Journal of Structural Biology 191 (2015) 39–48

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Tissue-specific sequence and structural environments of lysine acetylation sites

Nermin Pinar Karabulut [a], Dmitrij Frishman [a,b,c,*]

[a] Department of Genome Oriented Bioinformatics, Technische Universität München, 85354 Freising, Germany
[b] Helmholtz Zentrum Munich, German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, 85764, Neuherberg, Germany
[c] St Petersburg State Polytechnical University, St Petersburg 195251, Russia

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form 29 May 2015
Accepted 1 June 2015
Available online 3 June 2015

Keywords:
Evolution
Sequence analysis
Protein structure
Posttranslational modifications

ABSTRACT

Lysine acetylation is a widespread reversible post-translational modification that regulates a broad spectrum of biological activities across various cellular compartments, cell types, tissues, and disease states. While compartment-specific trends in lysine acetylation have recently been investigated, its tissue-specific preferences remain unexplored. Here we present a comprehensive tissue-based analysis of sequence and structural features of lysine acetylation sites (LASs) based on the recent experimental data of Lundby et al. (2012). We show that acetylated substrates are characterized by tissue-specific motifs both in linear amino acid sequence and in spatial environments. We further demonstrate that the general tendency of LASs to reside in ordered regions and, specifically, in α-helices, is also subject to tissue specific variation. In line with previous findings we show that LASs are generally more evolutionarily conserved than non-LASs, especially in proteins with known function and in structurally regular regions. On the other hand, as revealed by metabolic pathway analysis, LASs have diverse cellular functions in different tissues and are frequently associated with tissue-specific protein domains. These findings may imply the existence of tissue-specific lysine acetyltransferases (KATs) and lysine deacetylases (KDACs).

1. Introduction

Lysine acetylation is a reversible posttranslational modification (PTM), which involves the transfer of an acetyl group to the epsilon-amino group of a lysine residue of the substrate protein. This modification was previously only known to target histones, but more recently a broad spectrum of proteins was identified as acetylated and de-acetylated by lysine acetyltransferases (KATs) and lysine deacetylases (KDACs), respectively, underscoring the important role played by lysine acetylation in diverse cellular processes including the regulation of subcellular localization, protein stability, enzymatic activity, nucleic acid binding, and protein–protein interactions. Studies of lysine acetylation mechanisms moved into the scientific limelight ever since their association with major diseases, such as cancer, was discovered.

Recent advancements in high-resolution mass spectrometry-based proteomics have led to identification of thousands of lysine acetylation sites (LASs) (Henriksen et al., 2012), rendering possible proteome-wide *in silico* analyses of their sequence context as well as theoretical predictions of LASs (Basu et al., 2009; Hou et al., 2014; Lu et al., 2011; Shao et al., 2012; Suo et al., 2012). Currently available data reveal significant diversity of amino acid sequences surrounding lysine acetylation sites, making it difficult to derive consensus acetylation motifs. This diversity might be due to the broad variety of KATs and KDACs encoded, for example, in the human and mouse genomes (22 KATs and 18 KDACs) as well as to non-enzymatic lysine acetylation (Choudhary et al., 2014). Most of the LASs known today have not yet been associated to their cognate KATs and KDACs due to the technical challenges in detecting KAT- and KDAC-specific acetylation sites by high-throughput *in vitro* acetylation assays. To close this gap, Li et al. made a commendable effort in manually assigning 384 known LASs to three selected KAT families (Li et al., 2012), which, however, is still a far cry from close to 5000 experimentally confirmed LASs known from literature as of 2012.

Beyond linear sequence motifs, it has been hypothesized that the local structural environments of lysines can influence their predisposition to be recognized by KATs. Indeed, Kim et al. (2006) found that in mouse proteins, acetylated lysines prefer α-helical

* Corresponding author at: Wissenschaftszentrum Weihenstephan, Germany. Department of Genome Oriented Bioinformatics, Technische Universität München, 85354 Freising, Germany.
E-mail address: d.frishman@wzw.tum.de (D. Frishman).

http://dx.doi.org/10.1016/j.jsb.2015.06.001
1047-8477/

conformation, avoid disordered regions, and typically reside on protein surface. At the same time Okanishi et al. (2013), while confirming the tendency of acetylated lysines to be exposed, did not find any relationship between acetylation propensity and local secondary structure in *Thermus thermophilus*. Both studies were performed on rather limited datasets of acetylation sites. Recent availability of much larger proteome-wide acetylation assays warrants a deeper look into the role of structure in shaping the substrate spectrum of KATs.

The enzymes that catalyze the PTM events have different expression levels in different tissues and cellular compartments. Comprehensive studies of protein glycosylation (Kaji et al., 2012), phosphorylation (Lundby et al., 2012a) and acetylation (Lundby et al., 2012b) revealed thousands of differentially modified sites, opening up the possibility that PTM sites may possess substantially different sequence and spatial properties across tissues, depending on which particular enzyme catalyzes a particular modification event. The existence of compartment-specific sequence signatures for phosphorylation (Chen et al., 2014; van Wijk et al., 2014) and lysine acetylation (Choudhary et al., 2009; Kim et al., 2006; Lundby et al., 2012b; Shao et al., 2012) has already been firmly established.

Here we present the first comprehensive analysis of global and tissue-specific sequence and structure properties of LASs based on recent experimental data presented by Lundby et al. (2012b). We assessed the extent of evolutionary conservation of LASs and its dependence on functional and structural properties of proteins by comparing rat, mouse, and *Caenorhabditis elegans* acetylomes. We further investigated tissue-specific functional roles and domain preferences of acetylated proteins.

2. Materials and methods

2.1. Data collection and preprocessing

The dataset used in our analysis contains 15,474 lysine acetylation sites (LASs) in 4541 proteins identified by high-resolution tandem mass spectrometry in 16 rat tissues: brain, heart, muscle, lung, kidney, liver, stomach, pancreas, spleen, thymus, intestine, skin, testis, testis fat, perirenal fat, and brown fat (Lundby et al., 2012b). For each lysine-acetylated peptide in each tissue we obtained information about the UniProt (Consortium, 2014) IDs of the best-matching proteins (one or more), the sequence position of the acetylated site, and the intensity values (summed up extracted ion current of all isotopic clusters associated with the peptide in the corresponding tissue).

In order to find the best-matching UniProt ID for each acetylated peptide we applied the following procedure: (i) All fragments were excluded from consideration. (ii) If there was only one UniProt ID associated with an acetylated peptide, and its sequence position and the sequence of the corresponding full-length protein in the UniProt database were known, then we directly used that protein. (iii) Otherwise, we aligned all pairs of proteins and then chose the pair having the maximum sequence identity out of all pairs sharing at least 90% sequence identity. The idea behind this approach is to find those UniProt proteins corresponding to the given peptide that show at least some consistency in terms of their overall primary structure. If no pair of proteins associated with the given peptide showed more than 90% sequence identity, this peptide was excluded from consideration. (iv) Finally, out of two aligned best-matching proteins we retained the longer one. We obtained 10,626 acetylation sites on 3541 proteins, each of them having only one best-matching UniProt ID. The decrease in the number of acetylation sites is due to not satisfying the above criteria, not finding the sequence of the corresponding full-length protein in the UniProt database, or not finding a lysine residue in the specified sequence position of the finally obtained protein.

2.2. Sequence (1D) environments of acetylated and reference (non-acetylated) lysine residues

The positive dataset of tissue-specific LASs consisted of all lysine acetylated sites displaying non-zero intensity values in the corresponding tissue. The negative (reference or non-LASs) set was generated by extracting all lysine residues not annotated as acetylated by Lundby et al. (2012b) and relating them to those tissues in which the protein harboring the reference site also has at least one experimentally observed LAS. Then, we generated 21-mer sequences (from position −10 to position +10) surrounding each site in both positive and negative datasets and performed homology reduction on these 21-mers using CD-HIT (Li and Godzik, 2006) at the 90% identity threshold. Note that some of acetylation and reference sites occur in more than one tissue. The resulting dataset, which we call LAS1D, is composed of non-redundant 21-mer sequences corresponding to 9868 LASs and 94,362 non-LASs (Table 1). The distribution of LASs and non-LASs in different tissues is given in Fig. 1.

We used the Two Sample Logo method (Vacic et al., 2006) for differential analysis of 21-mer occurrence in different tissues, using the corresponding LASs and non-LASs as positive and negative sample inputs, respectively. For example, LASs observed in brain were compared to non-LASs in brain. Amino acids were colored using the WebLogo defaults, and *t*-test with a cut-off *p*-value of 0.05 was used to select significantly enriched residues. The Motif-X online tool (Chou and Schwartz, 2011) was used to extract motifs from the 21-mer sequences of LASs, using LAS and non-LAS as the foreground and background datasets, respectively.

2.3. Lysine acetylation sites with known 3D structure

In order to analyze the properties of spatial (3D) environments of LASs we collected a dataset of proteins with known atomic structure containing lysine residues annotated as acetylated by Lundby et al. (2012b). Using the amino acid sequences of acetylated proteins as queries we extracted the total of 1689 related 3D structures from the Protein Data Bank (Berman et al., 2000)

Table 1
Data summary of lysine acetylation sites.

Datasets	Description	Number
Initial dataset	LASs	10,626
	Proteins	3541
Structure-based dataset	LASs	2566
	Proteins	856
LAS1D (non-redundant sequence-based)	LASs (positive set)	9868
	Non-LASs (negative set)	94,362
LAS3D (non-redundant structure-based)	LASs (positive set)	2218
	Non-LASs (negative set)	8777