Neurocomputing 206 (2016) 13–18

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A systematic exploration of the associations between amino acid variants and post-translational modifications

Gui-Min Qin [a,b], Yi-Bo Hou [c], Xing-Ming Zhao [a,*]

[a] Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[b] School of Software, Xidian University, Xi'an 710126, China
[c] School of Computer Science and Technology, Xidian University, Xi'an 710126, China

ARTICLE INFO

Article history:
Received 9 August 2015
Received in revised form
8 October 2015
Accepted 9 November 2015
Communicated by Chennai Guest Editor
Available online 30 May 2016

Keywords:
Amino acid variants
Cross-talks
Diseases
Post-translational modifications

ABSTRACT

The post-translational modifications (PTMs) are important to protein activities and play key roles in various kinds of biological processes. It has been well recognized that the alteration of PTMs may lead to diseases. However, it is still not clear how the PTMs are related to diseases. In this work, we systematically investigated the associations between PTM residues and disease related mutations. In particular, the common PTMs, including phosphorylation, ubiquitylation and acetylation, were considered here. By analyzing the PTM sites and the amino acid mutations, we found that the amino acid mutations co-occurring with PTM sites and PTM cross-talks tend to be deleterious mutations in diseases. Specifically, the deleterious single amino acid variants (SAVs) associated with cancer and muscular diseases tend to co-occur with phosphorylation residues. The PTM sites of proteins from nuclear envelope, protein complex and lysosome were found to be more likely adjacent to the deleterious amino acid variations. These findings provide insights into the associations between PTMs and diseases, and can help identify novel disease genes in the future.

1. Introduction

Protein post-translational modifications (PTMs) modulate protein activities and functions, and are widely involved in various biological processes. With the advance in high-throughput technologies, more and more PTM data are being generated. Accordingly, more and more PTM types have been recognized. Until now, more than 200 different types of PTMs have been recorded, ranging from small chemical modifications, e.g. phosphorylation and acetylation, to the change of complete proteins, such as ubiquitylation. These distinct PTMs are further found to have extensive cross-talks between each other [1,2]. For example, Noort et al. found that the deletion of two putative N-acetyltransferases disrupts protein phosphorylation in *Mycoplasma pneumoniae* [3]. Recently, it has been found that some amino acid mutations could alter the process of PTMs, e.g. some mutations could induce insertion or deletion of phosphorylation sites [4]. It has also been found that dysfunctional PTMs may lead to diseases [5–9]. For example, the aberrant phosphorylation modifications of some proteins were found to be related to cancer, especially for the regulator proteins of cell proliferation and transcription as well as the tumor suppressor TP53 [5,10]. In addition, the aberrant acetylation and ubiquitylation of some proteins were also found to be associated with cancer [11,12].

In the past years, the association between certain disease associated single amino acid variants (SAVs) and particular PTMs has attracted increasing attentions, where the SAVs may affect adjacent PTM residues and alter the PTM processes [4,13–17]. For example, Ser-16 of phenylalanine hydroxylase (PAH) is a phosphorylation site, the mutation of which causes phenylketonuria [16]. It has been found that the SAVs close to phosphorylation sites are more disruptive to the functions of proteins than ordinary mutations [13]. Thus Reimand et al. hypothesized that cancer drivers might have unexpected mutation rates in phosphorylation-associated regions within proteins [14]. In addition, the mutations of androgen receptor (AD) acetylation sites were found to cause Kennedy's disease [15]. Recently, Ren et al. computationally detected phosphorylation related SNPs with GPS tool and found that a large proportion of non-synonymous SNPs might affect protein phosphorylation [4]. By systematically investigating the relationship between cancer and phosphorylation SNPs [17], Wang et al. found that phosphorylation SNPs might influence cancer susceptibility by reconfiguring phosphorylation signaling. Despite these efforts, most of previous analysis only focused on a certain disease or PTM type, and a comprehensive exploration of the associations between PTM sites and SAVs is required.

* Corresponding author.
E-mail address: xm_zhao@tongji.edu.cn (X.-M. Zhao).

http://dx.doi.org/10.1016/j.neucom.2015.11.106
0925-2312/

In this work, we systematically investigated the associations between post-translational modifications and various diseases related amino acid mutations. In particular, three common PTM types, i.e. phosphorylation, ubiquitylation and acetylation, were considered. Except for single PTMs, the cross-talks between distinct PTMs were also considered here. We found that the SAVs co-occurring with PTMs and PTM cross-talks were more likely to be deleterious SAVs and cause diseases. These findings provide new insights into the molecular mechanisms underlying diseases, and may help identify new disease genes in the future.

2. Methods and materials

2.1. Data sets

The experimentally verified phosphorylation, ubiquitylation and acetylation residues were extracted from 7 public databases, including PhosphoSitePlus [18], PHOSIDA [19], Phospho.ELM [20], UniProt [21], HPRD [22], dbPTM [23] and CPLM [24]. Although the dbPTM database integrates PTM sites from 11 public databases and those manually curated from literature, the other 6 databases provide complementary PTM residues. Those PTM residues that cannot be mapped to manually annotated and reviewed proteins from UniProt were discarded. As a result, we obtained 196,765 phosphorylation residues of 16,865 proteins, 62,435 ubiquitylation residues of 9908 proteins and 22,990 acetylation residues of 7674 proteins. Table 1 shows the statistics of the PTM residues collected.

The SAV data were downloaded from Uniprot (released date: 09–Jul–2014) [21], including 24,727 deleterious SAVs and 37,935 polymorphisms SAVs in 12,537 Uniprot/Swiss-Prot proteins. These 24,727 deleterious SAVs were further divided into distinct groups according to the disease types defined in [25]. In addition, for the proteins considered here, we took into account their cellular component information which may help better understand how the PTMs are related to diseases. The annotations for proteins were downloaded from the Gene Ontology database [26]. We adopted the annotations from the Generic GO slims which give a broad overview of the gene ontology content without detailed GO terms. Here, 21 cellular components with each containing more than 100 proteins were kept for further analysis.

2.2. Associations of post-translational modified proteins with diseases

We firstly investigated the associations between PTMs and SAVs by checking whether they occur together. An SAV was regarded to co-occur with a PTM residue if it was located in the flanking window of the PTM residue, where the 15 amino acids window was defined to cover both the upstream and downstream 7 residues of the PTM site [4,14]. The SAVs that co-occur with phosphorylation residues were denoted as pSAVs, and those associated with acetylation and ubiquitylation were denoted as aSAVs and uSAVs, respectively. Especially, if an SAV is deleterious, its co-occurring PTM residues would be regarded as deleteriously altered by this mutation.

Since one protein may contain multiple SAVs and PTM residues, we defined a score to evaluate whether a protein with a certain PTM type tends to be related to diseases. Here, we took a phosphorylated protein as an example.

$$Score_{PhoSAV} = \frac{N_{PhoSAV}}{L} \tag{1}$$

Where $Score_{PhoSAV}$ is a score used to represent the likelihood of the phosphorylated protein related to diseases, N_{PhoSAV} is the number of pSAVs in the corresponding protein, and L is the length of the protein sequence. Since a protein with long sequence is more likely to have more phosphorylation residues, we took into account the length of the sequence so that the length bias can be avoided. These scores were further normalized to [0, 1] as described below.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Where x_{min} and x_{max} are the minimum and maximum scores obtained with Eq. (1), respectively. The scores of ubiquitylation and acetylation modified proteins, i.e. $Score_{UbiSAV}$ and $Score_{AceSAV}$, were defined in a similar way. Accordingly, similar scores were defined and normalized for proteins containing ordinary SAVs, regardless of whether these SAVs co-occur with PTM residues.

To further investigate how the SAVs co-occurring with PTM residues are related to diseases, we investigated the correlation between SAVs and PTM residues. For each protein and a specific PTM type, we defined N_A as the number of all PTM residues in the protein, N_M as the number of PTM sites with adjacent SAVs, and N_D as the number of PTM sites with adjacent deleterious SAVs within the protein. Assuming n is the number of proteins containing phosphorylation residues, the following vectors were defined.

$$\mathrm{AN} = (N_{A_1}, N_{A_2}, \cdots, N_{A_n}) \tag{3}$$

$$\mathrm{MN} = (N_{M_1}, N_{M_2}, \cdots, N_{M_n}) \tag{4}$$

$$\mathrm{DN} = (N_{D_1}, N_{D_2}, \cdots, N_{D_n}) \tag{5}$$

Then the Pearson correlation coefficients (PCCs) were calculated for the pairs of AN and MN as well as the pairs of MN and DN. We also calculated the PCCs for AN and DN.

2.3. Associations of PTM cross-talks with diseases

Since the PTM residues adjacent to each other tend to be functionally interrelated [27], we further investigated the PTM cross-talks to see whether their mutations were related to diseases. In this paper, a cross-talk was regarded to exist between a pair of distinct PTM residues if one of them is located in the flanking region of the other. The 21 amino acids window size was defined for a PTM residue to cover both the upstream and downstream 10 residues [28], and another PTM type would be regarded to cross-talk with this one if its PTM residue is located in the window. We investigated whether the PTM cross-talk co-occurring SAVs were significantly enriched in deleterious SAVs with Fisher's exact test.

Table 1
Statistics of PTM sites collected.

PTM Type	Databases	PTM data	# of proteins of distinct PTM residue types
Phosphorylation	PhosphoSitePlus, PHOSIDA, Phospho. ELM, UniProt, HPRD, dbPTM	# of proteins:16,865 # of residues:196,765	Ser: 111,384 Tyr: 35,229 Thr: 50,129
Ubiquitylation	PhosphoSitePlus, UniProt, dbPTM, HPRD, CPLM	# of proteins: 9908 # of residues: 62,435	Lys: 62,434
Acetylation	PhosphoSitePlus, PHOSIDA, UniProt, HPRD, dbPTM, CPLM	# of proteins: 7674 # of residues: 22,990	Lys: 20,722 Met: 787 Ser: 424 Thr: 87