

SCIENTIFIC REPORTS



OPEN

Identifying Human SIRT1 Substrates by Integrating Heterogeneous Information from Various Sources

Zichao Zhai¹, Ming Tang², Yue Yang¹, Ming Lu¹, Wei-Guo Zhu^{2,3,4} & Tingting Li^{1,5}

Most proteins undergo different kinds of modification after translation. Protein acetylation is one of the most crucial post-translational modifications, which causes direct or indirect impact on various biological activities *in vivo*. As a member of Class III HDACs, SIRT1 was the closest one to the yeast sir2 and drew most attention, while a small number of known SIRT1 substrates caused difficulties to clarify its function. In this work, we designed a novel computational method to screen SIRT1 substrates based on manually collected data and Support Vector Machines (SVMs). Unlike other approaches, we took both primary sequence and protein functional features into consideration. Through integrating functional features, the Matthews correlation coefficient (MCC) for the prediction increased from 0.10 to 0.65. The prediction results were verified by independent dataset and biological experiments. The validation results demonstrated that our classifier could effectively identify SIRT1 substrates and filter appropriate candidates for further research. Furthermore, we provide online tool to support SIRT1 substrates prediction, which is freely available at <http://bioinfo.bjmu.edu.cn/huac/>.

Post-translational modifications (PTMs) are a general phenomenon occurred among all species and known to be essential for regulating diverse functions and dynamically coordinating signaling networks¹. There is a variety of different types of PTMs, such as acetylation, ubiquitination, methylation and phosphorylation². Among these modifications, acetylation is the process of transferring an acetyl functional group from one molecule to another, and deacetylation is the reverse reaction of acetylation. These two processes are mediated by histone acetyltransferases (HATs) and histone deacetylases (HDACs)³. Based on their homology to yeast proteins, the eighteen histone deacetylase are divided into four groups. Class I, II, IV includes eleven classical Zn⁺-dependent HDACs, while class III HDACs' function depend on NAD⁺ as a cofactor^{4,5}. In mammals, the class III HDACs contains Sirtuin family from silent mating type information regulator 2 homolog 1 (SIRT1) to SIRT7. Among them, SIRT1, homolog to the yeast Sir2 protein, is a member of the mammalian sirtuin protein (SIRT1-7) family and has diverse significant roles in various biological processes that encompass genomic stability, tumorigenesis, energy metabolism and cellular stress resistance⁶. SIRT1 can deacetylate a variety of substrates which extends biological functions. As reported, SIRT1 is shown to be able to deacetylate histones and preferentially deacetylate H3K9⁷ and H4K16⁸ *in vitro*, interacts and deacetylates H1K26 and mediates heterochromatin formation⁹. In addition to histone substrates, more and more non-histone proteins had been reported to serve as SIRT1 substrates. For example, SIRT1 regulates genomic stability by deacetylating the DNA damage repair related proteins such as Ku70¹⁰, NBS1 (Nijmegen breakage syndrome 1)¹¹, Werner syndrome protein (WRN)¹². SIRT1 modulates gluconeogenesis in the liver through deacetylation of some important factors such as CRT2 and FOXO1¹³. In

¹Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China. ²Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education); State Key Laboratory of Natural and Biomimetic Drugs; Beijing Key Laboratory of Protein Posttranslational Modifications and Cell Function; Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China. ³Peking-Tsinghua University Center for Life Science; Peking University, Beijing, 100191, China. ⁴School of Medicine; Shenzhen University, Shenzhen, 518060, China. ⁵Institute of Systems Biomedicine, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China. Zichao Zhai and Ming Tang contributed equally to this work. Correspondence and requests for materials should be addressed to W.-G.Z. (email: zhuweiguo@szu.edu.cn) or T.L. (email: litt@hsc.pku.edu.cn)

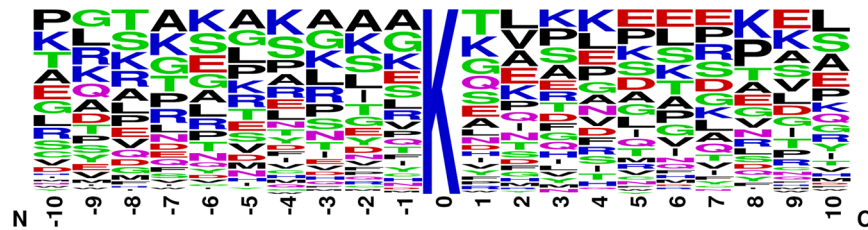


Figure 1. Frequency of amino acids in each position of SIRT1 substrates. The height of symbols within the stack indicates the relative frequency of each amino at that position.

addition, SIRT1 regulates immune system through deacetylation of NF- κ B¹⁴ and FOXP3¹⁵. All sorts of information remind us the importance and necessity of conducting research on SIRT1 and its deacetylation substrates.

As we know, enzymes affect biological process through interacting with substrates, and clarifying substrate specificity is the fundamental step to develop methods for finding SIRT1 substrates. While using oriented peptide libraries containing acetylated lysine, Blander *et al.* found that substrates of SIRT1 have no obvious motif specificity¹⁶. In 2012, based on Stable Isotope Labeling by Amino acids in Cell culture (SILAC), SIRT1 Wild-type and Knockout cells using HPLC-MS/MS analysis, were applied to proteomics for identification and quantification of lysine acetylation, which revealed SIRT1-response Lys acetylome and deacetylation related cellular pathways^{17,18}. This method was effective to screen SIRT1 substrates, but still time-consuming and expensive. Bioinformatics prediction method is a feasible approach to filter SIRT1 substrates.

In this study, we developed a SIRT1 substrate prediction method based on the manually collected substrate sites of SIRT1. Through integrating heterogeneous biological information from various sources, we constructed the SIRT1 substrate classifier based on Support Vector Machines¹⁹. Then, SIRT1 substrates were filtered from known acetylated peptides and 5,684 putative substrate sites from 1,630 proteins were obtained. Finally, we evaluated the efficiency of our prediction method through comparing with independent dataset and biological experiments.

Results

Sequence feature analysis and prediction performance with only sequence feature. Previous studies revealed two opposite opinions about whether substrates of SIRT1 have sequence specificity. One opinion is that substrates of SIRT1 share little sequence specificity¹⁶. In contrasting reports, the residues proximal to the SIRT1 deacetylated lysine sites are crucial to modifications^{20,21}. Here, we used WebLogo²² to visualize amino acids frequency in each position of SIRT1 substrates. After trimming 129 manually collected lysine centered SIRT1 substrate peptides to 21 amino acids length, 118 substrate peptides from 51 proteins were remained for subsequent analysis (Supplementary Table S1). From the amino acids distribution of these 118 peptides, it seems that there is no significant motif recognized by SIRT1, although there are some positions existing a little amino acid preference. As shown in Fig. 1, lysine (K) and alanine (A) are abundant from -7 to -1 near the N-terminus. Near the C-terminus, threonine (T) is most likely to be recognized at position 1. Except position 8, glutamic (E) is preferred from 5 to 9.

Then we further estimated the prediction performance merely based on primary sequence. To test the prediction performance with different sequence length, the compiled 118 SIRT1 substrate sequences of 51 proteins were trimmed to 11, 13, 15, 17, 19, and 21 amino acids length, which were still lysine centered and treated as positive datasets. To avoid exaggeration in prediction accuracy, we reduced the redundancy of sequence similarity between peptides. So, peptides with more than 70% sequence similarities in the positive samples left only one. Negative datasets, whose sample size and sequence length are the same as corresponding positive datasets, were randomly selected from the background set and redundancy of sequence similarity was avoided at the same time. The background set is composed of the whole lysine centered sites in human proteome after excluding proteins in positive datasets (extracted from Swiss-Prot database, version Release 20150323). Then, 4/5 positive and negative sample peptides were taken as training set. Meanwhile, the remaining 1/5 positive and negative sample peptides were taken as testing set. The above test processes were performed 100 times and negative datasets would be re-selected in each time of test. The final performance was the average of those 100 tests.

As the prediction performance shown in Table 1, we found sensitivity (Sn) varied from 41% to 48% and specificity (Sp) varied from 57% to 73%. In Fig. 2, the MCC of all sequence length performed not well (from 0.06 to 0.15) and the best performance came from 21 amino acids sequence length. We also used Random Forests and Neural Networks to construct SIRT1 substrate prediction models with sequence features, while their performance were even worse than SVM (Supplementary Table S2). The above analysis indicated the poor performance of identifying SIRT1 substrates with only sequence features.

Prediction performance with sequence and functional features. Unideal prediction performance with only sequence reminded us that protein functional features might be useful in improving the prediction performance. Thus, we collected the functional annotations from following databases: (1) Biological Process (BP), (2) Cellular Component (CC) and (3) Molecular Function (MF) annotations from the Gene Ontology (GO) database (version 20150313); (4) protein-protein interaction (PPI) information from the STRING database (version 9.1); (5) protein functional domain feature from the Pfam database (release 27.0). Then for each protein, a large number of functional annotations could be extracted as functional features from these five types of annotations. To

Seq Length	Sensitivity	Specificity
11-aa	48%	57%
13-aa	48%	57%
15-aa	44%	65%
17-aa	43%	70%
19-aa	42%	71%
21-aa	41%	73%

Table 1. Sn and Sp of prediction with different sequence length (SVM).

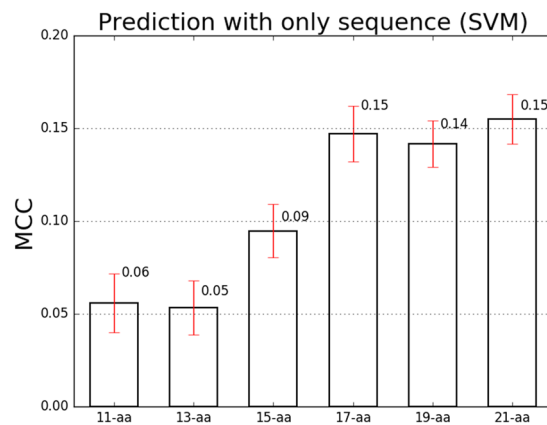


Figure 2. Prediction performance of SIRT1 substrates with different sequence lengths based on SVM. Error bar charts containing mean value and standard error were utilized here to visualise MCC results.

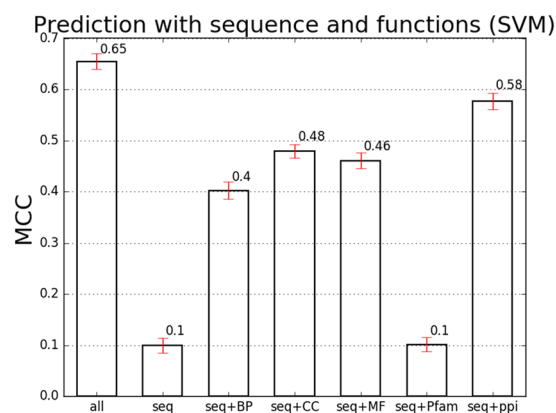


Figure 3. Prediction performance of SIRT1 substrates with sequence and functional features based on SVM. Error bar charts containing mean value and standard error were utilized here to visualise MCC results.

reduce the dimensionality of the feature space, only the over- or under-represented functional annotations would be selected. Hypergeometric tests were utilized to detect over- or under-represented functional annotations based on SIRT1 substrate proteins in the positive set, with the human proteome downloaded from the Swiss-Prot database as a background. Terms with Bonferroni corrected p -values less than $1e-2$ were considered as significant ones. If there were too many significant terms, top 100 significant terms with lowest p -values would be used. Then those selected functional annotations were used as functional features.

To estimate the prediction performance of integrating the above five types of functional features, and contributions brought by each type of functional features, seven feature combinations were obtained, as shown in Fig. 3. The test strategy was similar to that in the estimation of prediction performance with only sequence features. However, it has to be noted that two sites in one protein sharing same functional features, and if two substrate sites from one protein were separated into training set and testing set, it would cause an overestimation of prediction with functional features. So the positive and negative datasets were randomly divided into two parts (4/5 and 1/5) at protein level and then corresponding sites were extracted as the training and testing sets. In addition, the functional features were re-selected based on the 4/5 positive training samples in each time of 100 tests.

Func Groups	Sensitivity	Specificity
Seq	53%	56%
+BP	53%	84%
+CC	70%	78%
+MF	60%	84%
+Pfam	52%	57%
+STRING	63%	92%
+all	71%	92%

Table 2. Sn and Sp of prediction with sequence and functions (SVM).

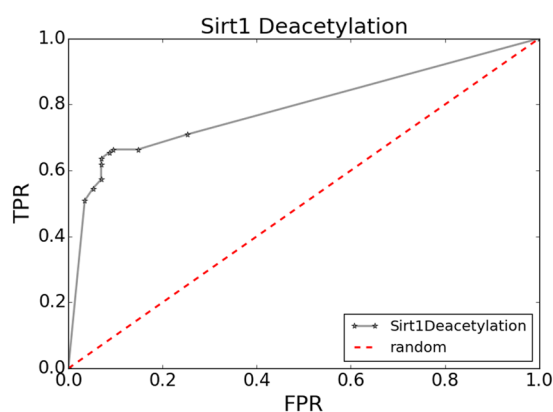


Figure 4. ROC curve of SIRT1 substrates prediction. Points on the curve were calculated with different vote number as cutoff. FPR: false positive rate, TPR: true positive rate.

As shown in Fig. 3, compared with the prediction performance with only sequences, integrating all functional features increased the MCC from 0.10 to 0.65. The Sn and Sp of integrating all features were 0.71 and 0.92 (Table 2). Furthermore, from the prediction performance of integrating sequence features with different kinds of functional features, it is obvious that most of the functional features contribute to the improvement of the prediction performance except those from the Pfam database. Compared with sequence features alone, integrating the PPI information increased MCC from 0.10 to 0.58 (Fig. 3). BP, CC, and MF annotations brought improvement of MCC to 0.40, 0.48 and 0.46 respectively (Fig. 3). These results indicated that integration of functional features can greatly improve the prediction performance.

Building SIRT1 substrates classifier and application to the known human acetylation sites. We constructed the final prediction SVM models with all known positive samples and negative samples randomly selected from the background set. Significant functional features were enriched based on all positive samples (Supplementary Table S3). The sample size of the background set is much larger than that of positive samples. We found that the selection of different negative sample sets would influence the final prediction results. When we made predictions with classifier which include only one SVM model, the overlap of two prediction results is only 74.24%. To make the prediction results repeatable, we adopted the strategy of multiple models. Each SVM model was built with the same positive sets and different randomly selected negative sets, and the final prediction results were decided by their positive prediction votes. It means that we selected a suitable positive prediction vote as cutoff, and sites with positive prediction votes no less than the cutoff would be predicted as putative candidates. We found that when the classifier included 9 SVM models and 5 of 9 models return positive predictions as the cutoff, the overlap of two predictions reached 90.07%. So, the final classifier adopted 9 models.

Furthermore, we explored different influence from different vote number to prediction performance. With different vote number as cutoffs, we obtained the points on ROC curve (Fig. 4) by the average performance of 100 rounds of evaluation tests. To balance the Sn and Sp, those candidates with no less than 5 of 9 models returning positive predictions were predicted as putative substrates of SIRT1. In addition, users can control the false positive rate by selecting more stringency cutoffs. For example, when we set the cutoff as 9, the Sn and Sp reached 51% and 97% respectively (Supplementary Table S4). Meanwhile, a web server (<http://bioinfo.bjmu.edu.cn/huac/>) was provided for the prediction of SIRT1 substrates to facilitate the users in related fields.

With the SIRT1 substrate classifier, we firstly screened putative SIRT1 substrates from human acetylation sites downloaded from the PhosphoSitePlus database. After excluding those acetylation sites of the known substrate proteins of SIRT1, 17,572 acetylation sites from 6,338 human proteins were obtained (Supplementary Table S5). With vote number 5 as the threshold, 5,412 sites from 1,590 proteins were predicted as the substrates of SIRT1 (Supplementary Table S6), whose proportion of sites is 30.80%. If our method can discriminate the substrates of SIRT1 from others, then we would expect the proportion of predicted substrates from the acetylation sites are

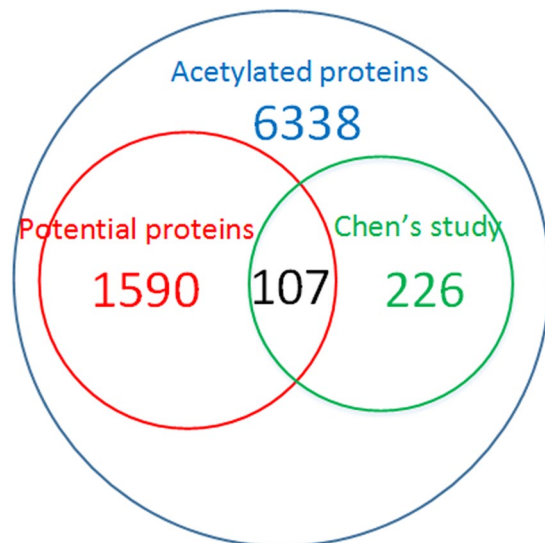


Figure 5. Overlap between our prediction results and Chen's study.

higher than that from the random lysine sites. To test this hypothesis, we randomly selected 19,949 lysine sites from 10,050 proteins (Supplementary Table S7). As for the random set, 3,112 sites from 1,644 proteins were predicted to be the substrates of SIRT1 (Supplementary Table S8), whose proportion of sites is 15.60%. The proportion of predicted substrates generated by the acetylation sites was significantly higher than that generated by the random lysine sites (p -value $< 2.2e-16$ by Pearson's Chi-squared test with Yates' continuity correction). With vote number 9 as the threshold, the proportion of predicted substrates from the acetylation sites were still significantly higher than that from the random lysine sites (23.55% VS 10.76%, p -value $< 2.2e-16$ by Pearson's Chi-squared test with Yates' continuity correction). These results indicated that our method could discriminate the substrates of SIRT1 from others.

Performance evaluation with independent dataset and experimental validation. In a previous study, a group of putative SIRT1 substrate sites in mouse were generated by Chen *et al.*¹⁸. Totally 270 SIRT1 substrate proteins that contained lysine sites exhibiting 2-fold change of acetylation level between WT and KO cells were collected (Supplementary Table S9). Those 270 proteins were taken as the putative substrate proteins of SIRT1. After ortholog mapping of those mouse proteins to human by BioMart and excluding those proteins in our positive set, 260 proteins were left. Of those 260 putative SIRT1 substrate proteins by Chen *et al.*, 226 were in the list of the 6,338 human acetylation proteins. Though those 226 putative substrates might contain some indirect substrates of SIRT1, we still expected a significant proportion of those 226 proteins can be classified as the substrates of SIRT1 by our method. As shown in Fig. 5, the overlap between the 1,590 predicted SIRT1 substrates by our method (with the vote number 5 as the cutoff) and the 226 putative SIRT1 substrates by Chen *et al.* is 107 (p -value $< 7.138e-15$ by Pearson's Chi-squared test with Yates' continuity correction), which revealed that our method is effective in filtering SIRT1 substrates.

Besides compared with other study, we designed biological assays to confirm the prediction results. According to the laboratory condition and the prediction results, we selected 9 proteins with highest voting scores (9 votes), which included SMARCA5 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5), MTA2 (metastasis associated 1 family, member 2), HMGA1 (high mobility group AT-hook 1), EGFR (epidermal growth factor receptor), CHD4 (chromodomain helicase DNA binding protein 4), GOT2 (glutamic-oxaloacetic transaminase 2), CDK6 (cyclin dependent kinase 6), FANCD2 (Fanconi anemia complementation group D2) and MSH6 (mutS homolog 6). The acetylation status of these 9 proteins were monitored after overexpression of SIRT1. When SIRT1-Flag tagged expression plasmid was transfected into Human colon cancer HCT116 cell line, the acetylation levels of SMARCA5, MTA2, HMGA1, EGFR, CHD4 and GOT2 decreased as equal amounts of the proteins were immunoprecipitated (Fig. 6). The other three proteins CDK6, FANCD2 and MSH6 did not exhibit any change, at least under the condition of overexpression of SIRT1 (see Supplementary Fig. S1). For the proteins that could not be deacetylated by the way of overexpressing SIRT1 plasmid, there might be several reasons. First, these proteins really could not be deacetylated by SIRT1, which means they were not SIRT1 substrates. Second, these proteins could be deacetylated by SIRT1 but not in the cell line we used. Third, because of the feature of different proteins, the deacetylation by SIRT1 were not detected under the experimental state we built. In addition, sometimes deacetylation happened under specific cell signal stimulation, for example, DNA damage treatment or growth factors. The overexpression of Class III HDACs might be insufficient to induce deacetylation in some conditions. In order to exclude the possibility of non-specific deacetylation, 7 acetylated proteins predicted not to be substrates were tested. These 7 acetylated proteins ACSL1 (acyl-CoA synthetase long-chain family member 1), USP47 (ubiquitin specific peptidase 47), OTUB1 (OTU deubiquitinase, ubiquitin aldehyde binding 1), KDM2A (lysine demethylase 2A), JADE2 (jade family PHD finger 2), FABP5 (fatty acid binding protein 5), INTS3 (integrator complex subunit 3) were selected from the lists with score 0, the results

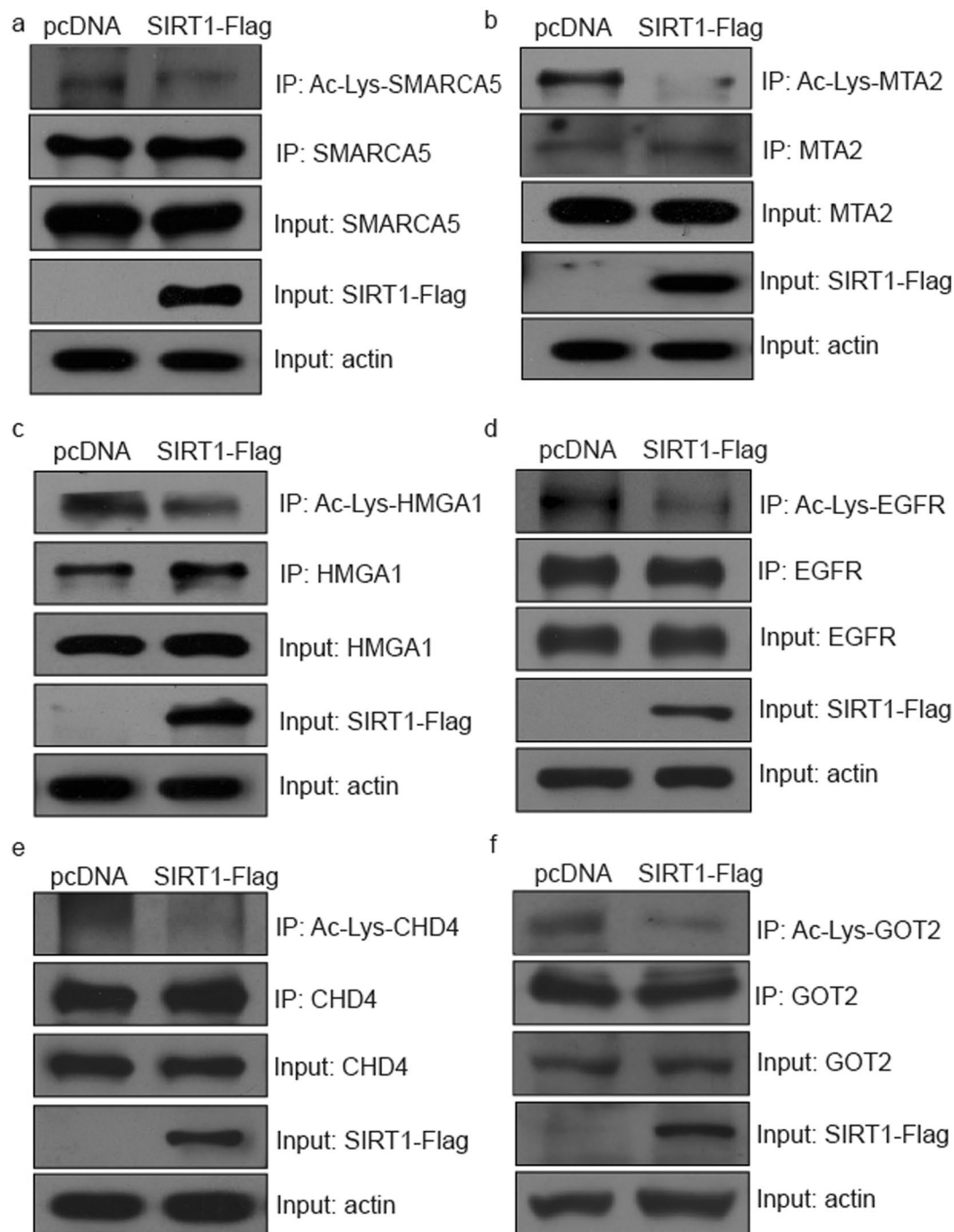


Figure 6. (a–f) Detection of predicted deacetylation substrates by SIRT1 with immunoprecipitation and Western blotting. Acetylation levels of SMARCA5 (a), MTA2 (b), HMGA1 (c), EGFR (d), CHD4 (e) and GOT2 (f) decreased after SIRT1 overexpression. pcDNA: pcDNA-vector transfection. SIRT1-Flag: Flag-tagged SIRT1 plasmid transfection as indicated in the figure. Equal amounts of indicated proteins were immunoprecipitated, followed by western blotting with pan-lysine acetylation antibody, used to detect the acetylation of immunoprecipitated proteins.

were consistent with our predictions that the acetylation levels of these 7 acetylated proteins did not show any change after SIRT1 overexpression (see Supplementary Fig. S2).

Discussion

In this study, we proposed a computational method to filter substrates of SIRT1 through integrating sequence features and various functional features. Primary sequence feature has been widely used to predict PTM substrates. But some types of PTMs do not have significant sequence specificity, like the substrates of SIRT1 in this study. Poor performance of prediction with only sequence features demonstrated that classifiers with

sequence features alone are insufficient to recognize SIRT1 substrates (Supplementary Table S2). The performances of prediction based on SVM gained significant improvement after integrating functional features with sequence features (Fig. 3). We also tested the prediction performance of Random Forests and Neural Networks after integrating functional features with sequence feature. We found that with either Random Forests or Neural Networks, the prediction performance could be improved greatly (Supplementary Table S10), and Random Forests even performs better than SVM. These results demonstrated that no matter what kind of classifiers were adopted, the benefit of combining functional information with primary sequences in the prediction of SIRT1 substrates was obvious. In addition, since there is no software available for SIRT1 substrate prediction, comparisons with bioinformatics approaches for identifying substrates of other acetylation enzymes were also conducted. We made prediction of SIRT1 substrates with the ASEB (Acetylation Set Enrichment Based method)²³ and GPS-PAIL (GPS Prediction of Acetylation on International Lysines)²⁴ methods and comparing the prediction performance based on Sn, Sp and MCC (details of the experimental processes can be found in the Supplementary Information). The Sn, Sp and MCC of ASEB were 0.19, 0.92 and 0.1498. The Sn, Sp and MCC of GPS-PAIL were 0.13, 0.92 and 0.0730. Both of these two methods made prediction based on primary sequence information and their prediction performance were much worse than our new method, whose Sn, Sp and MCC were 0.71, 0.92 and 0.6549. All these results confirmed the benefit of combining functional information with primary sequences in predicting SIRT1 substrates.

Among the five types of functional information we used, PPI brought the most significant improvement to prediction performance (Fig. 3). As shown in Supplementary Table S3, 626 proteins were found to be significantly over-represented with *p*-value cutoff 0.01, which means that a significant number of SIRT1 substrates interact with these 626 proteins. The protein with the most significant *p*-value was SIRT1 itself, which is reasonable. The other top significant proteins include p300 and p53. It has been found that p300 cooperates with SIRT1 to regulate the histone H2K56 acetylation homeostasis *in vivo*²⁵. For p53, many papers revealed that SIRT1 exerted multiple effects on p53 by directly deacetylating p53 or managing other factors such as Tip60 which was deacetylated by SIRT1 and acetylates p53^{26–28}. With those reasonable functional terms as prediction features, it is not surprising to improve the prediction performance greatly. We then screened SIRT1 substrates from the known acetylation sites and obtained a reasonable list of potential SIRT1 substrates.

Altogether, our method presented here provides novel and helpful insights for biologists to investigate the substrates of SIRT1, and should facilitate the cognition of SIRT1 deacetylation in health and disease.

Materials and Methods

Data preparation. We manually collected substrates of SIRT1 from published literature with key words, SIRT1*[Title/Abstract] AND deacetylate*[Title/Abstract], in PubMed, and 2,470 papers were returned until April 9th, 2015. After reviewing, 129 SIRT1 substrate sites from 54 human proteins were obtained (Supplementary Table S1). For deacetylated sites, we compiled their peptides with central deacetylated lysine and surrounding 10 amino acids on both sides, 118 substrate peptides from 51 proteins were remained (Supplementary Table S1), which were taken as positive samples here. The source reference of each deacetylated sites can be found in Supplementary Table S1 and researchers can gain more details from the original articles. Negative samples were randomly selected from the background set.

Sequence and functional features coding. Here, machine learning package scikit-learn²⁹ (scikit-learn-0.16.1) was used to construct SVM models with the radial basis function (RBF) kernel. To code these functional features for SVM classifier, for those selected functional terms, if the proteins have the feature, the corresponding bit would be represented by 1, otherwise represented by 0. For the sequence features, we simply defined a 20-bit binary tuple for each amino acid, for instance, lysine (K) is represented by a 20-dimensional vector [0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0]^T. Therefore, if the peptide length of candidate sequence is 21 and N functional features were enriched, the dimension of input vector is 20*21+N.

Statistical parameters for evaluating. In our statistical analysis, we abbreviated True Positive, True Negative, False Positive and False Negative as *TP*, *TN*, *FP* and *FN*. Sn, Sp and MCC are defined as equations 1, 2 and 3:

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TN}{TN + FP} \quad (2)$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (3)$$

Cell culture and transfection. Human colon cancer cell line HCT116 used in this research were purchased from the American Type Culture Collection. The cells were cultured in McCoy's 5A Medium supplemented with 10% foetal bovine serum (FBS) and 1% antibiotics. All cell lines were maintained in

a humidified incubator at 37 °C with 5% CO₂. One 100mm plate using lipotransfectamine 2000 (20 µl; Invitrogen) and plasmids (8 µg) to perform transfection. The transfected cells were then harvested at 48 h post-transfection.

Co-immunoprecipitation and Western blot. Cell extracts were prepared by lysing cells in Nonidet P-40 buffer (50 mM Tris-HCl, pH 7.5, 300 mM NaCl, 1% Nonidet P-40, 0.1% SDS, 2 mM EDTA, 10 mM sodium butyrate, 1% cocktail (Roche, Basel, Switzerland)). SMARCA5, MTA2, HMGA1, EGFR, CHD4 and GOT2 were immunopurified from clarified supernatant with the indicated antibodies incubated at 4 °C overnight. Then, 30 µL of protein G or A Sepharose slurry (GE healthcare, NY, USA) was added and incubated for 2 h at 4 °C. The beads were washed by Nonidet P-40 buffer three times at 1,000 rpm 4 °C for 1 min. The precipitated components were analyzed by Western blotting as previously described with minor modifications³⁰. Anti-pan-Lysine acetylation antibody was used to detect the acetylation of immunoprecipitated proteins.

References

- Lothrop, A. P., Torres, M. P. & Fuchs, S. M. Deciphering post-translational modification codes. *FEBS letters* **587**, 1247–1257, doi:10.1016/j.febslet.2013.01.047 (2013).
- Shahbazian, M. D. & Grunstein, M. Functions of site-specific histone acetylation and deacetylation. *Annual review of biochemistry* **76**, 75–100, doi:10.1146/annurev.biochem.76.052705.162114 (2007).
- Yang, X. J. & Seto, E. HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene* **26**, 5310–5318, doi:10.1038/sj.onc.1210599 (2007).
- North, B. J. & Verdin, E. Sirtuins: Sir2-related NAD-dependent protein deacetylases. *Genome biology* **5**, 224, doi:10.1186/gb-2004-5-5-224 (2004).
- de Ruijter, A. J., van Gennip, A. H., Caron, H. N., Kemp, S. & van Kuilenburg, A. B. Histone deacetylases (HDACs): characterization of the classical HDAC family. *The Biochemical journal* **370**, 737–749, doi:10.1042/BJ20021321 (2003).
- Finkel, T., Deng, C. X. & Mostoslavsky, R. Recent progress in the biology and physiology of sirtuins. *Nature* **460**, 587–591, doi:10.1038/nature08197 (2009).
- Imai, S., Armstrong, C. M., Kaerberlein, M. & Guarente, L. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature* **403**, 795–800, doi:10.1038/35001622 (2000).
- Vaquero, A., Sternglanz, R. & Reinberg, D. NAD⁺-dependent deacetylation of H4 lysine 16 by class III HDACs. *Oncogene* **26**, 5505–5520, doi:10.1038/sj.onc.1210617 (2007).
- Vaquero, A. *et al.* Human SirT1 interacts with histone H1 and promotes formation of facultative heterochromatin. *Molecular cell* **16**, 93–105, doi:10.1016/j.molcel.2004.08.031 (2004).
- Jeong, J. *et al.* SIRT1 promotes DNA repair activity and deacetylation of Ku70. *Experimental & molecular medicine* **39**, 8–13, doi:10.1038/emm.2007.2 (2007).
- Yuan, Z., Zhang, X., Sengupta, N., Lane, W. S. & Seto, E. SIRT1 regulates the function of the Nijmegen breakage syndrome protein. *Molecular cell* **27**, 149–162, doi:10.1016/j.molcel.2007.05.029 (2007).
- Li, K. *et al.* Regulation of WRN protein cellular localization and enzymatic activities by SIRT1-mediated deacetylation. *The Journal of biological chemistry* **283**, 7590–7598, doi:10.1074/jbc.M709707200 (2008).
- Liu, Y. *et al.* A fasting inducible switch modulates gluconeogenesis via activator/coactivator exchange. *Nature* **456**, 269–273, doi:10.1038/nature07349 (2008).
- Yeung, F. *et al.* Modulation of NF- κ B-dependent transcription and cell survival by the SIRT1 deacetylase. *The EMBO journal* **23**, 2369–2380, doi:10.1038/sj.emboj.7600244 (2004).
- van Loosdregt, J. *et al.* Regulation of Treg functionality by acetylation-mediated Foxp3 protein stabilization. *Blood* **115**, 965–974, doi:10.1182/blood-2009-02-207118 (2010).
- Blander, G. *et al.* SIRT1 shows no substrate specificity *in vitro*. *The Journal of biological chemistry* **280**, 9780–9785, doi:10.1074/jbc.M414080200 (2005).
- Chen, Y., Colak, G. & Zhao, Y. SILAC-based quantification of Sirt1-responsive lysine acetylation. *Methods in molecular biology* **1077**, 105–120, doi:10.1007/978-1-62703-637-5_7 (2013).
- Chen, Y. *et al.* Quantitative acetylation analysis reveals the roles of SIRT1 in regulating diverse substrates and cellular pathways. *Molecular & cellular proteomics: MCP* **11**, 1048–1062, doi:10.1074/mcp.M112.019547 (2012).
- Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council* **10**, 988–999, doi:10.1109/72.788640 (1999).
- Borra, M. T., Langer, M. R., Slama, J. T. & Denu, J. M. Substrate specificity and kinetic mechanism of the Sir2 family of NAD⁺-dependent histone/protein deacetylases. *Biochemistry* **43**, 9877–9887, doi:10.1021/bi049592e (2004).
- Garske, A. L. & Denu, J. M. SIRT1 top 40 hits: use of one-bead, one-compound acetyl-peptide libraries and quantum dots to probe deacetylase specificity. *Biochemistry* **45**, 94–101, doi:10.1021/bi052015l (2006).
- Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188–1190, doi:10.1101/gr.849004 (2004).
- Li, T. *et al.* Characterization and prediction of lysine (K)-acetyl-transferase specific acetylation sites. *Molecular & cellular proteomics: MCP* **11**(M111), 011080, doi:10.1074/mcp.M111.011080 (2012).
- Deng, W. *et al.* GPS-PALL: prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Scientific reports* **6**, 39787, doi:10.1038/srep39787 (2016).
- Das, C., Lucia, M. S., Hansen, K. C. & Tyler, J. K. CBP/p300-mediated acetylation of histone H3 on lysine 56. *Nature* **459**, 113–117, doi:10.1038/nature07861 (2009).
- Luo, J. *et al.* Negative control of p53 by Sir2alpha promotes cell survival under stress. *Cell* **107**, 137–148 (2001).
- Tang, Y., Luo, J., Zhang, W. & Gu, W. Tip60-dependent acetylation of p53 modulates the decision between cell-cycle arrest and apoptosis. *Molecular cell* **24**, 827–839, doi:10.1016/j.molcel.2006.11.021 (2006).
- Peng, L. *et al.* SIRT1 negatively regulates the activities, functions, and protein levels of hMOF and TIP60. *Molecular and cellular biology* **32**, 2823–2836, doi:10.1128/MCB.00496-12 (2012).
- Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- Zhu, W. G. *et al.* 5-aza-2'-deoxycytidine activates the p53/p21/Waf1/Cip1 pathway to inhibit cell proliferation. *The Journal of biological chemistry* **279**, 15161–15166, doi:10.1074/jbc.M311703200 (2004).

Acknowledgements

We thank Zheng Wu for the help in collecting SIRT1 substrate data. This work was supported by the National Natural Science Foundation of China [No. 31371337 and No. 91319302] and the National High-Tech Research and Development Program (Grant No. 2012AA020401) of China.

Author Contributions

T.L. and W.Z. designed the study. Z.Z., M.L. programmed the computational method. M.T. handled the verifying experiments. Z.Z., T.L., M.T. and Y.Y. wrote the manuscript. All authors contributed to data interpretation and approved the final manuscript. T.L. and W.Z. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Additional Information

Supplementary information accompanies this paper at doi:[10.1038/s41598-017-04847-7](https://doi.org/10.1038/s41598-017-04847-7)

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017