not all Mtp-proteins form functional switch by changing their protein partners; some of them might function as switch by transforming the status of RNA or DNA binding. Thus, Mtp-proteins may be associated with disease from aberrant Mtp-DNA interactions but not just abnormal PPIs.

**Exploratory Analysis on Mtp-Proteins: Energetic Effects of PTMs on the Stability of Protein–Protein Binding**

It has been shown that phosphorylation affects the stability of protein complexes through changing their binding energy. To examine the energetic effects of other types of PTM, particularly those caused by potential combinatorial interplay between multiple PTMs, on the stability of protein–protein binding, we mapped target sites of different PTMs onto protein structures retrieved from the Protein Data Bank. Alanine scanning was also performed to estimate the contribution of target residues to complex stability (Supplementary File 5 in the Supporting Information). For all PTM types analyzed, substitutions of the target sites with alanine did not alter the binding energy much, with the average difference in binding energy ranging from −0.6 to +0.9 kcal/mol (Table 2). This suggests that the individual PTM types analyzed may not affect PPI through changing the binding energy significantly. Different modified residues can be specifically recognized by different protein modules, such as pSer/Thr binding by PBD and aceLys binding by Bromodomain. So, through creating or altering the docking sites for protein modules, the PTM types analyzed may regulate the recognition or selectivity in protein–protein binding.

Although the majority of target site substitutions did not contribute significantly to the complex stability, some of them could still change the binding energy by more than +2 kcal/mol (destabilizing the complex) or −2 kcal/mol (stabilizing the complex) (Table 3). A significantly higher proportion of substitutions in phosphorylation sites were observed to alter the binding energy by more than +2 kcal/mol ($X^2 = 284.55, p < 0.01$), but compared with N-linked glycosylation and methylation, a significantly lower proportion of substitutions on phosphorylation sites affected the binding energy by more than −2 kcal/mol ($X^2 = 26.60, p < 0.01$). Hot spots are defined as the residues comprising only a small fraction of interfaces yet accounting for the majority of the binding energy, and alanine scanning mutagenesis causes large differences in binding energy by usually more than 2 kcal/mol. Thus, our results indicate that phosphorylation is a major cause of energetic effects on the stability of protein complexes.
multiple PTMs acting on a single protein can enhance the e
and amplify the signal in cellular signaling networks by a
= 125.99, 
examined binding energy changes on the combinatorial 
combinatorial interplay between multiple PTMs, we also 
disease.
in normal cellular processes as well as in the development of 
that the greatest change in binding energy was obtained by 
File 5b in the Supporting Information). The results showed 
interaction between phosphorylation and acetylation through 
target sites.
which implies that the evolutionary optimization acts not only on the 
residues of PTMs are gradually optimized during evolution, 
ments on stability are considered to be signi
significantly higher than any single role of phosphorylation and 
Accordingly, it is unwise to increase binding 
restraints from not only the physicochemical feasibility but also 
component for the number of interactions (r = 0.622). Although a single protein could be extremely large and can have surface areas of thousands of Å² units, it receives 
Our results showed that Mtp-proteins were signi
enriched in protein complexes and preferred to act as hubs in 
PPI networks with the capability of producing more interactions. How can Mtp-proteins bind so many different protein partners?
sequences of the Mtp-proteins might be one of the reasons. 
We found that PTM-containing proteins were longer than 
those that do not contain PTMs (average 728 residues versus 
458 residues). However, by studying the Pearson correlation 
coe cient between the length of protein and the corresponding 
number of partners, it reveals that the length is not a major 
component for the number of interactions (r = 0.006; p = 
0.622). Although a single protein could be extremely large and can have surface areas of thousands of Å² units, it receives 
restraints from not only the physicochemical feasibility but also 
the cell and genome. Obviously, it is unforeseeable to increase binding sites simply by increasing the protein length. Moreover, we detected that most PTM events were embedded in IDRs and most Mtp-proteins carried many IDRs. Indeed, disorder in a protein can increase the kinetics of PPI and complex formation through enhancing the capture radius of the chain and also 
elonging solvent-accessible surfaces compared with globular 
folded protein.108,109 IDRs can also signal through highly specific but short-lived association.110 It is comprehensible that IDRs can produce a broad range of different conformations, and little energy is needed to switch among them. Hence, IDRs themselves can facilitate or impact PPI. Meanwhile, IDRs can not only provide implantation positions for PTMs but also allow more pronounced effects of PTMs with their own characteristics.
PTM itself might be another reason. Indeed, many 
interactions are turned on/off in a PTM-dependent man
18,111 However, by examining the energetic effects of PTMs on the stability of protein–protein binding, we found that most single PTM events were only correlated with limited effects on stability, while combinatorial PTMs could exert a significantly higher effect. On the basis of the observation that 
PTM sites with known functions are more conserved, it is 
speculated that a substantial fraction of PTM sites are 

Table 2. Change of Binding Energy Caused by Alanine Substitution in Different Residue Groups

<table>
<thead>
<tr>
<th>residue group</th>
<th>phosphorylation</th>
<th>acetylation</th>
<th>methylation</th>
<th>N-linked glycosylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>0.10</td>
<td>−0.13</td>
<td>−0.10</td>
<td>−0.49</td>
</tr>
<tr>
<td>R2</td>
<td>0.26</td>
<td>−0.07</td>
<td>−0.22</td>
<td>−0.63</td>
</tr>
<tr>
<td>R3</td>
<td>0.79</td>
<td>0.87</td>
<td>0.26</td>
<td>0.83</td>
</tr>
</tbody>
</table>

“ΔG is the difference of binding energy by substituting the residues with alanine in the target complexes of PTM types analyzed. "R1" represents the residues replaced being the target residues of different PTMs; "R2" represents the residues replaced being identical to the target residues but not located on the target sites; and "R3" represents the residues in the target complexes being considered as a whole without taking into account what the residue is or which site it is on. 'Based on t test, "**" denotes p < 0.001.

Table 3. Percentage of Alanine Substitution Contributing Significantly to the Complex Stability

<table>
<thead>
<tr>
<th>change of ΔG (kcal/mol)</th>
<th>phosphorylation</th>
<th>acetylation</th>
<th>methylation</th>
<th>N-linked glycosylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;−2 (complex stabilizing)</td>
<td>4.67%</td>
<td>3.60%</td>
<td>7.85%</td>
<td>11.35%</td>
</tr>
<tr>
<td>1−2</td>
<td>10.24%</td>
<td>3.60%</td>
<td>4.77%</td>
<td>7.80%</td>
</tr>
<tr>
<td>&gt;2 (complex destabilizing)</td>
<td>11.95%</td>
<td>2.48%</td>
<td>0.22%</td>
<td>4.26%</td>
</tr>
</tbody>
</table>

Effect on stability is considered to be significant when the change of binding energy is >2 kcal/mol.

suggest that phosphorylation events prefer to target hot spots for stabilizing the complex, while N-linked glycosylation and methylation tend to be involved in sites for complex destabilizing.

Additionally, there was a significant difference among the binding energy changes by substitutions in the target sites (Group R1) in different PTM types (ANOVA: F = 82.84, p < 0.0001) (Table 2). This further suggests that different PTM types are biased toward increasing or reducing the stability of PPI. The changes in binding energy were also significant among the three groups. On one hand, this suggests that the target residues of PTMs are gradually optimized during evolution, which is congruent with previous findings.107 On the other hand, a significant difference between R2 and R3 further implies that the evolutionary optimization acts not only on the target sites of PTMs, but also on the residues, which are the same as the target amino acid, although they are not the known target sites.

To evaluate the energetic contribution of potential combinatorial interplay between multiple PTMs, we also examined binding energy changes on the combinatorial interaction between phosphorylation and acetylation through modeling after arousing multiple mutations (Supplementary File 5b in the Supporting Information). The results showed that the greatest change in binding energy was obtained by substituting one phosphorylation site and one acetylation site at 
the same time. The combined effect on binding energy was significantly higher than any single role of phosphorylation and acetylation (Combined: 1.322 kcal/mol, phosphorylation only: 0.098 kcal/mol, acetylation only: −0.134 kcal/mol; ANOVA: F 125.99, p < 0.0001). This suggests that a combinatorial interplay between different PTMs may dramatically affect the stability of PPI. Thus, not only through creating or changing the docking sites for protein modules to confer function switch, multiple PTMs acting on a single protein can enhance the effect and amplify the signal in cellular signaling networks by affecting the binding energy. This further elucidates why Mtp-proteins, as key elements in cellular signaling networks, are more pivotal in normal cellular processes as well as in the development of disease.

DISCUSSION

Our results showed that Mtp-proteins were significantly enriched in protein complexes and preferred to act as hubs in PPI networks with the capability of producing more interactions. How can Mtp-proteins bind so many different protein partners?

Sequences of the Mtp-proteins might be one of the reasons. We found that PTM-containing proteins were longer than those that do not contain PTMs (average 728 residues versus 458 residues). However, by studying the Pearson correlation coefficient between the length of protein and the corresponding number of partners, it reveals that the length is not a major component for the number of interactions (r = 0.006; p = 0.622). Although a single protein could be extremely large and have surface areas of thousands of Å² units, it receives restraints from not only the physicochemical feasibility but also the cell and genome. Obviously, it is unforeseeable to increase binding sites simply by increasing the protein length. Moreover, we detected that most PTM events were embedded in IDRs and most Mtp-proteins carried many IDRs. Indeed, disorder in a protein can increase the kinetics of PPI and complex formation through enhancing the capture radius of the chain and also enlarging solvent-accessible surfaces compared with globular folded protein.108,109 IDRs can also signal through highly specific but short-lived association.110 It is comprehensible that IDRs can produce a broad range of different conformations, and little energy is needed to switch among them. Hence, IDRs themselves can facilitate or impact PPI. Meanwhile, IDRs can not only provide implantation positions for PTMs but also allow more pronounced effects of PTMs with their own characteristics.

PTM itself might be another reason. Indeed, many interactions are turned on/off in a PTM-dependent manner.15–18,111 However, by examining the energetic effects of PTMs on the stability of protein–protein binding, we found that most single PTM events were only correlated with limited effects on stability, while combinatorial PTMs could exert a significantly higher effect. On the basis of the observation that PTM sites with known functions are more conserved, it is speculated that a substantial fraction of PTM sites are
brane-associated proteins are di
class mass spectrometric analysis of membrane-bound and mem-
peptide ions by electrospray ionization, sample preparations for
Thus, despite only weak influence on the binding energy by single PTM events, multiple PTMs acting on a single protein could enhance the effect and amplify the signal in cellular signaling networks.
So, the sequence itself and PTM harbored in Mtp-proteins can endow the Mtp-proteins more important roles, such as acting as hub in PPI, partially explaining why these proteins are more prone to be associated with diseases. However, to what extent other factors, such as the concentration of partners, a wider expression profiling of Mtp-proteins, and even the cell environment, can impact PPI, especially in the development of disease, still warrants further studies.
Alternative splicing plays a key role in the expansion of proteome diversity and organismal complexity. Current estimates suggest that ~90% of human genes are alternatively spliced.116,117 Recently, distinct splice variants and pathway were found to be enriched in different breast cancer subtypes; many of such variants were not completely matched to any known protein sequence and were found only in tumor samples.118 Through using PPI assay, Ellis et al. showed that regulated alternative exons frequently remodel interactions to establish tissue-dependent PPI networks.120 These examples illustrate the specific roles of splice variants in tissue differentiation and disease progression as well as their potential as biomarkers for diagnosis and drug target discovery. Interestingly, tissue-specific spliced segments often contain disordered regions and are enriched in PTM sites.121 This not only demonstrates that splice isoforms probably have different PTM patterns but also would increase the regulatory dimension above the system level: transcriptomic diversity may result in proteomic rewiring through arousing changes in PTM sites. Thus, those isoforms differentially expressed in a particular disease are likely to trigger or participate in diseases through altering PTM sites and then the corresponding regulatory circuits. It is possible that multiple PTM types allow aberrant splicing to produce dysfunctional isoforms, accounting for the association of Mtp-proteins with diseases. However, it is still challenging to characterize the functions of different isoforms to date. To differentiate functions of alternatively spliced isoforms, many computational methods have been developed, for instance, through interrogating public RNA-seq data at the transcript level or integrating proteomic analysis with structure-based conformational predictions.122–124 Considering that alternative splicing may change PTM sites, we propose to integrate PTM data in future computational predictions to reveal the functions of splice variants.
Another point that we would like to note here is that there may be some biases in our PTM data sets. First, because the association with lipids would impede the isolation and solubilization of proteins in buffers suitable for mass spectrometry and efficient generation of positively charged peptide ions by electrospray ionization, sample preparations for mass spectrometric analysis of membrane-bound and membrane-associated proteins are difficult.79 Although many improved methods have been developed,125–127 such bias is unavoidable. Second, for the detection of PTMs using mass spectrometry, a very high peptide coverage of the modified peptides is needed. This, in turn, is dependent on the abundance of proteins and may therefore result in a bias against proteins of low abundance. To examine to what extent protein abundance could influence our observed pattern, we compiled a data set containing proteins with abundance data derived from Schwannhauser et al.128 (see Supplementary File 6a in the Supporting Information) and analyzed the proteins with a high abundance (higher than the median abundance). The result demonstrated a trend similar to our finding (Supplementary File 6a, Figure S1 in the Supporting Information). Specifically, for protein “group 0”, in which all proteins carry no known PTM event, the proportion of disease-related proteins involved was only 47%. The value increased to 74% in “group 2” and up to 93% in “group 5”. Significant differences were detected not only between “group 0” and “group 1” (X2 = 9.72, p < 0.01), but also between “group 1” and “group >1” (X2 = 6.79, p < 0.01). These suggest that for the data set containing proteins with a high abundance our reported trend is valid. Moreover, we also analyzed the whole data set with both highly and less abundant proteins, and a similar trend was also observed (Supplementary File 6a, Figure S2 in the Supporting Information). These results provide support to our findings. Perhaps, because the expression levels of proteins are changing in both the temporal and spatial scales, only very rarely there will be proteins whose expression level is always low. Also, UniProt is the central hub for the collection of functional information on proteins but not a single snapshot or data set of one particular experiment. As such, the bias against proteins of low abundance should be buffered to some extent and may not be a major component for the result in our study.

CONCLUSION
In conclusion, we found that proteins harboring target sites of combinatorial multiple PTM types are preferred to be associated with disease. By integrating proteomic, protein complex, PPI network, functional annotation, subcellular localization, and binding energy information, we further analyzed the features and functions of Mtp-proteins. Many particularities of Mtp-proteins were revealed, including their enrichment in complex, preference to act as hubs in PPI networks, biased functions, and preferred subcellular localizations. These analyses not only demonstrate the particularities of Mtp-proteins but also explain why Mtp proteins are prone to be associated with diseases. Many PTMs are tracked as cancer markers or molecular targets for disease therapy, such as methylated HSP70 and acetylation of the surface tumor antigens.36,129 Thus, our study, on one hand, explores broader and more in-depth understanding of Mtp-proteins and, on the other hand, provides ground for developing molecular targets in many life-threatening diseases therapies.

ASSOCIATED CONTENT
Supporting Information
Change in binding energy caused by alanine substitution in different protein groups. Supplementary File 5b. Change in binding energy through modeling after arousing multiple mutations. Supplementary File 6a. Construction methods of data set containing proteins with abundance information and corresponding analysis results. Supplementary File 6b. Data set of proteins with abundance information. This material is available free of charge via the Internet at http://pubs.acs.org.

**REFERENCES**


