Proteome-wide prediction of acetylation substrates

Amrita Basu[a], Kristie L. Rose[b], Junmei Zhang[c], Ronald C. Beavis[d], Beatrix Ueberheide[e], Benjamin A. Garcia[f], Brian Chait[e], Yingming Zhao[g], Donald F. Hunt[b], Eran Segal[h,1], C. David Allis[a,1], and Sandra B. Hake[i,1]

[a]Laboratory of Chromatin Biology, Rockefeller University, New York, NY 10065; [b]Departments of Chemistry and Pathology, University of Virginia, McCormick Road, Charlottesville, VA 22904; [c]Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390; [d]Biomedical Research Centre, University of British Columbia, Vancouver, BC, Canada V6T 1Z3; [e]Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, Rockefeller University, New York, NY 10065; [f]Department of Molecular Biology, Princeton University, Princeton, NJ 08544; [g]Ben May Department for Cancer Research, University of Chicago, Chicago, IL 60637; [h]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel; and [i]Department of Molecular Biology, Adolf-Butenandt-Institut and Center for Integrated Protein Science Munich, Ludwig Maximilian Universität 80336, Munich, Germany

Contributed by C. David Allis, June 18, 2009 (sent for review May 21, 2009)

Acetylation is a well-studied posttranslational modification that has been associated with a broad spectrum of biological processes, notably gene regulation. Many studies have contributed to our knowledge of the enzymology underlying acetylation, including efforts to understand the molecular mechanism of substrate recognition by several acetyltransferases, but traditional experiments to determine intrinsic features of substrate site specificity have proven challenging. Here, we combine experimental methods with clustering analysis of protein sequences to predict protein acetylation based on the sequence characteristics of acetylated lysines within histones with our unique prediction tool PredMod. We define a local amino acid sequence composition that represents potential acetylation sites by implementing a clustering analysis of histone and nonhistone sequences. We show that this sequence composition has predictive power on 2 independent experimental datasets of acetylation marks. Finally, we detect acetylation for selected putative substrates using mass spectrometry, and report several nonhistone acetylated substrates in budding yeast. Our approach, combined with more traditional experimental methods, may be useful for identifying acetylated substrates proteome-wide.

histone | nonhistone | prediction | acetylation | PredMod

More than 40 years ago, Allfrey et al. (1) reported a strong correlation between increased levels of histone acetylation and elevated levels of gene expression. Since then, the field of chromatin biology has advanced considerably with remarkable progress made into mechanistic insights of histone modifications and their biological functions. Histones are abundant nuclear proteins known to contain a wealth of posttranslational modifications (PTMs) including, among others, acetylation, methylation, and phosphorylation. These PTMs may contribute to "epigenetic signatures" that play a role in diverse biological processes. Of the known PTMs, acetylation has the capacity to destabilize the chromatin polymer through charge neutralization of the basic lysine residue potentially harboring structural consequences for higher-order chromatin structures (cis effects) (2–4). Furthermore, acetylation recruits specialized "effector" proteins that in turn affect chromatin structure (trans effects) (3), as has been proposed in the histone code hypothesis (5).

Lysine acetylation in histones was the first PTM identified to be regulated by a highly balanced enzyme system that contains lysine acetyltransferases (KATs) and histone deacetylases (HDACs), which are responsible for governing a steady-state balance of acetylation (6, 7). Certain KATs have been shown to also acetylate nonhistone transcription-related proteins, and finally, acetylation has emerged to play a critical role in human biology and disease. Promising advances have been made recently in developing drug therapies that target HDACs for certain cancers (8). A computational tool that is predictive of acetylation events could contribute to a more complete understanding of what substrates are physiologically relevant, as more insights are gained into acetylation-mediated pathways.

Conventional experiments [e.g., mutagenesis, antibodies, and mass spectrometry (MS)] have typically been used to identify acetylated lysines in substrate proteins. These methods are often laborious, time intensive, and expensive. Therefore, a robust computational prediction tool is desirable to reduce the number of experiments needed to identify potential PTM sites in proteins of interest. Past computational studies suggest that there are canonical motifs in acetylated substrates proteome-wide (9). Our approach sets out to test whether novel acetylation marks can be predicted using a combined experimental and computational approach. Our analysis focuses on histones because these are widely studied, heavily acetylated substrates. Briefly, we train a "classifier" from histone sequences in an unbiased manner, assign nonhistone sequences into the clusters defined in the training phase, and finally generate predictions based on the acetylation states of the histone lysines within the cluster assigned. We report the results of a computational approach, combined with experimental validation, and present a unique software tool, PredMod, which may assist in predicting candidate acetylation sites proteome-wide.

Results

Training Set and Key Assumptions. We used histones as a training set because of the wealth of information known about their PTM patterns and well-developed purification and analytical detection methods, and focused on the major human core histones bearing a total of 56 lysines (H2A: 13; H2B: 19; H3: 13; H4: 11) (Fig. 1*A*). To date, MS and antibody data suggest that there are 23 "validated" acetylated lysines and 33 lysines that have not yet been observed as acetylated in human histones based on literature [supporting information (SI) Table S1]. We sought to uncover additional acetylation sites within the "not observed" class of lysines in a systematic, rigorous manner via our computational method. We selected parameters that could influence our ability to predict acetylation sites on histones by making a series of assumptions. First, we focused our attention on short stretches of amino acids N- and C-terminal of all 56 lysines. Because structural studies of published KAT domains coupled with peptide substrates typically do not exceed 14–20 aa in length (10, 11), a sliding window of a maximum number of 12 residues flanking each lysine was chosen (Fig. 1*B*). Residues most proximal to the lysine were given the highest weight (Fig. 1*B*), assuming that these residues are most important for enzyme recognition, as several studies have shown (10, 11). Second, we

Author contributions: A.B., B.C., E.S., and C.D.A. designed research; A.B., K.L.R., J.Z., and B.U. performed research; A.B., R.C.B., B.A.G., and Y.Z. contributed new reagents/analytic tools; A.B., D.F.H., B.C., and S.B.H. analyzed data; and A.B., E.S., C.D.A., and S.B.H. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: sandra.hake@med.uni-muenchen.de, alliscd@mail.rockefeller.edu or eran.segal@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/cgi/content/full/0906801106/DCSupplemental.

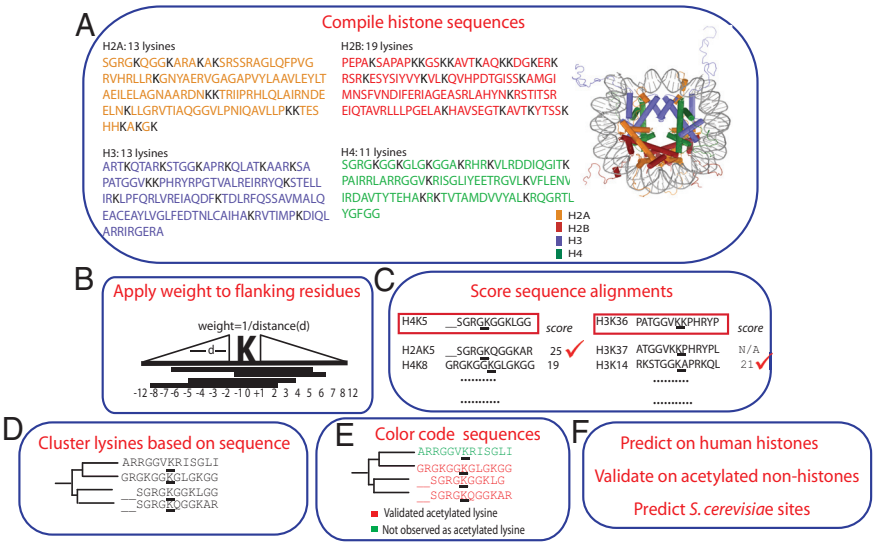

Fig. 1. Schematic of the overall computational and experimental approach. (*A*) Human core histone proteins (H2A: orange; H2B: red; H3: blue; H4: green) containing 56 lysines (black) were taken as input data for computational training. (*B*) A sliding window of amino acids (black bars) flanking the input lysine (at position 0) is used to train the model. Not all window lengths are shown. Weights (calculated as inversely proportional to distance [d]) are applied to amino acids based on the distance from the input lysine to the amino acid in positions −12 to +12. (*C*) BLAST sequence alignments are performed between all 56 lysines and surrounding sequences, and the highest scoring alignment is selected to begin the clustering analysis. Shown are sequences H4K5 and H3K36 (boxed in red) spanning positions −6 to +6 and their highest scoring match (denoted by a checkmark). Note that H4K5 and H2AK5 do not have 6 residues flanking the lysine N-terminally; scores are normalized based on length in these cases. (*D*) Lysines clustered together based on sequence alignment scores creating a fully predictive hierarchical tree (4 sequences are shown here; all 56 sequences are shown in Fig. 2). (*E*) Sequences are color coded according to published data on their modification state. Red: validated evidence of the lysine being acetylated; green: this lysine was not observed as being acetylated in literature. (*F*) After establishing PredMod, predictions were made on lysines in human core histones. The algorithm was then validated using a set of human acetylated proteins reported in literature, substrates detected using a pan-acetyl IP approach, and a yeast proteome-wide dataset. Finally, predictions were made on yeast nonhistone sites and validated in vivo.

varied standard BLAST sequence alignment parameters, including gap penalty, extension, insertion, and deletion scores (Fig. 1*C*). For lysines in the extreme N- and C-terminal region, such as H3K4 or H2AK129, we normalized the raw alignment score based on the length of the sequence. Additionally, both orientations of the protein sequence (N-terminal to C-terminal or vice versa) were weighted equally. For sequences with lysines that are located in close proximity to each other, such as H3K36 and H3K37, we restricted our alignment matrix so that these sequences did not receive an alignment score. This restriction prevented our training set to be overrepresented with sequences from overlapping fragments of the same protein. Finally, we compensated for structural accessibility by penalizing buried lysines and improving the score of accessible lysines (12). This, however, did not influence our ability to predict acetylation sites on histones, and therefore was not included in our further computations.

We performed a hierarchical clustering of core histone lysines based on the sequences surrounding each of these given lysines. All 56 histone core sequences were aligned to one another, creating a matrix of pairwise alignment scores, generating a hierarchical tree of histone sequences (Fig. 1*D*). We next classified each lysine into 1 of 2 categories based on its acetylation status reported in literature: "validated" (23 lysines) or "not observed" (33 lysines) (Table S1). Finally, we visually categorized each of the 56 lysines by color coding our tree based on the acetylation status of each lysine (Fig. 1*E*).

To assess how robust our clustering was and how well it could actually predict lysine acetylation, we took all 56 lysines and performed a leave-one-out cross-validation (LOV) (13) by iteratively excluding one lysine from our training set. Next, we reconstructed the hierarchical tree with the remaining 55 lysines and incorporated the excluded single lysine observation as test data. For each set and combination of predefined parameters (stated above) and in a single run, we performed a LOV analysis to examine the predictive power on all 56 lysines to discover which set of parameters best optimized classification power. If 2 lysines were in overlapping fragments of the same protein, we excluded both of these lysines from our training set when either lysine was a test case. We took each test lysine (total of 56) and traversed through our training tree to find which subgroup of sequences our target sequence formed the tightest cluster with.

A receiving operating curve (ROC) analysis was performed on our test dataset (Fig. S1), where the statistics measure used was the area under curve (AUC). An AUC of 1 represents a perfect prediction, and an AUC of 0.5 random predictions. Each point on a single curve of the ROC plot was calculated by measuring the false positive versus true positive rate of the performance on all 56 lysines for a given parameter(s) under a cutoff alignment score. If the test lysine clustered within a group of validated acetylated lysines (Fig. 2*A*, red) above the cutoff score, the lysine was predicted to be acetylated. Conversely, if the test lysine clustered within a group of not-observed lysines (Fig. 2*A*, green) above the alignment score, the lysine was predicted as not acetylated. The default status of the lysine when it did not fall into the above criteria was not acetylated. The best ROC plot achieved an AUC of 0.80, and the parameters in this case included 6 weighted residues to both the left and right of the tested lysine (Fig. S1). A threshold for prediction was also determined based on this plot. To test the significance of this score, we applied the previous procedure to 1,000 random permutations of the labels of the observed and not-observed lysines. The median AUC in these permutations was 0.64, and the maximum score was 0.79; thus, our AUC was statistically significant ($P < 0.001$).

Computational Prediction of Novel Human Histone Acetylation Marks and in Vivo Validation by Mass Spectrometry. After hierarchical clustering of all our lysine-embedded histone sequences, we next

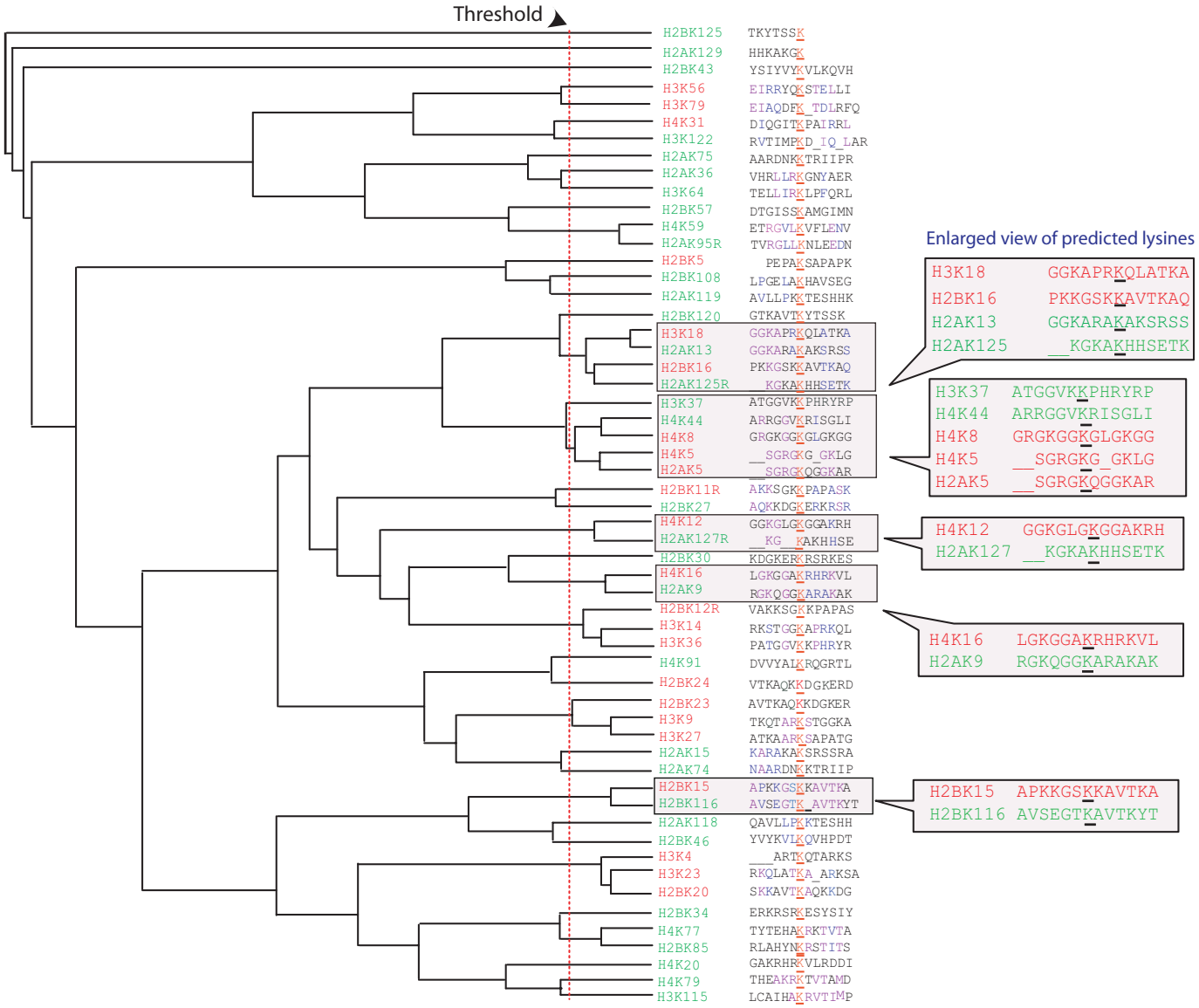

Fig. 2. Computational prediction of human histone acetylation sites. Predictive tree of all 56 lysines from human core histone sequences using hierarchical clustering (see *SI Text* for details). Histone lysines (in red or green) are color coded according to published data on their modification state as described in Fig. 1*E*. For each pair of sequences under a single node, amino acids are colored in light purple (identical residues) or dark blue (in accordance with the BLOSUM matrix) (25). Underlined red lysines represent the residue that was used for training the algorithm. Dashed red vertical line represents the selected threshold used to make predictions. Gray boxes represent a zoomed-in view of lysines that cluster together. An *R* next to the lysine indicates that a C- to N-terminal arrangement was used in the alignment.

sought to predict novel acetylation sites in the human core histones. As our tree illustrates in Fig. 2, not-observed lysines that clustered tightly with validated acetylated lysines (green sequences in gray captions) were potential acetylation targets because of their similar sequence constitution. Based on the threshold, determined by the ROC plot, we selected these as candidate sites. The previous method predicted 7 unique acetylation sites in the human core histones; 4 in H2A (K9, K13, K125, K127), 1 in H2B (K116), 1 in H4 (K44), and 1 in H3 (K37) (Fig. 2). This large number of predictive sites was unexpected because histones have been intensely investigated for PTMs in recent years. To test whether these predicted lysines are acetylated in vivo, we used an MS-based approach to examine histone peptides from human cell lines that were asynchronously growing and treated without any HDAC inhibitors (see *SI Text*). All peptides containing the predicted lysines were identified, and importantly, 4 of our 7 predicted acetyl-lysines were experimentally validated: H2AK9, H2AK13, H2AK125, and H2AK127 (Figs. S2 and S3). Histones H3 and H2B from sodium butyrate-treated human cells also showed H3K37 and H2BK116 acetylation, but because these marks were observed only under these special conditions (see *Discussion*), we did not count them as validated.

In summary, we correctly predicted 4 of the 7 acetyl-lysine sites, suggesting that our algorithm is capable of identifying acetylation sites in human histone proteins.

Nonhistone Sequence-Based Dataset Prediction and Validation. Because our computational analysis revealed a high level of sequence homogeneity among acetylated lysines within histone proteins, leading to the successful prediction of unique modified residues, we next wondered if our approach might also enable us to predict nonhistone acetylation sites.

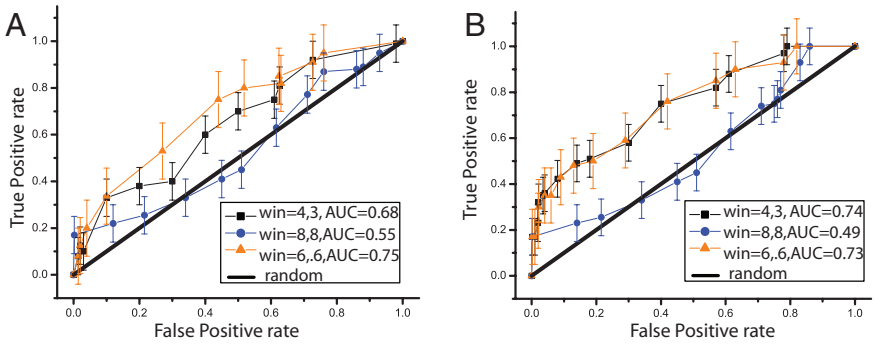

Fig. 3. Prediction performance on human nonhistone substrates. ROC curve for human pan-acetyl IP substrate test set (*A*) and literature-validated human acetylated proteins (*B*). The y axis represents the true positive rate, and the x axis the false positive rate. Win = (x,y) denotes the length of residues spanning the lysine; x: number of residues N-terminal to the lysine; y: number of residues C-terminal to the lysine. Diagonal line represents a random prediction.

In our first approach, we included a dataset that contained both nuclear and cytosolic proteins from HeLa cells, which were immunoprecipitated with a pan-acetyl antibody (Fig. S4*A* and Table S2) and identified by MS (14). The precipitate contained peptides with a total of 1,413 lysines, and 51 previously validated acetylation sites. With PredMod, we were able to predict 34 (67%) of these sites correctly (Fig. S4*A*) when they were surrounded by 6 residues to the left and right (AUC = 0.75, sensitivity S_n = 0.66, specificity S_p = 0.94) (Fig. 3*A*, orange curve). In total, 6% (85) of the total number of lysines were predicted that were not validated as acetylated (Fp < 6%). Fp is a maximum false positive rate; a true negative count cannot be accurately determined because many of these lysines could potentially be acetylated, but not detected under the experimental procedures used.

In our second dataset, we compiled a list of 32 proteins containing 1,378 lysines with 73 of these reported in literature to be acetylated in vivo and/or in vitro (Fig. S4*B* and Table S3). With PredMod, we predicted 39 of 73 (53%) lysine marks accurately with Fp < 6.5% (AUC = 0.74, S_n = 0.58, S_p = 0.93) when these were surrounded by six residues to the left and right (Fig. 3*B*, orange).

Both test datasets exhibited a decrease in performance when larger numbers of residues N- and C-terminal to the target lysine were used (Fig. 3, blue line), suggesting that KATs may recognize a smaller and defined set of residues. Overall, findings from both approaches revealed that our selected parameters for histones were also valid for the prediction of acetylated nonhistone substrates using an ROC analysis approach.

Analysis of Acetylation Motifs. We next sought to understand which amino acids play a critical role in acetylation site selection, and asked whether there were preferences for certain amino acids near the target acetylated lysines in our datasets. Notably, when we examined the surrounding residues (six residues to the left and right) of a validated acetylated lysine versus a not-observed one in human histone and nonhistone proteins we discovered an enrichment for small residues (G/A in pink), lysines (K in green), and phosphorylatable residues (S/T in blue) (Fig. 4). To test whether the observed enrichment of G, K, S was statistically significant, we determined the frequency of these residues flanking a lysine in the entire human proteome. We noticed that on average, these residues were of significantly higher frequency in our datasets than in the human proteome. We used the hypergeometric test to measure the statistical relevance of this observation (Table S4). Our findings show that the most significant *P* values were found in the category of small residues ($P < 0.01$ in multiple flanking positions; Fig. 4, tick marks), suggesting that small amino acids, perhaps due to their sterically undemanding side chains, could accommodate the flexibility of the substrate, thus allowing protein docking and catalysis. This observation was in agreement with a previous study (9), which revealed that glycine preceding lysine was common among acetylated lysines. In conclusion, we were able to identify a significant enrichment of mainly small amino acids and lysines surrounding validated acetylated lysines in comparison with not-observed ones, suggesting that KAT enzymes have a general need for specific residues for recognition and/or activity. These observations are in agreement with studies of several KATs with test substrates (10, 11).

***S. cerevisiae* Proteome-Wide Prediction and in Vivo Validation.** The previous predictions were performed with human proteins, and we therefore wondered whether our algorithm would also be able to predict acetylation sites in proteins from other organisms. Because histone acetylation has been studied extensively in budding yeast, we assessed the performance of our model on a proteome-wide dataset that included acetylated peptides in *S. cerevisiae* (15) (see *SI Text*). In addition, we experimentally validated our predicted acetylation sites in candidate yeast nonhistone proteins in vivo.

In our first approach, we examined in vitro a proteome-wide dataset of acetylated peptides of *S. cerevisiae* that contained 356 peptides, including acetylated histone peptides (see *SI Text*). This dataset allowed us to approximate the number of yeast acetylation events on a global level (0.6%; see *SI Text*), and the substrates themselves allowed us to further validate our prediction algorithm. We filtered these protein-derived peptides according to their cellular compartment (nuclear vs. cytoplasmic) (16), and correctly predicted 43% of acetylation events on nuclear proteins (79 lysines total; AUC = 0.71, S_n = 0.41, S_p = 0.92, Fp < 4%) and 30% on the cytoplasmic proteins (248 lysines total; AUC = 0.70, S_n = 0.31, S_p = 0.90, Fp < 5%). We also noted that nuclear yeast proteins showed a similar enrichment for small residues surrounding the target lysine, as found in the human substrates (Fig. S5).

In our second approach, we validated our predictions on 3 yeast candidate proteins that had previously not been published to contain acetylated sites: Spt6 (17), Sir3 (18), and Eaf7 (19). We expressed and purified our tap-tagged candidate proteins in *S. cerevisiae* (Fig. 5*A*) and subsequently subjected them to MS. With PredMod, we predicted 15 sites to be acetlyated of 416 total lysines in our 3 candidate proteins combined. Four of these, within our top 6 ranked predicted sites (Fig. 5*B*), were validated as acetylated by MS and therefore predicted correctly (Fig. S6). The total number of acetylated lysines in the yeast proteome is ≈0.6%; therefore, our in vivo hit rate of ≈25% is of reasonable accuracy.

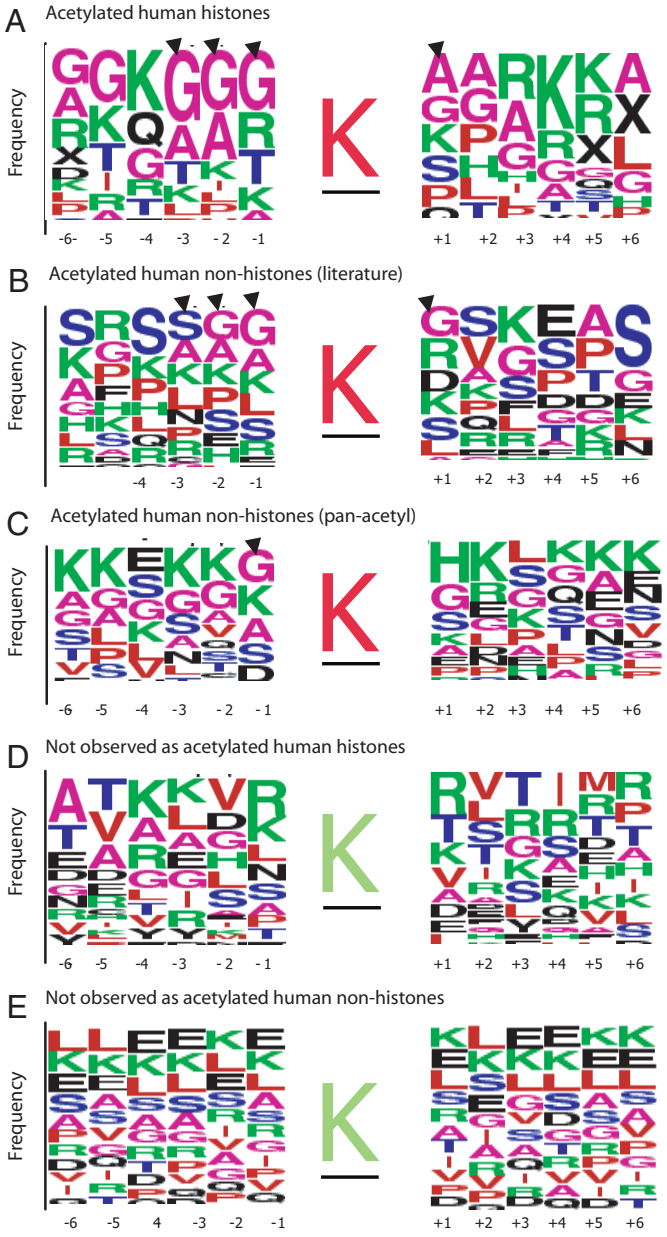

Fig. 4. Frequency distribution of amino acids surrounding lysines in human histone and nonhistone proteins. Frequency of amino acids (y axis) spanning positions −6 to +6 (x axis) in validated acetylated lysines in histone proteins (23 lysines) (*A*), validated acetylated lysines within proteins in literature (73 lysines) (*B*), validated lysines in the pan-acetyl IP substrates (51 lysines) (*C*), not observed as acetylated lysines in histones (33 lysines) (*D*), and not observed as acetylated lysines in proteins as reported in literature and not observed as acetylated lysines in pan-acetyl IP substrates (3,493 lysines) (*E*). Residues in green: basic; red: hydrophobic; pink: small; blue: S/T; black: all other residues. Underlined red K: lysine that has been validated experimentally as acetylated; underlined green K: lysine that has not been experimentally observed as acetylated. *X* denotes that no amino acid was present in that position. Tick marks represent residues described in text.

These findings show the power of PredMod for identifying bona fide acetylation sites in nonhistone proteins, and further display the strength of using histone sequences as a useful guide for nonhistone acetylation prediction.

Discussion

Acetylation plays a crucial role in the function of multiple cellular pathways (20). To better understand these processes, it

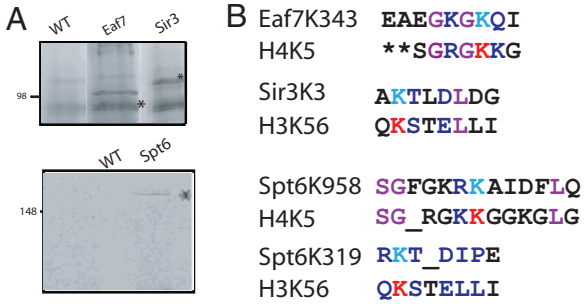

Fig. 5. Novel predictions and in vivo validation of *S. cerevisiae* nonhistone proteins. (*A*) Coomassie-stained gel of TAP-tag pull-down purification of yeast proteins Eaf7, Sir3, and Spt6. Asterisk denotes bands that were isolated and inspected for acetylation by MS. (*B*) Sequence alignment of candidate proteins with identical or similar histone regions. Light purple amino acid pairs represent identical residues, and dark blue pairs represent residues that can be evolutionarily substitutable in accordance with the BLOSUM matrix. Correctly predicted lysines are indicated in light blue. Red lysines are acetylated histone residues.

is first necessary to identify sites of acetylation within a protein of interest. Our prediction program, PredMod, is a first and promising step in finding novel acetylation sites, although we did not achieve 100% prediction capacity. We envision several possibilities as to why some predicted lysines have not yet been detected experimentally. First, the MS approach has limited detection and sensitivity capabilities and cannot recover peptides that are acetylated at only low levels. Second, lysines could be modified only in distinct environmental conditions, cell cycle stages, and cell types, and are therefore undetectable in the cell extracts we used. Additional histone acetylation sites were detected by MS/MS when HeLa cells were pretreated with HDAC inhibitors, and we retrained our algorithm with this data. Preliminary findings from this analysis show that the predictive power of our overall approach is not altered significantly, thereby increasing further confidence in the power of our approach (Table S5). Third, acetylation might be inhibited by adjacent PTMs (negative crosstalk), and therefore the responsible KAT might be prevented from binding to or accessing its target site. Finally, acetylation is a dynamic, transient modification, and MS results may therefore depend on a time-specific acetylation state whose kinetic properties have not been adequately captured by our experimental parameters. Of interest is the class of lysines in histones that was not predicted by our algorithm, yet detected by MS, and which might indicate a different class of KATs that need special sequence surroundings (see *SI Text*).

Our findings suggest that the sequence environment in both histone and nonhistone proteins contributes to the likelihood of acetylation. Consistent across both human and yeast acetyl datasets, we noticed an enrichment of small residues, particularly glycine, and charged amino acids flanking validated acetylated lysines (Fig. 4 and Fig. S5). It is possible that we are perhaps achieving a higher accuracy for nuclear proteins because the KAT substrates (histones) we used for training PredMod mostly reside in the nucleus. Our findings also suggest that nuclear versus cytoplasmic KATs could possess unique substrate recognition profiles, as illustrated by the differences in preferred flanking residues C-terminal to the lysine (Fig. S5 *B* and *C*).

As more substrates in the acetyl-proteome are discovered (20), it is likely that the predictive power of our approach will be strengthened, leading to more accurate confidence in the predicted site. The power of our approach is underscored by the fact that we are predicting significantly higher than random on nonhistones given our set of limited training data. Our training dataset is 10–100× in magnitude lower than other PTM datasets

including acetylation (9, 21, 22), yet our approximate sensitivity measure of 60% is comparable and often higher than other prediction algorithms that achieve as low as 16%−18% sensitivity (9). It would be interesting to see whether similar approaches could be applied to the prediction of other widespread histone modifications, such as lysine methylation.

Overall, our findings suggest that KATs target specific sequence patterns, and that the predictive knowledge about histone acetylation provides a useful platform for studying both histone and nonhistone lysine acetylation. Our model and findings represent a step toward gaining a framework for predicting lysine acetylation sites in both human and yeast proteomes. It will be of interest in future studies to see whether our algorithm is also capable of predicting lysine acetylation sites in many other organisms.

Materials and Methods

Cell Lines. Mammalian cell lines were grown in Iscove's DMEM supplemented with 10% FCS and penicillin/streptomycin at 37 °C and 5% CO_2.

Histone Isolation. Nuclei were isolated and histones acid-extracted from asynchronously growing, untreated cells as previously described (23). See *SI Text* for further details.

MS Analysis of Histones. Experimental details are described in *SI Text*.

MS Analysis of Yeast Nonhistone Acetylation Sites. Tagged cells of our nonhistone proteins were lysed under cryogenic conditions. Tandem TAP-tag purification was performed on candidate yeast proteins as described (24), and eluates run on SDS-PAGE gels and stained with Coomassie. Protein bands were in-gel digested with trypsin or chymotrypsin, and peptides extracted. Details of these methods are provided in *SI Text*.

Datasets. Training set: 56 human and *S. cerevisiae* core histone lysine sequences were collected from the Swiss-Prot database (http://ca.expasy.org/sprot/). Test set: source of nuclear protein and pan-acetyl antibody datasets are described in *Results*. For information on the budding yeast proteome-wide dataset, see *SI Text*.

Hierarchical Clustering Analysis. We performed hierarchical clustering on the sequences surrounding each of the 56 histone lysines. All 56 sequences were aligned to one another, creating a matrix of pairwise alignment scores; our metric was based on these pairwise scores. Sequence alignment scores were computed by performing BLAST local alignments using the NCBI BLAST 2.0 server. A standard BLOSUM62 evolutionary substitution matrix was applied (25).

Statistical Analysis. ROC calculations are described in the main text. Hypergeometric probability calculation: $\Pr = \binom{m}{K}\binom{N-m}{n-K}\binom{n}{N}$ (N, all lysines in human proteome; K, number of times the particular residue is seen flanking in each position in human proteome; n, total number of lysines in each independent validation dataset; m, number of times the particular residue is seen flanking in each position in validation dataset). Sensitivity (S_n) was calculated as the total number of correctly identified acetylation sites from the positive dataset divided by the total positive dataset. Specificity (S_p) was calculated as the total number of negative sites that were not predicted to be acetylated divided by the total negative dataset size. For additional information, please see *SI Text*.

Sequence Logos. Sequence logos for displaying the flanking residue distribution of all lysines in our training and test datasets were created according to ref. 26.

Software URL. Our acetylation prediction software, PredMod, can be found at www.cs.cornell.edu/w8/~amrita/predmod.html (see *SI Text*).

ACKNOWLEDGMENTS. We thank members of the Allis Lab for their constructive discussions. We especially thank A. Ruthenburg, M. Lachner, S. Whitcomb, and L. Banaszynski for careful reading of the manuscript. A.B. is a Tri-Institutional Computational Biology Fellow. This work was supported by a National Institutes of Health Merit Award (to C.D.A.), the Deutsche Forschungsgemeinschaft (German Research Foundation), and the Center for Integrated Protein Science Munich (S.B.H.).

1. Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci USA* 51:786–794.
2. Verreault A, Kaufman PD, Kobayashi R, Stillman B (1998) Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase. *Curr Biol* 8(2):96–108.
3. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. *Nat Struct Mol Biol* 14(11):1025–1040.
4. Grant PA (2001) A tale of histone modifications. *Genome Biol* 2(4):REVIEWS0003.
5. Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* 403(6765):41–45.
6. Brownell JE, et al. (1996) Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* 84(6):843–851.
7. Pflum MK, Tong JK, Lane WS, Schreiber SL (2001) Histone deacetylase 1 phosphorylation promotes enzymatic activity and complex formation. *J Biol Chem* 276(50):47733–47741.
8. Marks PA (2007) Discovery and development of SAHA as an anticancer agent. *Oncogene* 26(9):1351–1356.
9. Schwartz D, Chou MF, Church GM (2009) Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol Cell Proteomics* 8(2):365–379.
10. Marmorstein R (2001) Structure and function of histone acetyltransferases. *Cell Mol Life Sci* 58(5–6):693–703.
11. Marmorstein R (2001) Structure of histone acetyltransferases. *J Mol Biol* 311(3):433–444.
12. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 389(6648):251–260.
13. Cooper GF, et al. (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 9(2):107–138.
14. Kim SC, et al. (2006) Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Mol Cell* 23(4):607–618.
15. Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* 5(8):1843–1849.
16. Huh WK, et al. (2003) Global analysis of protein localization in budding yeast. *Nature* 425(6959):686–691.
17. Clark-Adams CD, Winston F (1987) The SPT6 gene is essential for growth and is required for delta-mediated transcription in Saccharomyces cerevisiae. *Mol Cell Biol* 7(2):679–686.
18. Gasser SM, Cockell MM (2001) The molecular biology of the SIR proteins. *Gene* 279(1):1–16.
19. Krogan NJ, et al. (2004) Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. *Proc Natl Acad Sci USA* 101(37):13513–13518.
20. Yang XJ, Seto E (2008) Lysine acetylation: Codified crosstalk with other posttranslational modifications. *Mol Cell* 31(4):449–461.
21. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4(6):1633–1649.
22. Saunders NF, Brinkworth RI, Huber T, Kemp BE, Kobe B (2008) Predikin and PredikinDB: A computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics* 9:245.
23. Shechter D, Dormann HL, Allis CD, Hake SB (2007) Extraction, purification and analysis of histones. *Nat Protoc* 2(6):1445–1457.
24. Puig O, et al. (2001) The tandem affinity purification (TAP) method: A general procedure of protein complex purification. *Methods* 24(3):218–229.
25. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol* 22(8):1035–1036.
26. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14(6):1188–1190.