Systematic identification of Class I HDAC substrates

Tingting Li*, Boyan Song*, Zheng Wu, Ming Lu and Wei-Guo Zhu

Submitted: 24th May 2013; Received (in revised form): 12th July 2013

Abstract
Lysine acetylation is a common post-translational modification of histone and non-histone proteins. This process has an important function in regulating transcriptional activities and other biological processes. Although several computer programs have been developed to predict protein acetylation sites, deacetylases responsible for known or predicted acetylation sites remain unknown. In this research, Class I histone deacetylases (HDACs) substrates were manually obtained, and sequence features of deacetylation sites were analyzed. We found that three members of Class I HDACs (HDAC1, HDAC2 and HDAC3) shared similar sequence features. Therefore, a method was proposed to identify the substrates of Class I HDACs. We evaluated the efficiency of the prediction based on P-value distribution analysis and leave-one-out test. To validate the result of the prediction, we overexpressed Class I HDACs in cells and detected the acetylation levels of potential substrates. In the experiment, five of the seven predicted proteins were deacetylated by Class I HDACs. These results suggested that our method could effectively predict protein deacetylation sites. The work has been integrated to the website ASEB, which was freely available at http://cmbi.bjmu.edu.cn/huac.

Keywords: Class I HDACs; deacetylation; prediction; sequence feature; specificity

INTRODUCTION
In eukaryotic cells, many biological processes are regulated by post-translational modifications such as phosphorylation, acetylation, methylation and ubiquitination. Lysine acetylation is a well-studied post-translational modification and has an important function in gene transcriptional regulation [1], protein–protein interaction (PPI) [2] and protein stability [3]. In 1964, Murray found that histones can be acetylated to regulate RNA synthesis [4]. Aside from histones, many non-histone substrates of acetylation, such as p53 [5], phosphatase and tensin homolog [6] and forkhead box O1 [7], have been identified. The number of acetylated proteins discovered has considerably increased in the past decades. In our previous study, >2000 acetylated human proteins were collected [8].

Numerous proteins are acetylated by >20 acetyltransferases that function with histone deacetylases (HDACs) to maintain the balance of protein...
acetylation levels [9]. Thus far, ~18 human HDACs have been identified. Considering sequence identity and subcellular distribution, HDACs are grouped into four classes, namely, Class I (HDAC1, HDAC2, HDAC3 and HDAC8), Class II-A (HDAC4, HDAC5, HDAC7 and HDAC9) and Class II-B (HDAC6 and HDAC10), Class III (SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6 and SIRT7) and Class IV (HDAC11) [10]. These HDACs can induce chromatin condensation and downregulate gene expression by removing acetyl groups from histones [11, 12]. Aside from histones, HDACs also have diverse non-histone substrates such as p53 [13], STAT3 [14] and YY1 [15]. Class I HDACs are localized in the nucleus [11, 12], whereas Class II HDACs can move between the nucleus and the cytoplasm [11, 12]. Our understanding about the substrates of Class II HDACs is limited except several well-known substrates such as Tubulin and Hsp90 [16, 17]. Class III HDACs are members of sirtuin family, which need nicotinamide adenine dinucleotide for enzymatic activity. They can regulate diverse biological processes, including DNA repair, cellular stress response and energy metabolism [18–22], by deacetylating NBS1, PGC1-α and FOXO3a. Studies are currently underway to reveal the functions of other HDACs and SIRTs.

Computational methods used to screen candidate sites before conducting experiments can simplify experimental procedures. Several methods to predict acetylation sites have been successfully established. Li et al. [23] developed LysAcet using support vector machine (SVM) method based on available acetylation information in 2009. Gnad et al. [24] improved the SVM method by providing more mass spectrometry identified acetylation sites to train SVMs. Xu et al. [25] developed EnsemblePail using an ensemble of SVM classifiers, which outperforms LysAcet. Compared with the aforementioned SVM-based methods, ASEB was another prediction method developed by our group, which can predict acetylation sites and the responsible KAT (lysine acetyl-transferase) families [8]. In contrast to studies on various methods to predict acetylation, studies on the prediction of HDACs involved in the balance of protein acetylation level are few. For instance, Smith et al. [26] predicted and verified some deacetylation sites of SIRT3 by peptide array and machine learning techniques. Although these methods can effectively predict deacetylation sites of certain deacetylases, the use of peptide array technique to predict deacetylation sites of every deacetylase is laborious. Increasing numbers of discovered substrates of deacetylases can help develop a prediction method based on validated data of previous experiments.

Considering the extensive studies on Class I HDACs and their important functions, this study focused on Class I HDACs. The first cloned HDAC was HDAC1 [27]. A series of Class I HDACs were subsequently identified. HDAC1 and HDAC2 are mainly localized in the nucleus, and knockout mice models have shown that the genetic deletion of HDAC1 and HDAC2 results in embryonic or perinatal lethality [28]. HDAC3 often forms a stable multiprotein complex with corepressors such as NCoR and SMART to perform its function [29]. Here, based on manually determined deacetylation sites of Class I HDACs, the characteristics of Class I HDAC substrates were initially analyzed. A method that could be used to predict relevant deacetylated proteins or sites was subsequently developed. In our previous study, large amounts of acetylation proteins and sites were obtained. With this method, the substrates of Class I HDACs were systematically determined from these known acetylation sites, and a list of potential substrates was obtained. We selected seven potential substrates from this list and evaluated their acetylation levels by immunoprecipitation assay and western blot after Class I HDACs were overexpressed in cells. Experimental results confirmed the presence of proteins; probable global transcription activator SNF2L1 (SNF2L), Histone H2A.1 (H2A.X), Tyrosine-protein kinase ABL1 (ABL1), Williams syndrome transcription factor (WSTF) and High mobility group protein HMG-I/Y (HMGYA) were deacetylated by Class I HDACs. To our knowledge, this study is the first to obtain and characterize Class I HDAC deacetylation sites systematically. The initial data presented in this study provided important bioinformatics information and a basis of experimental studies on deacetylation catalyzed by Class I HDACs. To facilitate application, we developed a webserver tool (http://cmbi.bjmu.edu.cn/huac) that may indicate possible deacetylation sites of Class I HDACs.

EXPERIMENTAL PROCEDURES

Data collection
We obtained data of proteins deacetylated by Class I HDACs by searching previous studies in PubMed. With ‘hdac* AND deacetylat*’ as key words, 835
papers were got by the end of 27 February 2012. We reviewed relevant papers and selected studies revealing identified deacetylation sites and HDAC information. Owing to technical difficulties, such as creating mutation plasmids and site-specific acetylation antibodies, only a small part of papers identified specific deacetylation sites. There are 23 papers providing specific deacetylation sites for HDAC1, HDAC2 and HDAC3 (Supplementary Table S1). We extracted and mapped deacetylated proteins by using UniProt Database for sequences and UniProt IDs. Each deacetylated site was carefully checked to ensure that the position was exactly the site described in the selected studies.

**Prediction methodology**

A method was developed to detect new sites that could be deacetylated by Class I HDACs. The concept used was similar to the ASEB program developed in our previous study [8]. Briefly, three steps were taken to obtain the final $P$-values for each query peptide sequence. Step 1, Calculate Similarity Scores. Similarity scores were calculated between the query peptide sequence and each sequence in two sets (predefined special peptide set ($S_k$) containing 36 known substrate peptides of Class I HDACs and predefined background peptide set ($S_b$) containing 10,000 randomly selected peptides) according to the Blosume 62 matrix. After normalizing all these scores, we ranked them from high to low. If the query peptide is more similar to the peptides in $S_k$, the relevant scores (defined as $r_i$) should be enriched at the top of the list. Step 2, Calculate Enrichment Score (ES). To determine the level of enrichment, a running sum statistic was calculated by walking down the score list. Take $R$ as the sum of $r_i$ for all sequences in $S_k$. The statistic increased $r_i/R$ when encountering a peptide in $S_k$ and decreased $1/10,000$ when encountering a peptide in $S_b$. The ES was defined as the maximum of the statistic. Step 3, Estimate Significance of ES. A total of 9999 peptide sets with the same size as $S_k$ were randomly selected from the background peptides. The aforementioned steps were repeated to calculate the ES for each set. Then, we ranked these 9999 ESs along with the ES($S_b$) from high to low. The final $P$-value was defined as $L/10,000$, where $L$ is the rank of ES($S_b$). A smaller $P$-value indicates a more significant chance that the query peptide is deacetylated by Class I HDACs.

In the leave-one-out test, a known peptide was selected at each time and the remaining 35 peptides were considered as the predefined special peptide set. $P$-values of the known peptides were then calculated. For comparison, 1000 peptides were randomly selected and analyzed using the 36 known peptides as the predefined special peptide set.

The web server for this method was available at http://cmbi bjmu edu cn/huac.

**Cell culture and transfection**

Human HEK 293T cell line and HepG2 cell line used in this research were purchased from the American Type Culture Collection. The cells were cultured in Dulbecco’s modified Eagle’s medium supplemented with 10% fetal bovine serum, seeded in 10 cm tissue culture plates and grown overnight until 80% confluence was reached. Plasmids (8 µg) and lipotransfectamine 2000 (20 µl; Invitrogen) were used to perform transfection in one plate. The transfected cells were then harvested at 48 h post-transfection.

**Co-immunoprecipitation and western blot**

The cells were lysed with 0.1% NP-40 buffer. The cell lysates were then incubated with antibodies overnight at 4°C. Immunocomplexes were pulled down by protein A beads, washed thrice in NP-40 buffer, and subjected to SDS–PAGE. To determine the acetylation level of proteins, pan acetylation antibody (Cell Signal Technology) was used.

**RESULTS**

Deacetylation sites were manually determined from published references. A total of 31, 19 and 22 sites were obtained for HDAC1, HDAC2 and HDAC3, respectively (Supplementary Table S1). HDAC8 can deacetylate histone proteins, including H2A, H2B, H3 and H4, but specific deacetylated sites on these proteins are unknown [30–32]. Thus, we only used the sites of HDAC1, HDAC2 and HDAC3 for further analysis. Most of the deacetylation sites were found in humans. In the following analysis, the deacetylation sites in humans, including 28, 19 and 20 sites of HDAC1, HDAC2 and HDAC3, respectively, were considered. Figure 1A shows that HDAC1, HDAC2 and HDAC3 shared many deacetylation sites. We also found that the deacetylation motifs of HDAC1, HDAC2 and HDAC3 were
similar. This result is consistent with the knowledge that such HDACs belong to the same class. Thus, the non-redundant combined set was used as the deacetylation set of Class I HDACs, which contains 36 substrate sites.

To characterize the deacetylation motif of Class I HDACs, we extracted peptide sequences containing 17 amino acids with deacetylated lysines surrounded by eight residues on both sides. From the sequence frequency analysis (Figure 1B) of WebLogo [33], we can see that Class I HDACs preferred G from –8 to +2 positions except at –4. At –1 and +1, K was also preferred. K was dominantly preferred at –3, +3 and +4 positions particularly at +4. This result indicates that Class I HDACs recognize special motifs around deacetylation sites.

Identification of substrates for Class I HDACs

Based on the peptide set of Class I HDACs, a method was developed to predict the substrates of Class I HDACs. This prediction method was established based on the sequence similarity. In brief, the deacetylation sites of Class I HDACs, and their surrounding amino acids (eight residues on each side) were considered as a deacetylated peptide set. We calculated the P-value to determine whether or not a given peptide can be deacetylated by Class I HDACs. A smaller P-value indicates a more significant probability that the given peptide could be deacetylated by Class I HDACs.

Deacetylation is the reverse process of acetylation. A non-acetylation site cannot be deacetylated. Thus, more substrate sites of Class I HDACs should be present in these known acetylation sites than in randomly selected lysine sites. In our previous study, >4000 human acetylation sites were collected [8]. If the proposed method is effective, then the score distribution of these known acetylation sites should be significantly different from that of randomly selected lysine sites. From the cumulative distributions of P-values of known lysine acetylation sites (the known deacetylation sites of Class I HDACs were removed, Supplementary Table S2) and 1000 randomly selected lysine sites (Supplementary Table S3), the known acetylation sites contained more substrates than the random sites (Figure 2) at $P = 0.0083$ (Mann–Whitney test).

The leave-one-out test was used to evaluate the performance of the proposed method that could be used to identify known deacetylated peptides of Class I HDACs. Each time, a known peptide was picked, and the others were treated as the predefined peptides set. The P-values for all known peptides...
were then calculated. For comparison, 1000 peptides found on lysine sites were randomly selected and analyzed. The randomly selected peptides may contain unreported deacetylation sites, indicating that the specificity was likely underestimated. The results of the test were listed in Table 1. The sensitivities and specificities of the method were evaluated at different cutoff levels. The specificities did not decline so much with the increased sensitivities. With the cutoffs $P/C20 \leq 1e^{-3}$ and $P/C20 \leq 1e^{-2}$, both specificities can reach as high as above 90%. With the cutoff $P/C20 \leq 1e^{-1}$, the specificity can still reach 83.70%. These results indicate that this method could identify deacetylation sites of Class I HDACs.

**Functional analysis of potential substrates of Class I HDACs**

With the proposed method, known human acetylation sites could be scored to find the potential substrates of Class I HDACs (Supplementary Table S4). Known acetylation sites with smaller $P$-values had a higher probability to undergo deacetylation induced by Class I HDACs. Thus, the functions of these known acetylation sites with lower $P$-values should be determined to reveal the functional preference of Class I HDAC substrates. The known human acetylation sites with $P < 0.05$ were compared with all of the known human acetylation sites. The results showed that the potential substrates of Class I HDACs were enriched in chromosome-related Gene Ontology (GO) functions (Figure 3). Biological process enrichment was analyzed using the online DAVID program [34, 35], and some redundant GO terms were removed. The overrepresented biological functions had Benjamini corrected $P < 1e^{-5}$.

**Validation of the potential substrates of Class I HDACs**

Considering the functions of Class I HDACs in gene transcription regulation, epigenetic modification and DNA damage repair, we selected several potential substrates from Supplementary Table S4, including SNF2L, H2A.X, ABL1, WSTF, HMGAI, heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC2) and inhibitor of growth protein 4 (ING4). In our assay, HDAC1, HDAC2 and HDAC3 were initially overexpressed in HEK 293T (Supplementary Figure S1) and HepG2 cell line. The acetylation level of the target protein was then detected using lysine-specific antibodies. The experimental results showed that SNF2L, H2A.X, ABL1, WSTF and HMGAI could be deacetylated by Class I HDACs (Figure 4). By contrast, HNRNPC2 and ING4 had no reaction with Class I HDACs (Supplementary Figure S2). One possible reason is that deacetylation of some proteins probably requires cell signal stimulation such as DNA damage or growth factors. The overexpression of Class I HDACs may be insufficient to induce deacetylation in some conditions.

Figure 4A shows that the overexpression of HDAC1 in HEK 293T cell line reduced the acetylation level of SNF2L, indicating that HDAC1 could deacetylate SNF2L as predicted. SNF2L is a member of the imitation switch family of ATP-dependent chromatin-remodeling factors, which can hydrolyze ATP to provide energy to weaken or disrupt histone-DNA contacts [36]. A 146 bp DNA is wrapped around an octamer of four histones (H2A, H2B, H3 and H4), forming nucleosome core particles. H2A.X is a variant of histone H2A [37]. Several studies have found

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Known</th>
<th>Sn</th>
<th>Random</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P/C20 \leq 1e^{-3}$</td>
<td>4</td>
<td>11.10%</td>
<td>14</td>
<td>98.60%</td>
</tr>
<tr>
<td>$P/C20 \leq 1e^{-2}$</td>
<td>10</td>
<td>2780%</td>
<td>54</td>
<td>94.60%</td>
</tr>
<tr>
<td>$P/C20 \leq 1e^{-1}$</td>
<td>16</td>
<td>44.40%</td>
<td>163</td>
<td>83.70%</td>
</tr>
</tbody>
</table>

**Figure 3:** Overrepresented biological functions of potential Class I HDAC substrates relative to known human acetylation sites. The bar corresponds to the negative log10 $P$-value (after Benjamini correction). table.
acetylation sites in H2A.X, but the deacetylase of H2A.X remains unclear. To verify our prediction, we overexpressed HDAC1, HDAC2 and HDAC3 in HEK 293T cell line and then determined the acetylation level of H2A.X by western blot assay (Figure 4B). The result indicated that HDAC1, HDAC2 and HDAC3 could deacetylate H2A.X.

ABL1 is a non-receptor tyrosine kinase that implicates in several cellular processes, such as actin remodeling, cell adhesion and DNA damage response [38]. It has been reported that ABL1 is acetylated by P300 in nuclear localization signal, which promotes ABL1 accumulation in cytoplasm [39]. Here, we found that overexpression of HDAC1, HDAC2 and HDAC3 in HEK 293T cell line can decrease the acetylation level of ABL1 (Figure 4C).

Figure 4D shows that overexpression of HDAC1, HDAC2 and HDAC3 in HEK 293T cell line can decrease the acetylation level of WSTF, especially HDAC2. Immunoprecipitation assay also detected the interaction between HDAC2 and WSTF (Supplementary Figure S3). WSTF is a part of the 1.3 megabases deletion in chromosome 7, which can be seen in nearly all patients with Williams–Beuren syndrome [40]. Recent studies have revealed that WSTF is involved in vitamin D mediated transcription and DNA replication as a component of chromatin remodeling complex [41, 42]. In addition,
WSTF is a tyrosine kinase that phosphorylates Tyr 142 of histone H2A.X, which suggests it has a role in DNA damage response [43].

HMGA1 mainly participates in gene transcription regulation. It is characterized by a DNA-binding motif named AT hook that binds to the AT rich sequence in minor grooves of DNA. In gene promoter or enhancer, HMGA1 and transcription factors form an enhanceosome that serves as a platform for recruiting RNA polymerase II and other factors [44–46]. Considering HMGA1 is highly expressed in neoplastic cells [47], we used HepG2 cell line in our assay and found overexpressed HDAC1, HDAC2 and HDAC3 can reduce HMGA1’s acetylation level (Figure 4E).

These experimental data showed that the prediction method could identify Class I HDAC substrates.

**The web server for the prediction of Class I HDAC substrates**

To facilitate the users in related fields, a web server was provided for the prediction of Class I HDAC substrates (http://cmbi.bjmu.edu.cn/huac). With the web server, each query lysine site can get a $P$-value (Figure 5A). A lower $P$-value indicates a higher probability of deacetylation by Class I HDACs. The distribution of $P$-values of possible sites can help users find the importance of potential sites. Thus, $P$-values were calculated from all of the lysine sites on human proteins and then ranked from lowest to highest. Based on $P$-value, the ranks of the lysine sites in human proteins could be determined. Such $P$-values could be used by users to evaluate predicted results (Figure 5B). To control false-positive predictions, users could consider lysine sites with $P$-values lower than the top 10%. In this case, the estimated specificity would possibly be >90%. Lysine sites with $P$-values lower than the top 10% are highlighted in the table showing the predicted results at http://cmbi.bjmu.edu.cn/huac (Figure 5A). For application, users could adjust the cutoff value according to the trade-off between discovering more putative deacetylation sites and obtaining fewer false-positive predictions.
In enzyme-catalyzed reactions, an enzyme often interacts with a specific substrate directly or indirectly via scaffold proteins. PPI information helps predict enzyme-mediated biological reactions. Research work has shown that PPI information can enhance the accuracy of protein kinase substrate prediction [48]. Recently, PPI interactions for HDACs had been expanded. Joshi et al. [49] found >200 proteins interact with HDACs (HDAC1–11). Here, we provided the shortest path ≤5 steps between Class I HDACs and the query protein in the PPI network (Figure 5C). The PPI information was downloaded from the PPI databases PINA and STRING [50, 51]. Users could then visualize the interaction between the query protein and Class I HDACs (Figure 5D) by clicking the network view link on the website. For the 592 potential substrate sites of Class I HDACs in Supplementary Table S4, 10 of them do not have shortest paths ≤5 steps from neither PINA nor STRING database; 42 of them do not have shortest paths from PINA database and 37 of them do not have shortest paths from STRING database.

**DISCUSSION**

In this study, experimentally identified deacetylated proteins and lysine sites of Class I HDACs were initially obtained. We then analyzed the sequence features surrounding the deacetylated lysines of Class I HDAC substrates and developed a method to predict whether the known acetylation sites can be deacetylated by Class I HDACs. We evaluated the performance of the proposed method by P-value distribution, leave-one-out analysis and experimental validation of predicted proteins. The results showed that this method could effectively predict deacetylation sites. We also provided the P-value distribution and the shortest path of PPI between Class I HDACs and the query protein. This method and the associated web server possibly facilitated the identification of Class I HDAC substrates.

With the proposed method, potential substrates of Class I HDACs could be determined. Functional analysis revealed that the potential substrates of Class I HDACs were enriched in chromosome-related GO functions relative to known acetylation sites (Figure 3). Class I HDACs may participate in RNA metabolic processes (Figure 3) in addition to well-studied transcriptional regulation. Our research may expand the knowledge of the function of Class I HDACs in RNA metabolic processes.

This study predicted the deacetylation sites of Class I HDACs. From the collected data and experimental results, different Class I HDACs exhibited specificities. Considering the increasing number of discovered deacetylation sites, we aimed to develop new programs for individual Class I HDACs. Given the limited available data, HDAC8 in Class I was not included in the current study. Data collection and web server update will be continuously conducted.

**SUPPLEMENTARY DATA**

Supplementary data are available online at http://bib.oxfordjournals.org/.

**Key Points**

- In this research, Class I HDACs deacetylation sites were manually collected and characterized, which shown they shared similar sequence features.
- A method was proposed to identify the substrates of Class I HDACs based on sequence features. To facilitate the users in related fields, a web server was provided online at (http://cmbi.bjmu.edu.cn/huac).
- With the proposed method, known human acetylation sites were scanned to find the potential substrates of Class I HDACs (Supplementary Table S4). In the validation experiment, five of the seven predicted proteins were deacetylated by Class I HDACs, which suggested that our method could effectively predict protein deacetylation sites.

**FUNDING**

The National Basic Research Program (Grant No. 2011CBA01104) of China, the National High-Tech Research and Development Program (Grant No. 2012AA020401) of China and the National Natural Science Foundation of China (Grant Nos. 60905014 and 31030041).

**References**

Identification of Class I HDAC substrates


