ARTICLE
Journal of Cellular Biochemistry 114:874–887 (2013)

Journal of Cellular Biochemistry

Influence of the Sequence Environment and Properties of Neighboring Amino Acids on Amino-Acetylation: Relevance for Structure-Function Analysis

Zeeshan Iqbal,[1] Daniel C. Hoessli,[1,2] Afshan Kaleem,[1,3] Jawaria Munir,[1] Muhammad Saleem,[4] Imran Afzal,[1] Abdul Rauf Shakoori,[5*] and Nasir-ud-Din[1,6**]

[1]*Institute of Molecular Sciences and Bioinformatics, Lahore, Pakistan*
[2]*Panjwani Institute of Molecular Medicine and Drug Research, University of Karachi, Karachi, Pakistan*
[3]*Department of Zoology, Lahore College for Women University, Lahore, Pakistan*
[4]*Department of Botany, University of the Punjab, Lahore, Pakistan*
[5]*School of Biological Sciences, University of the Punjab, Lahore, Pakistan*
[6]*HEJ Research Institute of Chemistry, University of Karachi, Karachi, Pakistan*

ABSTRACT

Proteins function is regulated by co-translational modifications and post-translational modifications (PTMs) such as phosphorylation, glycosylation, and acetylation, which induce proteins to perform multiple tasks in a specified environment. Acetylation takes place post-translationally on the ε-amino group of Lys in histone proteins, allowing regulation of gene expression. Furthermore, amino group acetylation also occurs co-translationally on Ser, Thr, Gly, Met, and Ala, possibly contributing to the stability of proteins. In this work, the influence of amino acids next to acetylated sites has been investigated by using MAPRes (Mining Association Patterns among preferred amino acid residues in the vicinity of amino acids targeted for PTMs). MAPRes was utilized to examine the sequence patterns vicinal to modified and non-modified residues, taking into account their charge and polarity. The PTMs data were further sub-divided according to their sub-cellular location (nuclear, mitochondrial, and cytoplasmic), and their association patterns were mined. The association patterns mined by MAPRes for acetylated and non-acetylated residues are consistent with the existing literature but also revealed novel patterns. These rules have been utilized to describe the acetylation and its effects on the protein structure-function relationship. J. Cell. Biochem. 114: 874–887, 2013.

KEY WORDS: ACETYLATION; CO- AND POST-TRANSLATIONAL MODIFICATIONS; SEQUENCE ANALYSIS; ASSOCIATION RULE MINING; CHARGE AND POLARITY; MULTIFUNCTIONALITY OF PROTEINS

The many in vivo functions of proteins depend on their capacity to accept different functional groups in their amino, hydroxyl, and carboxyl modification centers. The post-translationally modified proteins assume different functional configurations, and their proteolytic degradation products may assume again distinct functional configurations, as illustrated in the complement system and coagulation cascades. Multifunctional proteins cover the whole range of extra- and intra-cellular activities, required to carry out biological functions [Nasir-ud-Din et al., 2009].

Acetylation of proteins is an extensively studied modification, which occurs co- and post-translationally. This modification mainly controls gene expression, intra-cellular protein localization, and other regulatory functions. Acetylation on the N-terminal of mammalian proteins is irreversible [Polevoda and Sherman,

Abbreviations used: PTM, post-translational modification; CTM, co-translational modification; SPS, significantly preferred sites; APs, association patterns; SL, support level; CL, confidence level; NATs, N-α-acetyltransferases; HAT, histone acetyltransferase; Cp, cytoplasmic; Mc, mitochondrial; Nu, nuclear; Xp, miscellaneous.

Additional supporting information may be found in the online version of this article.

*Correspondence to: Abdul Rauf Shakoori, Distinguished National Professor and Director, School of Biological Sciences, University of the Punjab, Quaid-i-Azam Campus, Lahore 54590, Pakistan. E-mail: arshaksbs@yahoo.com, arshakoori.sbs@pu.edu.pk

**Correspondence to: Nasir-ud-Din, Institute of Molecular Sciences and Bioinformatics, 28-Nisbat Road, Lahore, Pakistan. E-mail: prof_nasir@yahoo.com, professor_nasir@yahoo.com, chairman@imsb.edu.pk

Manuscript Received: 17 August 2012; Manuscript Accepted: 15 October 2012

Accepted manuscript online in Wiley Online Library (wileyonlinelibrary.com): 23 October 2012

DOI 10.1002/jcb.24426 • © 2012 Wiley Periodicals, Inc.

2002] and takes place co-translationally. Indeed about 80% of the cytoplasmic proteins are known to be acetylated co-translationally [Kramer et al., 2009]. This mechanism was conserved throughout evolution, emphasizing its primordial role in determining protein function. N-α-acetyltransferases (NATs, a subfamily of the histone acetyltransferase [HAT] superfamily) are the enzymes responsible for *N*-terminal acetylation and are associated with the ribosome [Kramer et al., 2009].

The amino acid sequences of acetylated *N*-termini are highly diverse. For instance, the Met residue which is cleaved off is usually followed by amino acids having a short polar side chain (Gly, Ala, Ser, Cys, Thr, Pro, and Val). In human proteins, about 50% of Met–Lys, 96% of Met–Ala termini, and very few proteins with Val, Met, and Cys termini are acetylated co-translationally [Arnesen et al., 2009]. However, despite its widespread occurrence, the regulatory role of *N*-terminal acetylation is still not fully understood, though it is known to provide stability to the protein [Polevoda and Sherman, 2002], to confer protection against unfavorable proteolytic cleavages, non-enzymatic reactions [Boissel et al., 1985], and in some instances to trigger the growth hormone releasing factors [Polevoda and Sherman, 2002].

Acetylation of the ε-amino group of Lys residues is a very common post-translational modification (PTM) in histone proteins, transcription factors as well as in non-nuclear proteins [Yang and Seto, 2008]. The enzymes responsible for addition of the acetyl moiety to the ε-amino group of Lys are acetyltransferases (mainly HATs). Lys acetylation, unlike *N*-terminal acetylation, is a dynamic modification controlled by different acetylases and deacetylases [Yang, 2004]. Lys acetylation, by controlling interactions between DNA–protein and protein–protein, is instrumental in regulating transcriptional activities [Kouzarides, 2000]. Acetylation also regulates chromatin remodeling [Hake et al., 2007], cell proliferation [Kouzarides, 1999], apoptosis [Sykes et al., 2006], protein stability [Kouzarides, 2000; Yang and Seto, 2008], and nuclear localization [Yang and Seto, 2008]. Dysregulation of these dynamic processes may occur in the pathogenesis of many diseases, especially in that of cancer [Yang, 2004; Sykes et al., 2006] and diabetes [Gray and De, 2005].

The control of protein function is achieved by the combination of PTMs occurring on a given protein. On the hydroxyl function of Ser/Thr residues, phosphorylation and *O*-GlcNAc modification inversely regulate gene transcription [Comer and Hart, 2001; Gray and De, 2005] and cellular localization of proteins [Comer and Hart, 2001; Lefebvre et al., 2003]. A similar type of interplay exists between acetylation and other modifications such as methylation and phosphorylation, regulating chromatin modeling and transcriptional activities [Yang and Seto, 2008]. Acetylation and methylation in histone proteins may sometimes occur in a mutually exclusive fashion and thus either promote or inhibit gene expression [Lefebvre et al., 2003; Sarg et al., 2004]. In some instances, there is a competition between acetylation and methylation of ε-NH_2 of Lys. When Lys is mutated to Arg, no acetylation occurs in the protein in vivo [Rausa et al., 2004; Sarg et al., 2004], but only in vitro [Rausa et al., 2004]. In contrast, methylation of both Lys and Arg is well-known in vivo occurrence. In the tumor repressor p53 protein, acetylation combines with other modifications such as phosphorylation, to promote p53 association with p300, which in turn acetylates p53 in its *C*-terminal [Starheim et al., 2009]. The Lys residues in the *C*-terminal regions of histones also compete for ubiquitination, promoting nuclear export and protein degradation, while acetylation rather promotes nuclear localization and protein stability [Yang and Seto, 2008]. Thus a specific combination of different modifications may contribute to the multi-functionality of proteins.

The multifaceted properties of amino acids are responsible for the specificity and diversity of structural and functional annotation of proteins. Efforts have been made to characterize amino acids on the basis of similarities in properties such as ∞-helix, turn, β-strand, size and other physico-chemical properties [Kawashima et al., 2008]. The physico-chemical properties such as polarity and charge of amino acids should be taken into account to understand structure-function relationship. It has been observed that characteristic physico-chemical properties correspond to a specific cellular allocation (nuclear, cytoplasmic, inner and outer membrane) of bacterial proteins [Sjöström et al., 1995]. By the same token, it is instructive to consider the environment around an acetylation site in terms of polarity and charge. In the current study, we focused on polarity and charge of vicinal amino acids, in order to define their influence on acetylation sites.

Several computational studies have been performed to develop efficient and reliable prediction models for PTMs, utilizing the experimentally known modification data and data mining techniques that employed different mathematical/searching algorithms, implemented with one of several available machine learning schemes [Oyama et al., 2002; Creighton and Hanash, 2003; Ahmad et al., 2008a]. In silico, methods have significantly contributed to provide useful information about candidate modification sites in view of their experimental identification. As amino acids in the vicinity of a PTM site are critical in permitting or forbidding the incoming substituent on specific amino acid. The knowledge about amino acids surrounding modified sites and the correlation between them is critical for the development of a strong and reliable prediction model for modifications in proteins.

MAPRes (Mining Association Patterns among preferred amino acid residues in the vicinity of amino acids targeted for PTMs [Ahmad et al., 2008b]) is an efficient computational tool to develop a correlation between a modified site and its neighboring amino acids. MAPRes mines association patterns (APs) among statistically preferred amino acids in the vicinity of targeted residues for a given type of modification. The MAPRes algorithm is not only efficient for association analysis of modified sites, but also valuable for finding the confidence level (CL) and support level (SL) of such associations [Ahmad et al., 2008a]. The investigations based on physico-chemical properties of surrounding amino acids of modified residues are valuable for consensus development between structure and function. In this study, a polarity and charge-based classification was performed to investigate the influence of charge and polarity on the modifiable site. This newly classified data analyzed by utilizing MAPRes and detected significantly preferred polar, charge, and neutral amino acids around acetylated and non-acetylated residues. The applications and implications of our findings utilizing MAPRes have been verified regarding the acetylation potential of amino

groups by exploiting the existing literature and computational prediction models.

Mining APs for acetylated residues need to compile information on modified proteins (Protein ID, position, modified amino acid, and sequence). Several databases have information about co- and post-translational acetylation such as CPLA [Liu et al., 2011], SysPTM [Li et al., 2009a], HPRD [Keshava et al., 2009], PhosphoSitePlus [Hornbeck et al., 2004], and dbPTM 2.0 [Lee et al., 2006]. CPLA and PhosphoSitePlus are useful databases for the retrieval of the information only for Lys acetylation. SysPTM contains information about different types of PTM, while HPRD provides information associated only with Lys acetylation on human proteins. All the above-mentioned databases provide information about acetylated proteins but this information is not convenient for use with MAPRes. The required database should contain easily retrievable information that focuses on the acetylation site and its surroundings to successfully mine APs. The dpPTM 2.0 provide information about different PTMs (experimentally and predicted sites), which is in consensus with requirements for MAPRes. The version of dpPTM utilized has 1,173 acetylated proteins covering 458 PTM and 949 co-translational modification (CTM) sites (Table I).

In this work, in silico, studies performed to investigate the role of neighboring residues of post-translationally acetylated Lys and *N*-terminal co-translational acetylation with and without categorization of acetylated proteins, and according to sub-cellular location. The association rules and significantly preferred sites (SPS) were mined by MAPRes.

MATERIALS AND METHODS

In this study, evaluations were performed on different categories of acetylated proteins. At first, MAPRes was utilized for general consensus development between acetylated site and its neighboring amino acids by mining SPS and APs. The importance of physico-chemical properties was identified by generating a new dataset of classified amino acids regarding polarity and charge. In this classification, standard amino acids divided into five groups, according to nature of their side chain (R-group) property of the amino acids (Table II). In second step, sequence patterns were mined regarding the sub-cellular locations of acetylated proteins such as cytoplasmic (Cp), mitochondrial (Mc), nuclear (Nu), and miscellaneous[1] (Xp) proteins. The data statistics about sub-cellular proteins are in Table III. Furthermore, examination of the sub-cellular protein dataset performed according to the charge and polarity of the surrounding amino acids. Next, the primary sequence analyses of acetylated proteins were extended to the non-acetylated residues to search for possible patterns around non-acetylated residues. These investigations were performed on targeted sites with 10 straddling amino acids. The dataset of experimentally identified acetylated sites was searched from various databases and finally assembled from dbPTM [http://dbptm.mbc.nctu.edu.tw/download.php, version 2.0]. Data statistics for assembled proteins of experimentally known acetylated residues are in Table I. Ambiguities in the data were cleaned by using data cleaner utility developed in .net framework.

Validation of the APs was checked by comparing APs mined by MAPRes with the information obtained from existing computational prediction models and available published literature [Li et al., 2006; Basu et al., 2009; Li et al., 2009b].

TABLE I. Data Summary of Acetylated Proteins

	Total	PTM	CTM
Acetylated sites	1,407	458	949
After cleaning	1,406	458	948
Non-acetylated sites	136,706	26,606	110,100
N-terminal non-acetylated sites[a]	–	–	4,175
Proteins	1,173	226	947

[a]The PTM group only consists of Ac-Lys while the CTM group comprises Ac-Ala, Ac-Gly, Ac-Ser, Ac-Met, and Ac-Thr. The CTMs occur at the end of the proteins, but MAPRes collected all non-acetylated residues from the entire protein chain. However, this exercise did not provide a clear definition of significantly preferred amino acids around non-acetylated sites in CTM proteins. The optimization of the data was carried out by removing all non-acetylated sites at position greater than 10.

RESULTS

MAPRes was implemented to develop correlations among the targeted sites and their neighboring residues in acetylated proteins. As PTM and CTM control different types of protein functions, the association rules mined by MAPRes for each modification will be discussed separately.

PREFERENCE ESTIMATION AND APS FOR GENERAL AND CHARGE-SPECIFIC DATASETS

Preference estimations and association rule mining is an extensively used technique for large-scale dataset [Agarwal et al., 1993; Nikfarjam and Gonzalez, 2011]. The association rules are mined for protein sequence analysis in two steps: 1) frequency estimation for connected amino acids, 2) identification of highly preferred amino acids around targeted site. MAPRes has already been implemented on *O*-phosphorylation and *O*-glycosylation dataset for mining preferred sites and APs. Herein, MAPRes implemented to establish rules with and without considering the polarity and charge of vicinal amino acids of acetylated sites.

ANALYSIS OF ACETYLATION SITES WITHOUT SPECIFIC SUB-CELLULAR LOCATION: POST-TRANSLATIONAL ACETYLATION

The observed and expected frequencies of amino acids, at each position (−10 to +10), in the surroundings of targeted residues were tabulated (Suppl. Table I). The frequency diagram for acetylated Lys

TABLE II. Classification of Amino Acids for Charge-Specific Analysis

Chemical nature of side chain	Amino acids	Abbreviations
Non-polar, aliphatic R groups	G, A, V, L, I, P	N
Aromatic R groups	F, Y, W	A
Polar, uncharged R groups	S, T, C, M, N, Q	P
Negatively charged R groups	D, E	E
Positively charged R groups	K, R, H	O

[1]Miscellaneous protein are cell membrane, cell surface, endosome, endoplasmic reticulum membrane, cell function, focal adhesion, peripheral membrane endomembrane, golgi apparatus, lipid droplet, virion proteins and those which have unknown sub-cellular location.

TABLE III. Data for Ac-Lys According to Sub-Cellular Localization

	Cytoplasmic proteins (Cp)	Mitochondrial proteins (Mc)	Nuclear proteins (Nu)	Miscellaneous proteins (Xp)
Proteins	23	87	62	54
Modified sites	33	195	141	89
General				
No. of SPS	2	11	31	8
Total no. of APs	8	10	26	22
Identical APs	3	2	20	14
Charge specific				
No. of SPS	1	6	15	5
Total no. of APs	7	21	40	21
Identical APs	1	11	24	11

The consensus development for acetylated amino acids and their surrounding ones was achieved by analyzing the sequences of proteins belonging to the four categories of cytoplasmic (Cp), mitochondrial (Mc), nuclear (Nc), and miscellaneous (Xp) proteins. In the General dataset, we used MAPRes to determine the significantly preferred amino acids, while in the Charge-specific dataset, the polarity and charge of vicinal amino acids were taken into account.

(Ac-Lys) showed that Lys itself was the most frequent residue at various positions in its neighborhood, particularly at −4, +4, and +8 positions (Suppl. Table I). MAPRes searched and mined 29 unique APs by utilizing preference estimations of surrounded amino acids. MAPRes mined APs for Ac-Lys with CL ranging from 43.33% to 100% and 23 out of 29 patterns were mined at 100% CL (Table IV). It was observed that Lys is significantly preferred residue at 15 different positions in the vicinity of Ac-Lys (Table V). As described earlier, Lys itself is the most frequent residue around Ac-Lys and it is also the most observed residue in mined APs (Table VIa). Other residues are Ala and Gly that have a high rate of occurrence in searched patterns. Furthermore, Lys at +4 and −4 position mined at highest SL (Table VIa).

Analysis performed on the basis of polarity and charge of the amino acids indicated that non-polar amino acids occurred with highest frequency at every position in the vicinity of acetylated Lys (Suppl. Table II), but interestingly MAPRes also identified 9 out of 11 SPS for positively charged amino acids. Non-polar amino acids were found preferred only at +2 position around Ac-Lys but with highest SL (Table V). MAPRes suggested 12 APs around Ac-Lys, which have +2 position preferred for non-polar amino acids. Furthermore, MAPRes found positively charged amino acids in all APs except one. MAPRes found all these APs at 100% CL (Table VIb).

ANALYSIS OF ACETYLATION SITES WITHOUT SPECIFIC SUB-CELLULAR LOCATION: CO-TRANSLATIONAL ACETYLATION

Results for CTM sites showed that Met at −1 position is the most frequent and most preferred residue around the modified Ala, Gly, Ser, and Thr. At position +1, Asp is the most frequent for Ac-Gly and Ac-Met while Glu is most frequent at position +1 in the vicinity of Ac-Met and Ac-Thr (Suppl. Table I). In the environment of co-translationally modified residues, MAPRes identified 65 SPS and 67 APs with different CL and SL (Table IV). Furthermore, it was observed that Ala was significantly preferred at many positions in the vicinity of Ac-Ser and Ac-Ala, and Lys was significantly preferred at three different positions (+4, +6, and +10) in the environment of Ac-Ser and Ac-Gly. The APs mined by MAPRes for CTM sites (except Met) found Met at −1 position with SL up to 70%. Other residues concerning the APs are Ala (+1, +3, and +5) around Ac-Ala, Asp (+1), Glu (+2) and Pro (+2) around Ac-Met, Ala (+3 and +7) and Lys (+6) around Ac-Ser, Glu (+1) and Lys (+8) around Ac-Thr and Asp (+1) around Ac-Gly were highly preferred sites (Table VIa).

In the analyses performed on CTM sites with respect to polarity and charge of surrounded amino acids, 56 identical APs were found. Preference estimation around the CTM sites specified 23 SPS. It was observed that −1 position is significantly preferred for polar

TABLE IV. Results Generated by MAPRes for Co- and Post-Translationally Acetylated Sites

	PTM				CTM			
	General dataset		Charge-specific dataset		General dataset		Charge-specific dataset	
	Acetylated	Non-acetylated	Acetylated	Non-acetylated	Acetylated	Non-acetylated	Acetylated	Non-acetylated
No. of significantly preferred positions	26	61	11	30	65	168	23	45
No. of association patterns/rules (total)	32	52	29	63	128	160	137	379
No. of association patterns/rules (identical)	29	52	20	55	67	139	54	245
No. of association patterns/rules mined at 100% CL	23	36	20	50	44	16	39	146
Range of CL (%)	43.33–100	75.07–100	100	89.47–100	3.96–100	2.3–100	6.21–100	2.72–100

The significantly preferred positions were searched by MAPRes for all 20 amino acids around each modified sites (see details in Table VI) and their APs were mined. Rarely does MAPRes suggest a similar patterns at different SL, indicating that a specific rule is valid at several SLs.

TABLE V. Significantly Preferred Positions Mined by MAPRes

Amino acids	Acetylated sites						Non-acetylated sites					
	PTM	CTM					PTM	CTM				
	Lys	Ala	Gly	Met	Ser	Thr	Lys	Ala	Gly	Met	Ser	Thr
A	−6, −3, −2,	1, 2, 3, 4, 5, 6,	–	7,	3, 7, 8, 9, 10,	–	−8, −4, −3, −2, 1,	−5, −4, −3, −2, −1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,	−3, −2, −1, 3, 5, 6, 7, 8, 10,	1, 2, 3, 4, 6, 7, 8, 9, 10,	−2, −1, 1, 2, 3, 5, 6, 7, 9, 10,	−2, −1, 1, 4, 5, 7, 9, 10,
C	–	–	–	4, 6,	–	–	–	10,	5,	–	−1, 1, 7, 9,	–
D	–	1, 9,	1,	1,	1, 5,	–	−4, −3, 4,	–	–	2,	3,	3,
E	–	1,	3,	1,	1,	1, 10,	−10, −7, −6, −4, −3, −1, 1, 3, 4, 5, 7, 8,	–	–	2,	–	3,
F	–	–	9,	–	–	–	2,	–	–	–	1,	–
G	−7, −4, −2, −1, 3,	8,	5,	9, 10,	3,	–	–	1, 6, 8,	−3, −2, 1, 2, 3, 4, 5, 6, 7, 9,	–	7,	–
H	1,	–	–	–	–	–	–	–	–	–	–	–
I	–	–	–	–	–	–	−2, 1, 2, 6,	–	–	–	–	–
K	−9, −8, −7, −5, −4, −3, 1, 3, 4, 5, 6, 7, 8, 9, 10,	–	4, 6, 7,	–	4, 6, 10,	8,	−10, −9, −8, −7, −6, −5, −4, −3, −2, −1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,	7, 8, 10,	1, 2, 3, 4, 6, 7, 8, 10,	–	4, 9, 10,	8,
L	–	–	–	–	1,	–	−5, −2, −1, 2,	–	–	10,	−1,	–
M	–	−1,	−1,	–	−1,	−1,	–	−9, −8, −7, −6, −5, −4, −3, −2, −1,	−9, −8, −7, −6, −5, −4, −3, −2,	–	−9, −8, −7, −6, −5, −4, −3, −2, −1,	−9, −8, −7, −6, −5, −4, −3, −2,
N	–	–	–	1, 3,	–	–	−2,	–	–	–	–	–
P	–	4,	–	2,	–	–	–	1, 4, 7, 9,	–	3,	–	5, 10,
Q	–	2,	–	–	–	2,	–	–	–	3,	1,	3,
R	–	10,	–	10,	–	–	–	8,	8, 9,	–	4,	–
S	7,	1, 3,	–	5,	1, 2,	–	–	−3, −2, −1,	−3, −1,	1, 2, 3, 5,	−2, −1, 2, 6,	−2,
T	–	1, 4,	–	–	2, 9,	–	–	−1,	–	2, 3, 5,	8,	3,
V	–	7, 10,	2,	–	9,	–	−6, −5, −2, 9,	2, 6,	9,	–	–	–
W	–	–	–	–	–	–	−2, 2, 6, 7,	6, 10,	10,	–	–	–
Y	1,	–	–	–	–	–	−5,	–	–	–	–	–
Clssified amino acids			SPS for charge-specific dataset						SPS for charge-specific dataset			
N	−2,	3, 4, 7,	–	–	3, 9,	–	−2, 2,	1, 2, 3, 4, 5, 6, 8, 9,	2, 3, 4, 5,	1, 4, 9, 10,	2, 7, 9,	1, 10,
A	1,	–	9,	–	–	–	−2, 2,	–	–	–	–	–
P	–	−1, 1,	−1,	3, 5,	−1, 2,	−1,	–	−3, −2, −1,	−3, −2,	1,	−3, −2, −1,	−3, −2, 3,
E	–	1,	1, 3,	1,	1, 5,	1,	−10, −7, −6, −4, −3, −1, 1, 3, 4, 7, 8,	–	–	2,	–	3,
O	−9, −8, −5, −4, 1, 4, 8, 9, 10,	–	7,	–	6,	–	−9, −8, −7, −6, −5, −3, −1, 1, 2, 3, 5, 6, 7, 8, 9,	7, 8, 10,	1, 7, 8, 9, 10,	–	4, 10,	–

TABLE VI. Association Rules Mined for Acetylated Residues

Serial no.	(a) General dataset	Confidence level	Support level
Association rules for Lys			
1	<A,−2>	100	10
2	<A,−3>	100	10
3	<A,−6>	100	10
4	<G,−1>	100	10
5	<G,−2>	100	10
6	<G,3>	59.090908	10
7	<G,−4>	100	10
8	<G,−7>	100	10
9	<H,1>	100	10
10	<K,1>	100	10
11	<K,10>	58.653843	10
12	<K,3>	100	10
13	<K,−3>	100	10
14	<K,4>	56.71642	15
			10
15	<K,−4>	100	15
			10
16	<K,−4><K,4>	100	5
17	<K,4><K,8>	100	5
18	<K,−4><K,8>	100	5
19	<K,5>	100	10
20	<K,−5>	100	10
21	<K,−5><K,4>	100	5
22	<K,6>	43.333332	10
23	<K,7>	82.258064	10
24	<K,−7>	100	10
25	<K,8>	85	15
			10
26	<K,−8>	100	10
27	<K,9>	100	10
28	<K,−9>	100	10
29	<S,7>	100	10
Association rules for Ala			
1	<A,1>	100	15
2	<A,3>	50.943398	15
3	<A,3><A,5>	100	5
4	<A,5>	100	15
5	<A,5><G,8>	100	5
6	<A,6>	100	15
7	<M,−1>	46.518986	70
			65
			60
			55
			50
			45
			40
			35
			30
			25
			20
			15
8	<M,−1><A,1>	100	10
		100	5
9	<M,−1><A,2>	100	10
		100	5
10	<M,−1><A,3>	48.837208	10
		48.837208	5
11	<M,−1><A,4>	100	10
		100	5
12	<M,−1><A,5>	100	10
		100	5
13	<M,−1><A,6>	100	10
		100	5
14	<M,−1><D,1>	49.39759	10
		49.39759	5
15	<M,−1><D,9>	100	5
16	<M,−1><E,1>	48.27586	10
			5
17	<M,−1><G,8>	100	5
18	<M,−1><P,4>	100	5
19	<M,−1><Q,2>	73.52941	5
20	<M,−1><R,10>	100	5
21	<M,−1><S,1>	52.941177	10
			5
22	<M,−1><S,3>	100	5

(*Continued*)

TABLE VI. (*Continued*)

Serial no.	(a) General dataset	Confidence level	Support level
23	<M,−1><T,1>	100	5
24	<M,−1><T,4>	100	5
25	<M,−1><V,10>	100	5
26	<M,−1><V,7>	100	5
Association rules for Met			
1	<D,1>	39.7351	30
			25
			20
			15
2	<D,1><P,2><N,3>	100	10
3	<D,1><P,2><N,3><C,4><S,5><-<S,5><C,6><A,7>	100	5
4	<D,1><P,2><N,3><C,4><S,5><-<S,5><C,6><G,10>	100	5
5	<D,1><P,2><N,3><C,4><S,5><-<S,5><C,6><G,9>	100	5
6	<E,1>	40.99379	35
			30
			25
			20
			15
7	<P,2>	100	15
8	<P,2><N,3><C,4>	100	10
9	<S,5>	100	15
Association rules for Ser			
1	<A,3>	49.056606	15
2	<A,7>	69.333336	15
3	<K,6>	48.333332	15
4	<L,1><K,4><D,5><K,6><V,9>	100	5
5	<M,−1>	42.405064	70
			65
			60
			55
			50
			45
			40
			35
			30
			25
			20
			15
6	<M,−1><A,3>	51.162792	10
7	<M,−1><A,7>	100	10
8	<M,−1><A,9>	100	10
9	<M,−1><E,1>	40.229885	10
10	<M,−1><K,6>	83.72093	10
11	<M,−1><S,2>	100	10
12	<M,−1><T,2>	100	10
13	<S,2>	100	15
Association rules for Thr			
1	<E,1>	7.4534164	20
2	<K,8>	15.000001	25
			20
3	<K,8><E,10>	100	10
4	<M,−1>	7.120253	70
			65
			60
			55
			50
			45
			40
			35
			30
			25
			20
5	<M,−1><E,1>	11.494253	15
			10
6	<M,−1><E,10>	100	10
7	<M,−1><K,8>	100	15
			10
8	<M,−1><K,8><E,10>	100	5
9	<M,−1><Q,2>	26.470589	15
			10
Association rules for Gly			
1	<D,1>	11.258278	55
			50

(*Continued*)

TABLE VI. (*Continued*)

Serial no.	(a) General dataset	Confidence level	Support level
2	<D,1><G,5><K,6><K,7><F,9>	100	30
			25
3	<D,1><V,2><K,4><G,5><K,6><-<K,6><K,7><F,9>	100	20
4	<M,−1>	3.955696	70
			65
			60
			55
			50
5	<M,−1><D,1>	16.86747	45
			40
			35
6	<M,−1><D,1><E,3><G,5><K,6-<G,5><K,6><K,7><F,9>	100	20
7	<M,−1><D,1><E,3><K,4><G,5-<K,4><G,5><K,6><K,7><F,9>	100	15
8	<M,−1><D,1><K,4><G,5><K,6-<G,5><K,6><K,7><F,9>	100	20
9	<M,−1><D,1><V,2><E,3><K,4-<E,3><K,4><G,5><K,6><K,7><-<K,7><F,9>	100	10
			5
10	<M,−1><D,1><V,2><K,4><G,5-<K,4><G,5><K,6><K,7><F,9>	100	15

Serial no.	(b) Charge-specific dataset	Confidence level	Support level
Association rules for Lys			
1	<N,−2>	100	45
			40
			35
			30
			25
			20
2	<N,−2><O,1>	100	15
			10
3	<N,−2><O,10>	100	10
4	<N,−2><O,4>	100	10
5	<N,−2><O,4><O,8>	100	5
6	<N,−2><O,8>	100	15
			10
7	<N,−2><O,8><O,9>	100	5
8	<N,−2><O,9>	100	10
9	<O,1>	100	25
			20
10	<O,10>	100	20
11	<O,4>	100	20
12	<O,−4>	100	20
13	<O,−4><N,−2>	100	10
14	<O,−4><N,−2><O,8>	100	5
15	<O,−4><N,−2><O,9>	100	5
16	<O,8>	100	25
17	<O,8>	100	20
18	<O,−8>	100	20
19	<O,9>	100	20
20	<O,−9>	100	20
21	<O,−9><N,−2>	100	10
Association rules for Ala			
1	<N,3>	51.6129	50
			45
2	<N,3><N,4><N,7>	100	10
3	<N,4>	100	45
4	<N,7>	100	45
5	<P,−1>	46.52997	80
			75
			70
			65
			60
			55
			50
			45
6	<P,−1><E,1><N,3>	49.315067	10
7	<P,−1><N,3>	51.785713	40
			35

(*Continued*)

TABLE VI. (*Continued*)

Serial no.	(b) Charge-specific dataset	Confidence level	Support level
			30
			25
8	<P,−1><N,3><N,4>	100	15
			10
9	<P,−1><N,3><N,4><N,7>	100	5
10	<P,−1><N,3><N,7>	100	20
			15
			10
11	<P,−1><N,4>	100	40
			35
			30
			25
12	<P,−1><N,4><N,7>	100	20
			15
			10
13	<P,−1><N,7>	100	40
			35
			30
			25
14	<P,−1><P,1>	100	25
15	<P,−1><P,1><N,3>	100	10
16	<P,−1><P,1><N,3><N,4>	100	5
17	<P,−1><P,1><N,3><N,7>	100	5
18	<P,−1><P,1><N,4>	100	10
19	<P,−1><P,1><N,7>	100	10
Association rules for Gly			
1	<E,1>	5.900621	60
			55
2	<E,1><E,3><0,7>	100	25
3	<E,1><0,7><A,9>	100	30
			25
4	<P,−1>	3.9432175	80
			75
			70
			65
			60
			55
5	<P,−1><E,1>	8.988764	50
			45
			40
			35
6	<P,−1><E,1><A,9>	100	25
7	<P,−1><E,1><E,3>	100	25
8	<P,−1><E,1><E,3><0,7><A,9>	100	20
			15
			10
			5
9	<P,−1><E,1><0,7>	100	25
Association rules for Met			
1	<E,1>	39.130436	70
			65
			60
			55
			50
			45
			40
			35
			30
2	<E,1><P,3>	100	25
			20
3	<E,1><P,3><P,5>	100	15
			10
			5
4	<E,1><P,5>	100	25
			20
5	<P,3>	100	35
			30
6	<P,5>	100	35
Association rules for Ser			
1	<N,3>	48.387096	45
2	<N,3><N,9>	100	25
			20
3	<N,9>	100	50
			45
4	<P,−1>	42.42902	80
			75

(*Continued*)

TABLE VI. (*Continued*)

Serial no.	(b) Charge-specific dataset	Confidence level	Support level
			70
			65
			60
			55
			50
			45
5	<P,−1><E,1><N,3>	50.68493	10
6	<P,−1><E,1><N,3><N,9>	100	5
7	<P,−1><N,3>	48.214287	40
			35
			30
			25
			20
8	<P,−1><N,3><N,9>	100	15
			10
9	<P,−1><N,9>	100	40
			35
			30
			25
			20
			25
			20
10	<P,−1><P,2><E,5><N,9>	100	5
11	<P,−1><P,2><N,3>	100	10
12	<P,−1><P,2><N,3><N,9>	100	5
13	<P,−1><P,2><N,9>	100	15
			10
14	<P,2><E,5><O,6><N,9>	100	5
15	<P,2><N,3><N,9>	100	10
16	<P,2><N,9>	100	20
Association rules for Thr			
1	<E,1>	6.21118	35
			30
2	<P,−1>	7.0977917	75
			70
			65
			60
			55
			50
			45
			40
			35
			30
3	<P,−1><E,1>	8.988764	25
			20
			15
			10
			5

uncharged R group (P) amino acids in the vicinity of co-translationally modified sites except Met. Negatively charged R group (E) amino acids were preferred at +1 position, in the vicinity of Ac-Gly, Ac-Met, and Ac-Thr. Similar to PTM sites, non-polar amino acids have high frequency, almost at every position, but MAPRes found only 5 SPS for non-polar amino acids around CTM (Table V and Suppl. Table II).

ANALYSIS OF ACETYLATION SITE WITH SPECIFIC SUB-CELLULAR LOCATIONS: POST-TRANSLATIONAL ACETYLATION

The results concerning to the neighboring environment of Ac-Lys in different sub-cellular locations indicate that Lys at −1 position and His at +1 position were most frequent and preferred residues in Cp region (Table VII and Suppl. Table V). In the Mc proteins, Tyr and His were preferred residues at +1 position and Glu at −2 and −1 positions were found significantly preferred. In the Nu region, MAPRes found only three residues (Ala, Gly, and Lys), which were preferred at many positions (Table VII). Preference estimation made

TABLE VII. SPS for Ac-Lys in Different Sub-Cellular Localization

	Cp	Mc	Nu	Xp
Amino acids		General dataset		
A	–	–	−6, −3, −2, 2, 5,	–
C	–	2,	–	–
D	–	−3,	–	–
E	–	−2, −1,	–	–
G	–	–	−10, −7, −6, −5, −4, −3, −2, −1, 1, 3, 6,	–
H	1,	1,	–	1,
K	−1,	−8, 8, 10,	−9, −8, −7, −5, −4, −3, 1, 2, 3, 4, 5, 6, 8, 9, 10,	−8, −7, −4, −3, 7, 8, 9,
N	–	−10,	–	–
Q	–	−3,	–	–
Y	–	1,	–	–
Classified amino acids		Charge-specific dataset		
N	–	–	−3, −2,	–
A	–	1,	–	1,
P	–	–	–	–
E	–	−4, −2, −1,	–	–
O	1,	−8, 1,	−9, −7, −5, −4, −3, 1, 3, 4, 5, 6, 8, 9, 10,	1, 7, 8, 9,

on Xp proteins suggested Lys and His as preferred residues (Table VII). The association rules for Nu and Xp proteins indicates an approximately similar trend as that found in the general dataset of PTMs, while APs mined by MAPRes in the vicinity of Ac-Lys in Cp proteins showed that His and Lys are highly preferred at +1 and −1 positions, respectively. In the surroundings of Ac-Lys in Mc proteins Glu, Tyr, Asp, Glu, and Lys were found as highly preferred residues at different positions from −10 to +10 positions (Table VIIIa).

MAPRes mined 27 SPS and 47 identical APs for Ac-Lys at different sub-cellular locations regarding polarity and charge of surrounded amino acids (Table III). As in Table VII, positively charged amino acids were found preferred at various positions in proteins with different sub-cellular location. The aromatic R group has +1 position preferred in Mc and Xp proteins, negatively charged amino acids preferred only in Mc proteins while non-polar amino acids preferred in Nu proteins (Table VII). Frequency diagrams show that non-polar amino acids have highest frequency at many positions in surrounding of Ac-Lys proteins of different sub-cellular locations (Suppl. Table VI).

ANALYSIS OF ACETYLATION SITE WITH SPECIFIC SUB-CELLULAR LOCATIONS: CO-TRANSLATIONAL ACETYLATION

Distribution and investigation of the proteins according to sub-cellular location by the utilization of MAPRes were also performed in CTM proteins. In contrast to the original results, this exercise could not draw an extended picture of SPS and APs in different regions.

ANALYSIS OF NON-ACETYLATED LYS FOR POST-TRANSLATIONALLY ACETYLATED PROTEINS

MAPRes extracted 61 SPS and 52 APs for non-acetylated Lys in PTM proteins. Due to the diversity in non-acetylated Lys data, all APs were mined only at 5% SL. Frequency diagrams specified that Lys and Leu were most frequent residues at many positions in

TABLE VIII. Association Rules Mined According to Sub-Cellular Localizations

Serial no.	(a) General dataset	Confidence level	Support level
Association rules for Cp			
1	<H,1>	15.55556	20
			15
			10
2	<K,−1>	100	25
			20
			15
			10
3	<K,−1><H,1>	100	5
Association rules for Mc			
1	<E,−2><Y,1>	100	5
2	<D,−3>	100	10
3	<E,−1>	100	10
4	<E,−2>	100	10
5	<H,1>	62.22223	10
6	<K,10>	46	10
7	<K,8>	27.16049	10
8	<K,−8>	37.70492	10
9	<Y,1>	100	15
	<Y,1>		10
Association rules for Nu			
1	<G,−1>	100	20
2	<G,1><K,4>	100	10
3	<G,−2>	100	20
4	<G,−5><G,−1>	100	10
5	<G,−5><K,−4>	100	10
6	<K,−3><G,−2>	100	10
7	<K,4>	100	30
			25
			20
8	<K,−4>	76.66666	30
			25
			20
9	<K,−4><K,4>	100	15
			10
10	<K,4><K,8>	100	10
11	<K,−4><K,8>	75	10
12	<K,−4><K,9>	88.2353	10
13	<K,−5>	100	20
14	<K,−5><K,4>	100	10
15	<K,8>	49.38272	25
			20
16	<K,−8><K,−4>	81.81818	10
17	<K,−9>	100	20
18	<K,−9><K,−4>	100	10
19	<K,−9><K,−4><A,−3><G,6><K,10>	100	5
20	<K,−9><K,−5>	100	10
Association rules for Xp			
1	<H,1>	22.22222	10
2	<K,−3>	35	15
			10
3	<K,−3><K,9>	35.71429	5
4	<K,−4>	23.33333	15
			10
5	<K,−4><K,7>	100	5
6	<K,−4><K,8>	25	5
7	<K,7>	100	15
			10
8	<K,−7>	35.89744	15
			10
9	<K,7><K,8>	100	5
10	<K,−7><K,8>	35.71429	5
11	<K,8>	23.45679	20
			15
			10
12	<K,−8>	22.95082	15
			10
13	<K,−8><K,8>	27.77778	5
14	<K,9>	33.33334	15
			10

(*Continued*)

TABLE VIII. (*Continued*)

Serial no.	(b) APs for charge-specific dataset	Confidence level	Support level
Association rules for Cp			
1	<0,1>	10.31746	35
			30
			25
			20
			15
			10
			5
Association rules for Mc			
1	<A,1>	72.41379	20
			15
			10
2	<E,−1>	100	15
			10
3	<E,−2>	100	20
			15
			10
4	<E,−2><A,1>	100	5
5	<E,−4>	100	15
			10
6	<E,−4><0,1>	100	5
7	<0,1>	37.30159	20
			15
			10
8	<0,−8>	100	20
			15
			10
9	<0,−8><E,−2>	100	5
10	<0,−8><E,−4>	100	5
11	<0,−8><0,1>	100	5
Association rules for Nu			
1	<N,−2>	100	55
			50
			45
			40
			35
			30
2	<N,−2><0,1>	100	20
3	<N,−2><0,4>	100	20
4	<N,−2><0,8>	100	25
			20
5	<N,−3>	100	50
			45
			40
			35
			30
6	<N,−3><N,−2>	100	25
			20
7	<N,−3><N,−2><0,3>	100	15
8	<N,−3><N,−2><0,3><0,4>	100	10
9	<N,−3><N,−2><0,4>	100	15
10	<N,−3><0,4>	100	25
			20
11	<N,−3><0,8>	100	20
12	<0,1>	34.12698	30
13	<0,4>	100	35
			30
14	<0,−4>	100	35
			30
15	<0,−4><N,−2>	100	20
16	<0,−4><N,−3>	100	20
17	<0,−4><0,8>	100	20
18	<0,8>	69.51219	40
			35
			30
19	<0,9>	64.70589	30
20	<0,−9><N,−2>	100	20
21	<0,−9><N,−3><N,−2>	100	15
22	<0,−9><0,−4><N,−3><N,−2><0,8><0,9>	100	5
23	<0,−9><0,−4><N,−3><N,−2><0,9><0,10>	100	5
24	<0,−9><0,−5><N,−3><N,−2><0,3><0,4>	100	5

(*Continued*)

TABLE VIII. (*Continued*)

Serial no.	(b) APs for charge-specific dataset	Confidence level	Support level
Association rules for Xp			
1	<A,1>	27.58621	15
2	<A,1><0,8>	100	5
3	<0,1>	18.25397	25
			20
			15
4	<0,1><0,8>	36.84211	5
5	<0,1><0,9>	43.75	5
6	<0,7>	100	25
			20
			15
7	<0,7><0,8>	100	10
			5
8	<0,7><0,9>	100	5
9	<0,8>	30.48781	25
			20
			15
10	<0,8><0,9>	30	10
			5
11	<0,9>	35.29412	25
			20
			15

surrounding of non-acetylated Lys (Suppl. Table III). Preference estimation proposes Lys, Ala, Asp, Glu, Phe, Ile, Leu, Asp, Val, Trp, and Tyr at several locations (Table V). For polar, charged and neutral amino acids, MAPRes mined 30 SPS in total, 15 for positively, 11 for negatively, and 2 for each non-polar and aromatic R group amino acids. Frequency plots indicate that positively charged amino acids have the highest frequency near non-modified Lys (Suppl. Table IV). The APs found by MAPRes for non-acetylated Lys are in Suppl. Table VII.

ANALYSIS OF NON-ACETYLATED RESIDUES (ALA, GLY, SER, MET, AND THR) IN CO-TRANSLATIONALLY ACETYLATED PROTEINS

The frequency plot for CTM proteins shows that −1 and +1 positions are most frequent and preferred for Ala around co-translationally modified sites. Ala, Lys, and Met were shown preference at many positions around non-modified CTM sites (Table V). The rules suggested for non-acetylated Ala, Gly, Ser, Met, and Thr (only at *N*-terminal) are in Suppl. Table VIII.

VALIDATION OF ASSOCIATION RULES

Validity of the rules mined by MAPRes for acetylated and non-acetylated residues was checked by exploiting acetylation prediction models. LysAcet and PAIL predicted 520 and 713 sites, respectively, in 50 proteins and remaining Lys considered as non-predicted sites (Table IX). Peptides of 21 amino acids long with 10 amino acids on each side (−10 to −1 downstream and 1–10 upstream) were constructed for predicted and non-predicted sites. The APs mined by MAPRes searched in the above-described peptides and counted only those containing one or more association rules. The consistency percentage for predicted sites of LysAcet with MAPRes rules were 85% for modified sites and 92% for non-modified sites (Table IX). MAPRes rules were also searched in vicinity of predicted and non-predicted sites of PAIL and PredMod. As in Table IX, comparison with PAIL and PredMod results develop a high percentage of consistency with the results mined by MAPRes. Furthermore, such prediction models that can predict modification site on the bases of polarity and charge of the amino acids are not available. So it is very difficult to compare MAPRes rules (for charged based analysis) with existing computational techniques. But for approximate conclusion, the amino acids of the predicted dataset classified according to the list as in Table II and then searched all mined APs that were established on the bases of polarity and charge of vicinal amino acids. The percentage of consistency also shows a high level of conformity (Table IX).

DISCUSSION

N- and *O*-acetylations are regulated by many genes. A variety of genes is involved in *N*-acetylation and regulates it at co- and post-translational levels [Shao et al., 2007]. The diversity of *N*-acetylation at co- and post-translational level induces functional varieties that are well defined and further functional specificity is characterized by terminal and/or internal acetylation sites [Soppa, 2010]. The genes involved are known and well studied [Shandilya et al., 2009]. Functional regulation of proteins is controlled by the combinatorial relationship between the different covalent modifications on specific amino acids in the polypeptide chain. It is very time consuming and labor intensive to identify the location of modified residues using wet lab techniques without any prior knowledge.

TABLE IX. Comparison of the Patterns Mined by MAPRes with Existing Prediction Models

	LysAcet		PAIL		PredMod	
	General	Charge specific	General	Charge specific	General	Charge specific
No. of sites						
Predicted	520		713		77	
Non-predicted	508		332		203	
No. of peptides in which rules were found						
Predicted	444	450	609	589	69	67
Non-predicted	468	324	320	203	174	113
Percentage of conformity with patterns mined by MAPRes						
Predicted	85	87	85	83	90	87
Non-predicted	92	64	96	63	86	56

However, in silico investigations, to map modification sites and their environment, can considerably reduce the cost and time of wet lab efforts. Computational methods for analyses and prediction of post-translationally modified proteins and their surrounding amino acids facilitate the design of directed protocols for experimental verifications. MAPRes is a tool to explore SPS and APs around acetylated sites in primary sequence or domain of proteins and help to focus on the sites most likely to become acetylated. MAPRes can also take into account the polarity and charge of the amino acids to illustrate the role of different properties that allow or prohibit acetylation (Fig. 1). This proteomic survey of acetylated proteins is helpful to elucidate how the protein's functional properties may be governed by the 3D locations of PTMs.

In this study, APs for acetylation of the ε-amino function of Lys (PTM) and *N*-terminal acetylation (CTM) have been mined by MAPRes at different SL and CL for experimentally known acetylated proteins. MAPRes has already been utilized to mine the APs for *O*-phosphorylation and *O*-glycosylation [Ahmad et al., 2008ab] and currently used to mine the APs for different *N*-acetylated residues. It can be further utilized on *O*-acetylation, but the current investigation is only concerned with evaluating the influence of structural environment on *N*-acetylated residues. The varied nature of *N*-acetylated proteins makes it difficult to determine the functional significance by using wet lab approaches. Some proteins need this modification for their activity and stability [Kouzarides, 2000; Yang and Seto, 2008], whereas others do not require this modification to be functional. When compared with post-translational acetylation of internal Lys residues, which dictate the function of proteins, the role of *N*-terminal acetylation is also important as a determinant of function [Liu et al., 2011]. Additionally, *N*-terminal acetylation has also been shown to occur post-translationally on melanocyte stimulating hormone [O'Donohye et al., 1982]. This complexity of *N*-terminal acetylation warrants the use of in silico techniques to simplify the wet lab investigations. The computational techniques will facilitate analyses of *N*-acetylated proteins as more remains to be learned about functional regulation of such proteins.

The current study deals with *N*-terminal and ε-Lys acetylation. The results suggested by MAPRes for both types of modification are consistent with the existing literature. It is well understood that *N*-acetylation occurs at smaller, uncharged vicinal amino acid of terminal Met after cleavage of this initiator Met. But interestingly, if Met is followed by a bulky or charged amino acid, cleavage is prevented, and acetylation is decreased [Boissel et al., 1988]. The acetylation, however, can occur on Met itself [Liu et al., 2011]. It was shown by MAPRes that Asp/Glu are preferred residues on +1 position around Ac-Met. The mined APs for Ac-Met point out that Asp and Glu have 30% and 35% SL, respectively, and this SL increases up to 70% when analysis performed on the basis of

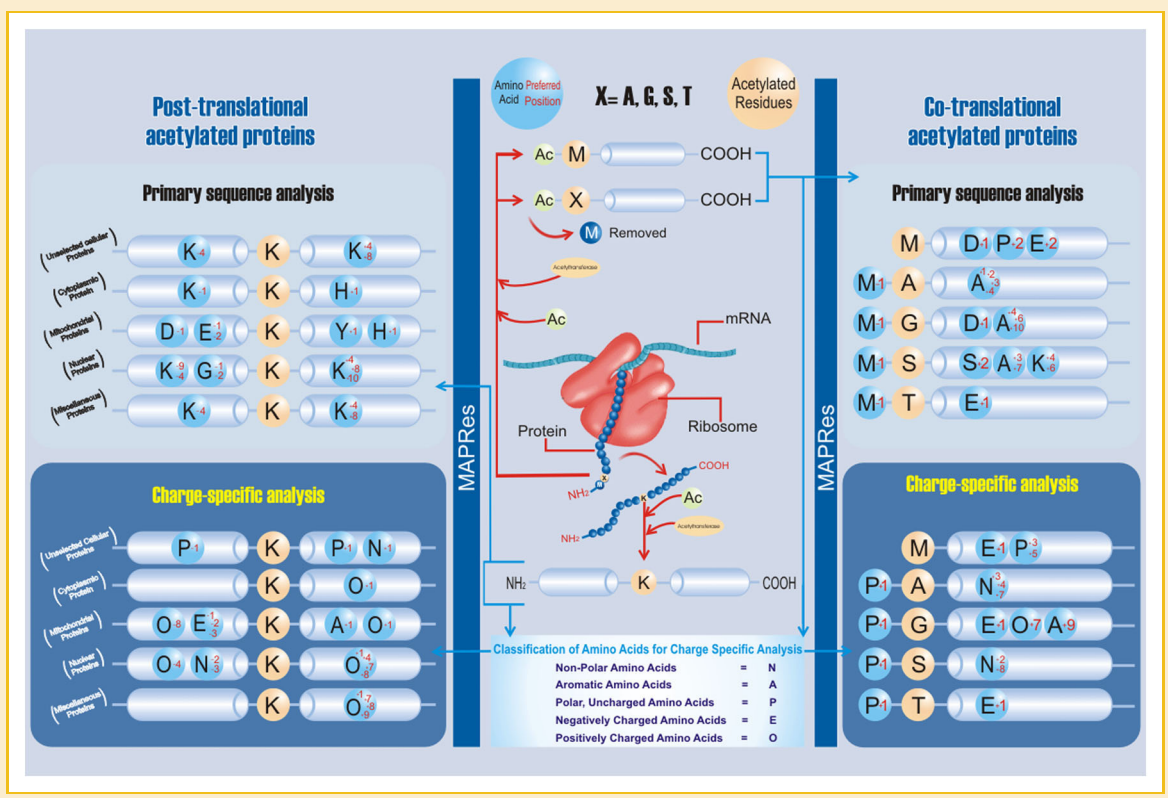

Fig. 1. Goals and achievements of the study. Central panel: General scheme of protein synthesis with emphasis on co-translational and post-translational acetylation. Left panel: Top: Amino acid requirements around acetylated lysines (K). Bottom: Charged residue requirements, for unselected, post-translational acetylated cellular proteins, cytoplasmic proteins, mitochondrial proteins and miscellaneous proteins. Right panel: Amino acid requirements following N-terminal acetylated residues (top) and charged residues (bottom) for the same protein categories as above. The MAPRes algorithm was utilized to generate the results summarized in this figure.

polarity and charge (Table VIa and VIb). Notably, these preferred negatively charged amino acids (Glu and Asp) are likely increase Met acetylation [Utsumi et al., 2001]. MAPRes also suggested that Asn on +1 and Ser on +5 position affect the modification site, as was also stated by Polevoda and Sherman [Polevoda and Sherman, 2000; Utsumi et al., 2001]. Another study with *Saccharomyces cervisiae* has shown that NAT B requires acidic amino acids next to Met [Polevoda and Sherman, 2000]. The APs proposed for Ac-Ser are Ala (+3, +7), Lys (+6), and Ser (+2), having highest SL after Met (−1). Moreover, Ser (+2) was also found with 100% CL, which indicates its importance around targeted Ser (Table VIa). It is worth noticing that Leu at +1 position has highest frequency, but its correlation was not developed with Met at −1 position, therefore no rule for Leu (+1) with Met (−1), as derived with MAPRes. A similar trend has been described by Polevoda and Sherman [2003] but only for mammalian proteins that have Ac-Met.

Furthermore, MAPRes concluded that around Ac-Thr and Ac-Ala, Met at −1 position and Asp/Glu at +1 are significantly preferred residues. This observation agrees with the work in which frequencies around N-acetylated residues were calculated from position +1 to +5, by Helbig et al. [2010]. The sequence analyses of co-translationally acetylated residues (Ala, Gly, Met, Ser, and Thr) suggested some unique trends for each modified residue. In Particular, the penultimate residue plays very important role on modification site [Utsumi et al., 2001; Lee et al., 2010]. Most interestingly, the negatively charged amino acids were significantly found at +1 position adjacent to co-translationally acetylated residues, and consensus of these preferences with modified residues was developed with high SL around Ac-Met, Ac-Gly and Ac-Thr. In the surroundings of other two modified sites (Ac-Ser and Ac-Ala), the APs were same as above but with relatively low SL (Table VIb). This indeed underscores the importance of polarity and charge of amino acids in the vicinity of acetylated residues.

Post-translational acetylation on Lys is known to influence regulatory functions of cellular proteins [Choudhary et al., 2009; Karve and Cheema, 2011]. Initially, Lys acetylation was only known for histone proteins, but later investigations revealed the fact that Lys acetylation occurs in non-histone proteins in various other cellular compartments [Sadoul et al., 2011]. The APs mined by MAPRes for post-translationally acetylated proteins show that Ac-Lys requires other Lys residues at upstream and downstream of the modified lysine (Table VIa). The simple amino acids (Ala and Gly) are also required around Ac-Lys to make it a potential acetylation site. The specific charge distribution caused by neighboring amino acids of an acetylated residue is important for the binding competence with the acetyltransferase. The relationship between the charge patch of the substrate and enzyme's active site has been discussed by Ren and Gorovsky [2001], who suggested that Ac-Lys required non-polar amino acids in its immediate vicinity (−3 to +3 position), while negatively charged amino acids were preferred next to these positions (Table VIb). In addition, this study suggested that Ser (+7) and His (+1) were often required around Ac-Lys.

Post-translational acetylation at different cellular locations imparts specific functional activities [Kim et al., 2006; Choudhary et al., 2009; Hirschey et al., 2009; Guan and Xiong, 2011; Karve and Cheema, 2011; Sadoul et al., 2011]. Acetylation of Nu proteins mainly regulates gene expression, nuclear transport, and actin nucleation [Choudhary et al., 2009; Karve and Cheema, 2011]. The regulation of longitivity and metabolic enzymes is controlled by acetylation of Mc proteins [Kim et al., 2006]. Various metabolic enzyme-associated dysfunctions are also linked to acetylation or deacetylation of Mc proteins [Kim et al., 2006; Hirschey et al., 2009; Guan and Xiong, 2011]. Acetylation of Cp proteins is found to regulate various coordinating events includes cytoskeleton dynamics, vesicle fusion, stress response, and intracellular trafficking [Sadoul et al., 2011].

To relate the sequence patterns for protein acetylation at different sub-cellular localizations, the rules for Nu, Mc, Cp, and Xp proteins were mined. The rules suggested by MAPRes, irrespective of the sub-cellular specificity of the acetylated proteins, have some similarities. Neighboring residues such as Glu (−1 and −2), Asp (−3), and Tyr (+1) in Mc proteins have significance only in the sub-cellular dataset but not in general dataset. A few similar patterns searched in both types of datasets such as Lys was found in all cellular localizations (except in cytoplasm), before and after −3 and +3 positions, respectively (Table VIIIa). Although, Lys had preference at +1 and +2 positions in Nu proteins (Table VII), consensus for these residues with Ac-Lys was not developed hence no rule is mined by MAPRes. It has been established in different studies that Lys is significantly preferred residue at +4 and −4 position in Nu proteins [Choudhary et al., 2009], which is consistent with our findings. Furthermore, Choudhary et al. [2009] described that Tyr and His are favored next to Ac-Lys in Mc proteins, while according to Kim et al. [2006] found His (+1) and Lys (−1) to have significance for Ac-Lys in Cp proteins, which are also consistent with our conclusions.

The association rule mining for acetylated proteins at specific sub-cellular location on the basis of polarity and charge of amino acids shows the preference only for positively charged amino acids at +1 position in Cp proteins (Table VII). Nu proteins are known for their high content of basic amino acids. This charge-specific analysis suggested that the Ac-Lys in Nu proteins had non-polar amino acids in their vicinity (−2 and −3 positions) at very high SL (Table VIIIb). Previous studies agreed that acetylated human Nu proteins (histone and non-histone) have preference for small and positively charged vicinal residues around Ac-Lys [Kim et al., 2006; Basu et al., 2009]. MAPRes identified the +1 position as significant for aromatic R group amino acids and the −1, −2, and −4 positions is usually occupied by negatively charged amino acids (Tables VII and VIIIb). In addition to agreeing with earlier studies [Kim et al., 2006; Basu et al., 2009; Choudhary et al., 2009], MAPRes identified some novel patterns.

It was realized that when inspecting APs mined for non-modified sites, the length of APs extends only up to one residue and that all the APs mined were at a maximum of 5% SL. Diversity in data, is of course one of the major cause of the above findings, which can be explained by looking at the frequency logos of non-acetylated Lys residues (Table IV and Suppl. Table III). Obviously, APs which support the modified sites should not support the non-modified sites. However, MAPRes extracted a few similar rules for acetylated and non-acetylated sites, and this is most probably due to the non-

optimized dataset for non-acetylated residues. The conformity of APs mined by MAPRes with existing methods provides a strong support for using MAPRes in this context (Table IX).

Several prediction tools have been developed based on the assumption that the nature of the penultimate amino acid determines, whether *N*-terminal acetylation should occur or not. In this work, APs of *N*-terminal acetylation have been mined and showed that the penultimate amino acid, together with vicinal amino acids, determine the acetylation of the α-amino group in the *N*-terminal amino acid. Post-translational acetylation on Lys can be predicted by using prediction models such as PAIL, LysAcet, and PredMod [Li et al., 2006; Basu et al., 2009; Li et al., 2009b]. The results (modified and non-modified) obtained from MAPRes for general and charge-specific data were cross-validated with PAIL and LysAcet, and showed a strong correlation (Table IX). The PredMod prediction server is especially designed for histone proteins, but can also be used for non-histone proteins. The prediction results obtained from PredMod also showed high level of conformity with results generated by MAPRes. The percentage of conformity for general dataset has ranged 85% to 90% and for charge-specific dataset it becomes 83% to 87% (Table IX). Such consistency highlights the adequacy of our data mining technique and of the MAPRes algorithm.

In conclusion, sequence analysis for pattern mining using MAPRes for general, sub-cellular, and charge-specific datasets of *N*-acetylated proteins established rules which are consistent with existing knowledge. The patterns mined by MAPRes are useful for establishing correlations between acetylated sites and surrounding amino acids but not for classification and prediction of modified sites. This study focuses on consensus development related to *N*-acetylated residues with vicinal amino acids in selected proteins, sub-cellular protein datasets, non-acetylated residues, and the charge of residues around the modified sites. The results of this study will be helpful for experimentalists of various biological sciences.

ACKNOWLEDGMENTS

Nasir-ud-Din acknowledges financial support from Pakistan Academy of Sciences and EMRO-COMSTECH/WHO for this work.

REFERENCES

Agarwal R, Imielinski T, Swami A. 1993. Mining association rules between sets of items in large databases. In: Proc. of the ACM SIGMOD International Conference on Management of Data, 207–216.

Ahmad I, Hoessli DC, Qazi WM, Khurshid A, Mehmood A, Walker-Nasir E, Ahmad M, Shakoori AR, Nasir-ud-Din . 2008. MAPRes: an efficient method to analyze protein sequence around post-translational modification sites. J Cell Biochem 104:1220–1231.

Ahmad I, Qazi WM, Khurshid A, Ahmad M, Hoessli DC, Khawaja I, Choudhary MI, Shakoori AR, Nasir-ud-Din . 2008. MAPRes: mining association patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications. Proteomics 8:1954–1958.

Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, Varhaug JE, Vandekerckhove J, Lillehaug JR, Sherman F, Gevaert K. 2009. Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. Proc Natl Acad Sci USA 106:8157–8162.

Basu A, Rose KL, Zhang J, Beavis RC, Ueberheide B, Garcia BA, Chait B, Zhao Y, Hunt DF, Segal E, Allis CD, Hake SB. 2009. Proteome-wide prediction of acetylation substrates. Proc Natl Acad Sci USA 106:13785–13790.

Boissel JP, Kasper TJ, Shah SC, Malone JI, Bunn HF. 1985. Amino-terminal processing of proteins: hemoglobin South Florida, a variant with retention of initiator methionine and N alpha-acetylation. Proc Natl Acad Sci USA 82:8448–8452.

Boissel JP, Kasper TJ, Bunn HF. 1988. Cotranslational amino-terminal processing of cytosolic proteins cell-free expression of site-directed mutants of human hemoglobin. J Biol Chem 263:8443–8449.

Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M. 2009. Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science 325:834–840.

Comer FI, Hart GW. 2001. Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II. Biochemistry 40:7845–7852.

Creighton C, Hanash S. 2003. Mining gene expression databases for association rules. Bioinformatics 19:79–86.

Gray SG, De Meyts P. 2005. Role of histone and transcription factor acetylation in diabetes pathogenesis. Diabetes Metab Res Rev 21:416–433.

Guan KL, Xiong Y. 2011. Regulation of intermediary metabolism by protein acetylation. Trends Biochem Sci 36:108–116.

Hake SB, Xiao A, Allis CD. 2007. Linking the epigenetic 'language' of covalent histone modifications to cancer. Br J Cancer 96:31–39.

Helbig AO, Gauci S, Raijmakers R, van Breukelen B, Slijper M, Mohammed S, Heck AJ. 2010. Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome. Mol Cell Proteomics 9:928–939.

Hirschey MD, Shimazu T, Huang JY, Verdin E. 2009. Acetylation of mitochondrial proteins. Methods Enzymol 457:137–147.

Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. 2004. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 4:1551–1561.

Karve TM, Cheema AK. 2011. Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. J Amino Acids 2011:207691.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. 2009. Human protein reference database-2009 update. Nucleic Acids Res 37:D767–D772.

Kim SC, Sprung R, Chen Y, Xu Y, Ball H, Pei J, Cheng T, Kho Y, Xiao H, Xiao L, Grishin NV, White M, Yang XJ, Zhao Y. 2006. Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol Cell 23:607–618.

Kouzarides T. 1999. Histone acetylases and deacetylases in cell proliferation. Curr Opin Genet Dev 9:40–48.

Kouzarides T. 2000. Acetylation: a regulatory modification to rival phosphorylation? EMBO J 19:1176–1179.

Kramer G, Boehringer D, Ban N, Bukau B. 2009. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol 16:589–597.

Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH. 2006. dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 1:D622–D627.

Lee FJ, Lin LW, Smith JA. 2010. Identification of methionine *N* alpha-acetyltransferase from *Saccharomyces cerevisiae*. Mol Cell Proteomics 9:928–939.

Lefebvre T, Ferreira S, Dupont-Wallois L, Bussière T, Dupire MJ, Delacourte A, Michalski JC, Caillet-Boudin ML. 2003. Evidence of a balance between phosphorylation and O-GlcNAc glycosylation of Tau proteins—a role in nuclear localization. Biochim Biophys Acta 1619:167–176.

Li A, Xue Y, Jin C, Wang M, Yao X. 2006. Prediction of N^{ε}-acetylation on internal lysines implemented in Bayesian Discriminant Method. Biochem Biophys Res Commun 350:818–824.

Li H, Xing X, Ding G, Li Q, Wang C, Xie L, Zeng R, Li Y. 2009a. SysPTM: a systematic resource for proteomic research on post-translational modifications. Mol Cell Proteomics 8:1839–1849.

Li S, Li H, Li M, Shyr Y, Xie L, Li Y. 2009b. Improved prediction of lysine acetylation by support vector machines. Protein Pept Lett 16:977–983.

Liu Z, Cao J, Gao X, Zhou Y, Wen L, Yang X, Yao X, Ren J, Xue Y. 2011. CPLA 1.0: an integrated database of protein lysine acetylation. Nucleic Acids Res 39:D1029–D1034.

Nasir-ud-Din , Kaleem A, Ahmad I, Walker-Nasir E, Hoessli DC, Shakoori AR. 2009. Effect on the Ras/Raf signaling pathway of post-translational modifications of neurofibromin: in silico study of protein modification responsible for regulatory pathways. J Cell Biochem 108:816–824.

Nikfarjam A, Gonzalez GH. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. AMIA Annu Symp Proc 2011:1019–1026.

O'Donohye TL, Handelmann GE, Miller RL, Jacobowitz DM. 1982. N-acetylation regulates the behavioral activity of α-melanotropin in a multitransmitter neuron. Science 215:1125–1127.

Oyama T, Kitano K, Satou K, Ito T. 2002. Extraction of knowledge on protein-protein interaction by association rule discovery. Bioinformatics 18:705–714.

Polevoda B, Sherman F. 2000. Nalpha-terminal acetylation of eukaryotic proteins. J Biol Chem 275:36479–364782.

Polevoda B, Sherman F. 2002. The diversity of acetylated proteins. Genome Biol 3 reviews0006. 1–0006.

Polevoda B, Sherman F. 2003. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. J Mol Biol 325:595–622.

Rausa FM III, Hughes DE, Costa RH. 2004. Stability of the hepatocyte nuclear factor 6 transcription factor requires acetylation by the CREB-binding protein coactivator. J Biol Chem 279:43070–43076.

Ren Q, Gorovsky MA. 2001. Histone H2A.Z. acetylation modulates an essential charge patch. Mol Cell 7:1329–1335.

Sadoul K, Wang J, Diagouraga B, Khochbin S. 2011. The tale of protein lysine acetylation in the cytoplasm. J Biomed Biotechnol 2011:970382.

Sarg B, Helliger W, Talasz H, Koutzamani E, Lindner HH. 2004. Histone H4 hyperacetylation precludes histone H4 lysine 20 trimethylation. J Biol Chem 279:53458–53464.

Shandilya J, Swaminathan V, Gadad SS, Choudhari R, Kodaganur GS, Kundu TK. 2009. Acetylated NPM1 localizes in the nucleoplasm and regulates transcriptional activation of genes implicated in oral cancer manifestation. Mol Cell Biol 29:5115–5127.

Shao Y, Lu J, Cheng C, Cui L, Zhang G, Huang B. 2007. Reversible histone acetylation involved in transcriptional regulation of WT1 gene. Acta Biochim Biophys Sin (Shanghai) 39:931–938.

Sjöström M, Rännar S, Wieslander Å. 1995. Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. Chemometr Intell Lab Syst 29:295–305.

Soppa J. 2010. Protein acetylation in archaea, bacteria, and eukaryotes. Archaea 16:820681.

Starheim KK, Gromyko D, Evjenth R, Ryningen A, Varhaug JE, Lillehaug JR, Arnesen T. 2009. Knockdown of human N alpha-terminal acetyltransferase complex C leads to p53-dependent apoptosis and aberrant human Arl8b localization. Mol Cell Biol 29:3569–3581.

Sykes SM, Mellert HS, Holbert MA, Li K, Marmorstein R, Lane WS, McMahon SB. 2006. Acetylation of the p53 DNA-binding domain regulates apoptosis induction. Mol Cell 24:841–851.

Utsumi T, Sato M, Nakano K, Takemura D, Iwata H, Ishisaka R. 2001. Amino acid residue penultimate to the amino-terminal Gly residue strongly affects two cotranslational protein modifications, *N*-myristoylation and *N*-acetylation. J Biol Chem 276:10505–10513.

Yang XJ. 2004. The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. Nucleic Acids Res 32:959–976.

Yang XJ, Seto E. 2008. Lysine acetylation: codified crosstalk with other posttranslational modifications. Mol Cell 31:449–461.