Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Mining, Analysis and Prediction of Non-Terminal Protein Acetylation Sites

Wyatt T. Clark,[1] Aaron M. Buechlein,[1] A. Keith Dunker,[2] Predrag Radivojac,[1*]
1) School of Informatics, Indiana University, Bloomington, IN 47408
2) Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN 46202

Abstract–**Protein acetylation of lysine residues is a reversible post-translational modification associated with the substitution of a hydrogen in the ε-amino group of the side chain by the acetyl group. Lysine acetylation has been shown to be involved in a variety of cellular functions, most prominently gene regulation and transcription in eukaryotes. However, its full extent, sequence biases and functional repertoire remain largely unknown. In this study, we present a methodology for semi-automated literature mining where new acetylated sites are automatically identified from full scientific papers and then manually verified. We analyzed currently known acetylation sites with respect to sequence, physicochemical and structural properties in the neighborhood of acetylated residues and based on the analysis constructed a prediction model for identification of acetylated lysines. Our literature mining effort resulted in a set of 502 sites currently undocumented in major repositories of post-translationally modified sites (SWISS-PROT, HPRD), while the analysis showed that acetylation sites preferentially occur within intrinsically disordered regions. The accuracy of the prediction model exceeded 80% in the cross-validation experiments.**

I. INTRODUCTION

A. Protein Acetylation

Acetylation of non-terminal lysines is a reversible post-translational modification in which a hydrogen of the ε-NH_2 group of the residue side chain is replaced by the acetyl group (-$COCH_3$). Soon after the discovery of N-terminal acetyl groups in calf thymus histones [1], ε-N-acetyllysine was discovered to exist in these histones as well [2]. In fact, most ε-lysine acetylation research in the past has focused on histones, however, this modification is also known to occur on a variety of other proteins including transcription factors, α-tubulin, nuclear receptors, and HMG proteins [3-7]. The DNA-binding protein HMG1 was the first non-histone protein found to be lysine acetylated [8]. In addition to the transcription and gene regulation, protein acetylation is also known to affect several other functions. Histone acetylation is involved in the regulation of nuclear processes including DNA replication, recombination, and repair, while acetylation of transcription factors affects protein stability, DNA-binding affinity, and nuclear localization [9, 10].

Two mechanisms have been proposed to explain the effects of lysine acetylation on protein function: the 'loss-of-function' and 'gain-of-function' mechanisms [10, 11]. With respect to the 'loss-of-function' mechanism, lysine acetylation neutralizes the positive charge on the lysine side chain. The modification is also thought to reduce the ability of the ε-amino group to be a hydrogen bond donor [10, 11] and thus can influence the protein's ability to interact with DNA, RNA, and other proteins. Concurrent with this, acetylation is most notably known for regulating transcriptional activity [12]. With respect to the 'gain-of-function' mechanism, acetylation of lysine provides a new face for protein interactions. It is known that bromodomains, which are found in many proteins, have the ability to recognize acetyllysines [10, 11].

Protein acetylation is enzyme mediated, with the signal being turned on and off by acetyltransferases and deacetylases. There are six families of Histone Acetyltransferases (HATs). The first family, the GNAT superfamily, is perhaps the best understood. Members of this family are grouped together because they share conserved acetylation-related structural motifs found in the yeast protein Gcn5. Along with Gcn5, other members of GNAT superfamily include P/CAF (P300/CREB binding protein associated factor), Hat1 (HATB), Elp3, and Hpa2 [7, 13]. The second family is the MYST family, which derives its name from its founding members: MOZ, Ybf2/Sas3, Sas2, and Tip60 [14, 15]. Other members of this family include Esa1, MOF, MORF, and Hbo1 [7, 13]. Members of the MYST family also share a portion of the GNAT motif [14, 15]. The third family is p300/CBP which includes p300 and its close homolog CBP (CREB Binding Protein) as members [7, 13]. The fourth family is the Nuclear Receptor Cofactors and contains three members: SRC-1, ACTR, and TIF2 [7, 13]. The fifth and sixth families so far contain only one member each, TAFII250 (TFIID) and TFIIIC, respectively [7, 13]. Roth et al. grouped TAFII250 and TFIIIC into one family called Basal Transcription Factors [7]. Although these acetyltransferases are termed HATs, implying that they acetylate histones, previous studies have shown that they have the ability to acetylate other proteins as well. Thus the term Factor Acetyltransferase (FAT) has been created. P/CAF, p300/CBP, and TAFII250 are among the acetyltransferases that have demonstrated the ability to acetylate proteins other than histones [13].

The first histone deacetylase was discovered in 1996 by Taunton et al. [16]. There are now 18 known deacetylases in humans. They can be grouped into two families, histone deacetylases (HDACs) and the Sir2-like deacetylases. The HDACs can be split into two classes [17]. The first class of HDACs is grouped together due to their high homology with RPD3, a yeast histone deacetylase. Members of this class are HDAC1, HDAC2, HDAC3, HDAC8, and HDAC11 [18-20]. The second class of HDACs is categorized together due to their strong homology with hda1, another yeast histone deace-

*To whom correspondence should be addressed at School of Informatics, Indiana University, 901 E. 10th Street, Bloomington, IN 47408. Email: predrag@indiana.edu

The authors wish it to be known that, in their opinion, Wyatt T. Clark and Aaron M. Buechlein should be regarded as joint First Authors.

1-4244-0710-9/07/$20.00 ©2007 IEEE

tylase. There are five members of the hda1-like class which includes HDAC4, HDAC5, HDAC6, HDAC7, HDAC9, and HDAC10 [18-20]. The Sir2-like family is grouped together because of their shared homology with Sir2, a transcriptional silencer in yeast. The Sir2-like family contains seven members in humans: SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, and SIRT7 [17].

The degree of lysine acetylation of diverse proteins is largely unknown [21], as may be implied from the sheer number of acetyltransferases and deacetylases. However, there are currently fewer than 300 proteins found to be lysine acetylated and several hundreds of known lysine sites within these proteins [9, 13, 22]. To more fully understand the range of regulatory functions of lysine acetylation in biological pathways, information is needed about the diverse protein types and locations of lysine residues that are acetylated.

Various experimental techniques have been employed in the past to study acetylation within proteins. However, these techniques can be both expensive and time-consuming. Our aim is to present the research community with an *in silico* method to quickly and inexpensively elucidate some knowledge about lysine acetylation within proteins. These predictions can then be used for discerning knowledge about biological pathways, protein-protein or protein-nucleic acid interactions involving lysine acetylation.

To our knowledge, two published predictors of lysine acetylation sites have been constructed prior to our work. The AutoMotif Server, created by Plewczynski et al., contains within it an acetylation prediction [23]. These predictions rely on the creation of regular expressions based on experimentally verified acetylation sites within proteins from SWISS-PROT. Recently, Li et al. published PAIL, an internal lysine acetylation predictor utilizing a Bayesian discriminant method [24]. Their predictor was created using a dataset of 89 proteins which included 246 acetylation sites. Predictions were based on the product of the probabilities of the residues surrounding lysine residues.

B. Intrinsically Disorder Proteins and their Prediction

Intrinsically disordered proteins have recently gained significant attention of the research community due to their existence as conformational ensembles and ability to carry out function in ways different from well-established lock-and-key and induced-fit theories [25-27]. Several research groups have constructed predictors of disordered regions with accuracy levels varying between 70% and 80%, measured using a balanced-sample accuracy. Our group has constructed several of these models [28-32] with increasing success over the years. These predictors have shown to be of practical interest not only for the direct identification of disordered proteins or reducing the costs of structural genomics studies, but also as inputs to other problems where the amount of available data has been limited. Among many other (for review see [33]), two such studies were related to prediction of protein modification sites, where predictions of disordered regions directly helped increased accuracy of phosphorylation [34] and methylation sites [35]. Our work in this study features the VSL2 predictor [32] as being especially important for the prediction of protein acetylation sites.

C. Outline of the Study

The goal of our approach was to enlarge databases of documented acetylation sites using semi-automated approaches, then study their sequence and functional properties. Finally, we construct a predictor that will be useful in future studies of protein acetylation as well as protein function in general. In Section II, we present our methodology for extracting lysine acetylation sites from scientific literature and then elaborate on the approach for data analysis and prediction. Section III provides experimental details and results of our study. Section IV summarizes this paper.

II. Materials and Methods

A. Mining Scientific Literature

We developed an algorithm for identifying protein acetylation sites from the scientific literature. The algorithm searches full papers and uses simple and intuitive techniques based on regular expressions. The task of our algorithm is to identify a set of scientific papers that are most likely to describe experimentally determined acetylation sites. We require that each article be prioritized so as to maximize the number of newly extracted acetylation sites. These papers are subsequently read in order to provide us with confident locations of acetylation sites. Thus, all newly identified acetylation sites are verified manually and associated with literature that represents traceable evidence of experimental support.

Given an article s that can be represented as a string of lowercase symbols, we first locate all words a containing the substring "acetylat", $A = \{a_1, a_2, \ldots, a_{|A|}\}$, and subsequently all strings r that can be described by a regular expression of the form

$$(\text{k} \vee \text{lys} \vee \text{lysine})\ (\mathit{space})^*\ (\mathit{digit})^+\ (\mathit{space})^* \qquad (1)$$

where $\mathit{digit} \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Symbol * represents the Kleene star and $x^+ = x^* - \varepsilon$, where ε is an empty string and x is an arbitrary non-empty string. We denote a set of strings described by the regular expression from (1) as $R = \{r_1, r_2, \ldots, r_{|R|}\}$.

Given sets A and R, the priority score S for article s can be calculated using the following expression

$$S = \sum_{i=1}^{|R|} score(r_i, A),$$

where

$$score(r_i, A) = f(|\, position(r_i) - position(a_k)\,|)$$

and

$$k = \arg\min_{j=1\ldots m} |\, position(r_i) - position(a_j)\,|.$$

To convert a set of distances between r_i and a_j into a score that is inversely proportional to the distance between the starting positions of strings, $position(r_i)$ and $position(a_k)$, we are using the following functional form

$$f(x) = c \cdot e^{-d \cdot x}, \qquad (2)$$

where x is the distance in characters between r_i and a_j, while c and d are positive constants. In other words, for each word $r \in R$, word $a \in A$ closest to r is identified and their distance is calculated. The overall score S for article s is calculated as cumulatively over all regular expressions in R. Constants c and d are determined empirically so as to provide easily interpretable scores.

The above-mentioned procedure represents the basis of our algorithm, however it can be susceptible to large repetitions of acetylation sites within papers or sites that are simply more frequently studied and thus repetitively mentioned by the authors. This is frequently the case with histones. Therefore, in a practical setting, this procedure can be modified to exclude certain other sites or proteins (Section III).

B. Datasets

Proteins used for analysis of acetylation sites and predictor construction were collected from three sources: (i) SWISS-PROT, (ii) Human Protein Reference Database (HPRD), and (iii) scientific literature. SWISS-PROT is one of the best annotated databases storing protein sequence, taxonomic and functional annotation [36]. Acetylated residues were extracted by examining the MOD_RES fields for each protein. All records containing N6-acetyllysine were subsequently collected, however, sites containing annotation “by similarity,” “probable” and “potential” were excluded. HPRD provides a large variety of information on human proteins, including the post-translational modification sites, protein-protein interactions and protein function [37]. Finally, articles found using our text mining approaches comprised the third source of acetylatable lysines. In determining whether a particular site can be acetylated we followed author’s statements and have not interpreted their data or assigned confidence of annotation.

A significant problem in any approach predicting protein functional annotation is confident determination of negative examples, i.e. non-acetylated sites. In the case of acetylation, many sites were shown to be acetylated by various experimental techniques, however given isolated *in vitro* experiments it is often unclear whether the remaining sites in the same protein can undergo acetylation. Therefore, negative sites are expected to contain a significantly larger fraction of incorrect class labels, even in the mass spectrometry experiments where cellular conditions, proteomics platform or identification software may play an important role. In total, we collected a set of 293 proteins with 683 positive sites and 7,714 negative sites.

C. Redundancy Removal

To prevent learning and accuracy estimation on sites with very similar neighborhoods, positive and negative sites were filtered. Here, a “site” is considered to be a sequence fragment consisting of 25 residues around the central Lys. Negative fragments were first sifted against the positive fragments, i.e. each negative fragment whose sequence neighborhood was more than 40% similar to any of the positive fragments was removed from the set. There are two reasons for this step: (i) identical or very similar proteins could be used in different experiments, potentially involving different acetyltransferases, and thus resulting in different sites being acetylated; and (ii) two proteins having high sequence similarity are expected to have similar functions, thus unlabeled fragments are removed form the dataset since they could be considered positives “by similarity”. In the second filtering step, no two sites in any of the sets were allowed to have sequence identity greater than 40%. This step enables us to remove clusters of data points with high density as potential artifacts arising from certain proteins being intensively studied by the experimental community. With the same rationale as above, if two sites were characterized by similar amino acid neighborhood, they can be considered positive/negative based on similarity. Finally, redundancy removal prevents overestimation of the prediction accuracy.

D. Data Representation

Each lysine in our set of proteins is represented as a labeled data point in vector space. Features are calculated using a set of concentric windows flanking the Lys residues and can be grouped into three categories: (i) amino acid content and physicochemical properties, (ii) predicted protein properties, and (iii) evolutionary conservation. More formally, given a protein sequence $P = r_1r_2...r_n$, where r_i represents a residue at position i, and a window of length w, a collection of features is calculated for the following subsequence $r_{[i-h,\ i+h]}$, where h = floor(w/2), is the half window.

Features based on amino acid content and physicochemical properties. The first set of amino acid features was derived for windows $w \in \{3, 7, 11, 21\}$ by calculating 20 amino acid compositions. In addition, we calculated sequence complexity [38], β-entropy [39], charge and aromatic content for the fragments. Positive charge within the window was calculated by finding the number of Lys and Arg residues, while negative charge was calculated by finding the number of Asp and Glu residues. By looking at the count of Phe, Tyr and Trp residues we calculated the aromatic content within the window.

Features based on predicted properties. We utilized several predictors that output residue-based scores and created features over windows $w \in \{1, 7, 11, 21\}$. For each protein we predicted intrinsic disorder using VL2 [29], VL3 [30], and VSL2 [32] models. In addition, we used a flexibility predictor by Vihinen et al. [40], predictions of the B-factor [41], hydrophobic moment values [42], a predictor of phosphorylation [34], and the charge-hydropathy ratio [43].

Feature based on evolutionary conservation. We estimated the evolutionary conservation of residues over windows $w \in \{1, 3, 11, 21\}$ by first creating a position specific scoring matrix (PSSM) for each sequence and then averaging PSSM values for each residue over w values. PSSMs were calculated using PSI-BLAST [44] against GenBank. Each PSSM-based feature provides a measure of conservation of a given residue within a window. A window w = 1 provides a conservation score at a given Lys residue.

E. Predictor Construction and Evaluation

Once each sequence fragment was mapped to a vector space, a predictor was constructed in the following two steps: (i) data preprocessing and (ii) model training. In the data preprocessing step, we performed a set of three transformations on our data. Features were first removed using a t-test based feature filtering. Those whose p-values were above threshold t_{fs} were removed from the further steps, where $t_{fs} \in \{1, 0.1, 0.01, 0.001\}$. Z-score normalization was then performed on the remaining features, followed by the principal component analysis. Support vector machines (SVMs) were trained on the preprocessed data using SVM^{light} software [45]. We explored both polynomial and Gaussian kernels using several values for the parameters (see Section III). SVM^{light} provides a solution to the dual problem in which one seeks to create a hyperplane that maximizes the margin of separation between positive and negative data points. This problem is formalized as a maximization of the following equation

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j l_i l_j K(\mathbf{x}_i, \mathbf{x}_j),$$

subject to constraints $\sum_{i=1}^{N} \alpha_i l_i = 0$ and $\alpha_i \geq 0$ for i = 1, 2, … N, where $\mathbf{x}_i$ is i-th of the N training data points, $l_i \in \{-1, 1\}$ is the class label, and K is the kernel function. Using a trained model, test data point $\mathbf{x}$ is classified by comparing the score

$$s_i = \sum_{i=1}^{N} \alpha_i l_i K(\mathbf{x}, \mathbf{x}_j)$$

to zero. A value of the positive (negative) score provides the likelihood that the data point belongs to the positive (negative) set.

Performance estimation was carried out using a per-protein leave-one-out strategy. In each step, one protein was selected into a test set, while all the remaining proteins constituted a training set. This approach allows us to model a realistic situation in which a new protein is provided to the predictor for estimation of potential acetylation sites.

Given a vector of scores $\mathbf{s} = [s_1, s_2 \ldots s_n]$ for each test protein, where $s_i \in (-\infty, +\infty)$, and vector $\mathbf{l}$ representing a set of true class labels $\mathbf{l} = [l_1, l_2 \ldots l_n]$, where $l \in \{-1, 1\}$, we calculated the sensitivity (sn), specificity (sp), precision (pr), accuracy (acc) and area under the receiver operating characteristic curve (AUC) as measures of predictor performance. Setting a threshold $t = 0$, sensitivity for each test protein was calculated as a ratio between $|\{i \mid s_i \geq t \wedge l_i = 1\}|$ and $|\{i \mid l_i = 1\}|$. Conversely, specificity is calculated as a ratio between $|\{i \mid s_i < t \wedge l_i = -1\}|$ and $|\{i \mid l_i = -1\}|$, while precision is obtained as $|\{i \mid s_i \geq t \wedge l_i = 1\}| \,/\, |\{i \mid s_i \geq t\}|$. Accuracy is calculated as the average of sensitivity and specificity, since the dataset was heavily imbalanced. Finally, AUC was calculated as an area under the curve representing sn as a function of $1 - sp$. This curve is generated when the threshold t is gradually shifted between the minimal and maximal prediction scores. Final accuracy is measured as a mean over all individual accuracies per protein, while the area under the curve was calculated over all test examples together, after the final leave-one-out step.

F. Reliability Prediction

The rationale for building a reliability predictor is to ensure valid statistical inference since models of protein acetylation may be trained and evaluated on a biased sample of protein data. To identify differences in data distributions, we constructed a *contrast classifier* [46] to distinguish between labeled examples, both positive and negative, and examples randomly selected from SWISS-PROT. Using the same set of predictor-building steps, with the exception of evolutionary information, each data point can be assigned a likelihood score of the class membership. There, a high score towards the class of unlabeled data indicates smaller reliability of statistical inference. A desired outcome for a reliability model is random prediction accuracy that would indicate that the labeled examples are good representatives of the unlabeled data.

III. RESULTS

A. Literature Mining

In order to build an adequate dataset for lysine acetylation sites, it was necessary to perform text mining on the current literature available and manually verify lysine acetylation sites. To date, there are 19,520 articles in PubMed that are returned when a search is performed for "acetylat*". A similar search was performed on more than 30 journal sites and the resulting articles were downloaded. Nature and Science Direct provided the largest number of articles contributing 2,147 articles and 1,519 articles, respectively at the time of the search. In order to provide a level of automation and provide insight as to which articles would be more fruitful, they were ranked using the scoring function presented in Section II.A. The constants c and d from expression (2) were assigned values of 10 and 0.005, respectively, so as to provide scores for articles that held greater meaning in distinguishing site-rich articles from articles that did not provide as many sites. Ranking of the articles was performed several times using various combinations of values for c and d until functioning values were determined.

The highest ranking article was from the Journal of Biological Chemistry, had a score of 151.9 and contained 10 unique acetylation sites. Later, another step to aid in site discovery efficiency was to try to exclude articles discussing histones. If the word "histone" or "H3" or "H4" were seen in combination greater than 120 times, the article was excluded from the ranking. The value of 120 was chosen to try to minimize the number of articles excluded that would have provided non-histone acetylation sites. With this ranking, the highest scoring article received a score of 116.4 and contained 8 unique acetylation sites. This article was obtained from the Biochemistry Journal. Looking at the top 25, the Journal of Biological Chemistry provided the most articles with 5, followed by Nature, Biochemistry, and Molecular Cell which each provided 4 articles. Cell provided 3 articles, Nucleic Acids Research provided 2 articles, the Journal of Proteome Research provided 1 article, International Journal of Mass Spectrometry provided 1 article, and Gene provided 1 article.

Ranking provided very useful information for prioritizing the articles; however, it was not perfect. In general, articles with scores greater than 30 showed potential for providing at least one site, but as scores approached 30, articles without acetylation sites began to appear. A common reason was that the regular expression from equation (1) was not selective enough and included other information, e.g. protein names. Also, due to how the ranking was performed, we were able to count sites only once if the sites were discussed in a consistent manner throughout the article. However, if the author referred to the sites in more than one way, the sites were potentially counted more than once and could have led to the article having a falsely high score. Articles also potentially fell in the ranking if the acetylation sites were not listed in the text of the article and were only provided in figure or as supplementary data. An example of such an occurrence was the 144th ranked article in the rankings that included articles referring to histones. This article provided 11 unique acetylation sites; however, its score was only 41.7. The article providing the largest amount of lysine acetylation sites was "Substrate and functional diversity of lysine acetylation revealed by a proteomics survey" by Kim et al. [22] and was obtained through Science Direct. This article provided 386 unique sites in 197 different proteins. In our ranking system this article received a score of 72.0 and was ranked 35th if articles containing histones were considered and 8th if histone containing articles were excluded. The vast majority of these sites were provided as supplementary data on the article website. Thus, this article received a falsely lower score in our ranking system.

Overall, our literature search with manual curation detected 541 acetylation sites with traceable evidence. In Figure 1 we compare this number of sites to the ones currently available in SWISS-PROT and HPRD as of August 2006. Since SWISS-PROT and HPRD provide additional curation in identifying acetylation sites, it may not be very surprising that our dataset was the largest. However, it is surprising that the number of overlapping sites between SWISS-PROT, HPRD and our literature search are 94% orthogonal. All the sites identified in this study are freely available upon request.

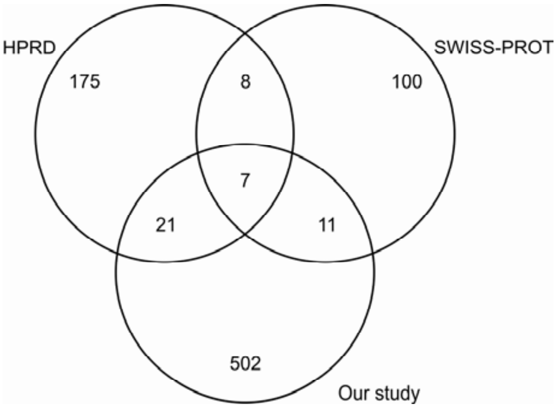

Figure 1. A comparison of the number of acetylation sites contained in HPRD, SWISS-PROT and those detected in our study.

B. Functional Analysis of Acetylated Proteins

To study functional annotation of acetylated proteins we searched the Gene Ontology (GO) database [47]. We used a non-redundant set of proteins in which sequence identity was <80% over the entire length. From the set of 290 acetylated proteins, 77 were found to be associated with a variety of GO terms. In the molecular function category, 112 functional terms were identified. The most prominent ones were transcription-related activities ("transcription regulator activity", "transcription factor activity", "transcription factor binding", "transcription cofactor activity", "transcription coactivator activity", "transcriptional activator activity") and catalytic activity ("transferase activity", "hydrolase activity" oxidoreductase activity", "acetyltransferase activity", "acyltransferase activity", "histone acetyltransferase activity"). However, a variety of other terms has been detected, including "unfolded protein binding", "DNA binding", "damaged DNA binding" etc. In the biological process category, nearly 300 different terms were observed, among which the ones involving the largest number of proteins were: "metabolism", "response to stimulus", "regulation of biological process", "regulation of physiological process", "transcription", "regulation of metabolism", "regulation of transcription", etc. Finally, a set of 53 terms from the category cellular component has been found. Most proteins were associated with terms "intracellular", "organelle", "membrane-bound organelle", "cytoplasm", "mitochondrion", "nucleus" and "membrane".

This analysis possibly implicates protein acetylation in a variety of regulatory functions, but also indicates that the dataset used in this study is diverse enough to provide valid statistical inference.

C. Statistical Analysis of Flanking Regions

Position-specific amino acid preferences were analyzed using Two Sample Logo software [48], located at http://www.twosamplelogo.org. Two Sample Logo calculates statistically significant differences between two sets of multiple sequence alignments. Residues whose p-value was below 0.05 are plotted in Figure 2, with the size of each residue proportional to the difference between the samples.

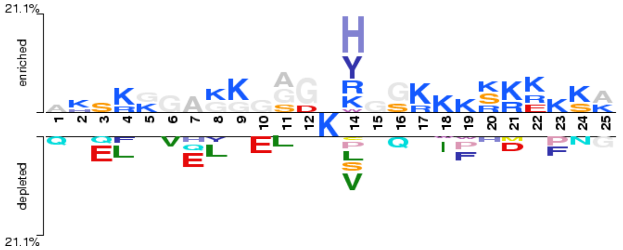

Figure 2. Two Sample Logo of the lysine acetylation sites. The upper panel represents residues enriched in the set of acetylated lysines, the lower panel represents residues depleted in the set of acetylated lysines, while the middle panel displays consensus residues between acetylated lysine sites and all remaining sites from the same set of proteins.

The upper panel of Figure 2 displays residues that are enriched in the positive sample, the lower panel displays those depleted in the positive sample, while the middle part displays the consensus residues. The most striking difference between positive and negative residues is the presence of positively charged residues around the acetylated lysine. Fifteen out of 24 positions are enriched in lysine, while 6 are enriched in arginine. The most interesting signature of acetylated lysines is the presence of histidine and tyrosine at position +1, and this signature is discussed by Kim et al. [22].

Previous studies have already identified clustering of lysines around acetylation sites. In particular, four sequence motifs have been reported to date: (i) KXXK [49-51], (ii) RXKK [50], (iii) GKXXP [7, 9] and (iii) H4 motif [52]. We applied motif search on the set of proteins in our dataset and identified 68 KXXK, 33 RXKK, 31 GKXXP and 4 H4 motifs. However, only 10 KXXK, 4 RXKK, 5 GKXXP and 3 H4 motifs were actually known acetylation sites, giving potentially high false positive rates.

Statistical analysis of non-position specific features is presented in Table I. For each feature described in Section II, we calculated the p-value of the null hypothesis that positive and negative examples were generated by the same distribution. For this purpose we assumed a normal distribution over all data points and applied the t-test. The top 10 positively and negatively correlated features, together with their p-values, are listed in Table I. We note here that each property is displayed only once, for the best performing window size (in parentheses).

TABLE I
TOP 10 FEATURES DESCRIBING LYSINE ACETYLATION. NUMBERS IN PARENTHESES INDICATE WINDOW SIZES. SEE TEXT FOR DETAILED EXPLANATIONS.

Positive correlation		*Negative correlation*	
Feature	*p-value*	*Feature*	*p-value*
VSL2B (21)	$9.9 \cdot 10^{-73}$	PSSM, L (21)	$1.4 \cdot 10^{-17}$
net charge (21)	$2.9 \cdot 10^{-19}$	β-entropy (21)	$2.9 \cdot 10^{-13}$
residue G (21)	$6.3 \cdot 10^{-19}$	PSSM, I (21)	$1.4 \cdot 10^{-11}$
B-factor (21)	$7.7 \cdot 10^{-19}$	entropy (21)	$2.0 \cdot 10^{-10}$
residue K (21)	$1.6 \cdot 10^{-18}$	PSSM, M (21)	$4.9 \cdot 10^{-10}$
residue H (3)	$1.0 \cdot 10^{-17}$	PSSM, F (21)	$1.8 \cdot 10^{-10}$
VL2-C (21)	$8.2 \cdot 10^{-14}$	residue L (21)	$2.6 \cdot 10^{-6}$
DisPhos (21)	$2.7 \cdot 10^{-14}$	PSSM, V (21)	$3.1 \cdot 10^{-6}$
Vihinen (21)	$4.8 \cdot 10^{-14}$	PSSM, Y (21)	$4.2 \cdot 10^{-6}$
VL3 (21)	$4.7 \cdot 10^{-13}$	residue I (21)	$1.4 \cdot 10^{-5}$

Apart from the enrichment of Gly, Lys and His around the acetylated sites, the most striking positive correlation was achieved by the VSL2B disorder predictor [32]. VSL2B is the fast version of the VSL1 and VSL2 predictors which achieved excellent performance during CASP6 experiment [31]. This feature is orders of magnitude more informative than any other individual feature and greatly increased prediction accuracy. The remaining positively correlated features were all related to high content of intrinsic disorder. Interestingly, the conservation of acetylated lysines was selected as the 11th best feature with the p-value of $1.7 \cdot 10^{-12}$ thus indicating functional similarity across the eukaryotic kingdom. Negatively correlated features are related to the presence of hydrophobic residues (Leu, Ile) and their evolutionary conservation. Thus, less conserved Leu, Ile, Met, Phe, Val, and Tyr in the flanking regions are a signature of acetylated lysines.

D. Acetylation and Intrinsic Disorder

Based on several examples from the literature, the relationship between several protein post-translational modifications and intrinsically disordered regions was suggested previously [27, 53]. Here, we investigate this relationship thoroughly for all acetylation sites and compare them with other lysine sites assumed not to be acetylated. Since the structural form of most of the acetylated proteins is unknown, we used currently most accurate predictors of intrinsically disordered regions [31, 32]. The average VSL2B score for acetylated sites was 0.58±0.01, while the score for the remaining sites was 0.34±0.01. The sites were also classified into near-terminal, if they were 12 residues or less away from the N- or C-terminus, or internal. The corresponding disorder scores for the near-terminal positive and negative sites were 0.78±0.04 and 0.53±0.01, respectively. Similarly, the scores for the internal positive and negative sites were 0.56±0.01 and 0.33±0.01, respectively.

E. Predictor Evaluation

As mentioned in Section II, the acetylation predictor was evaluated using per-protein leave-one-out strategy. However, within each protein, only non-redundant sites to the training set were considered. Several learning parameters were considered for training SVMs: for the polynomial kernel we varied degree of the polynomial between 1 and 3, while for the Gaussian kernel, σ was assigned values from $\{10^{-2}, 10^{-4}, 10^{-6}\}$.

The results shown in Tables II-II indicate that lysine acetylation is predictable with relatively high accuracy, matching or exceeding protein phosphorylation or methylation. Non-linear models with Gaussian kernel outperformed polynomial kernels. Figure 3 shows ROC curves for the best performing models from each class. We believe that these findings are extendable to other protein sequences since the accuracy of the reliability model did not exceed 58%. This indicates that the unlabeled lysines were generally not distinguishable from those present in our dataset.

TABLE II
CLASSIFICATION ACCURACY [%] FOR THE PREDICTION OF LYSINE ACETYLATION SITES USING SVM WITH POLYNOMIAL KERNEL; *sn* – SENSITIVITY, *pr* – PRECISION, *sp* – SPECIFICITY, *acc* = (*sn* + *sp*) / 2 – ACCURACY, *AUC* – AREA UNDER THE ROC CURVE.

degree	*Acetylation, polynomial kernel*				
	sn	*sp*	*pr*	*acc*	*AUC*
$p = 1$	88.0	71.1	49.8	79.5	84.0
$p = 2$	67.2	85.4	55.5	76.3	86.0
$p = 3$	88.2	70.8	49.4	79.5	83.8

TABLE III
CLASSIFICATION ACCURACY [%] FOR THE PREDICTION OF LYSINE ACETYLATION SITES USING SVM WITH GAUSSIAN KERNEL; *sn* – SENSITIVITY, *sp* – SPECIFICITY, *pr* – PRECISION, *acc* = (*sn* + *sp*) / 2 – ACCURACY, *AUC* – AREA UNDER THE ROC CURVE.

sigma	*Acetylation, Gaussian kernel*				
	sn	*sp*	*pr*	*acc*	*AUC*
$\sigma = 10^{-2}$	75.4	85.0	57.8	80.2	89.9
$\sigma = 10^{-4}$	90.8	74.0	52.1	82.4	86.8
$\sigma = 10^{-6}$	88.2	70.8	49.4	79.5	83.8

We compared classification accuracy of our system with AutoMotif server and PAIL. For these, we used only the newest sites [22] on which these servers were likely not trained. Out of 382 positive and 5,395 negative sites, AutoMotif server predicts 18 sites as positives and none of the negative sites as positives. Similarly, PAIL predicted 204 positives correctly, however it also predicted 2,633 negatives as positive sites. These results were obtained using high stringency threshold, while similar ones can be obtained using medium stringency.

IV. DISCUSSION

While lysine acetylation is emerging as an important and widespread post-translational modification, its full functional role is still incompletely understood. The goal of our study was to build methodology and collect a large number of experimentally determined acetylation sites, analyze their properties systematically and construct a prediction model that will be useful for applications in bioinformatics.

Our text mining algorithm is simple and based on regular

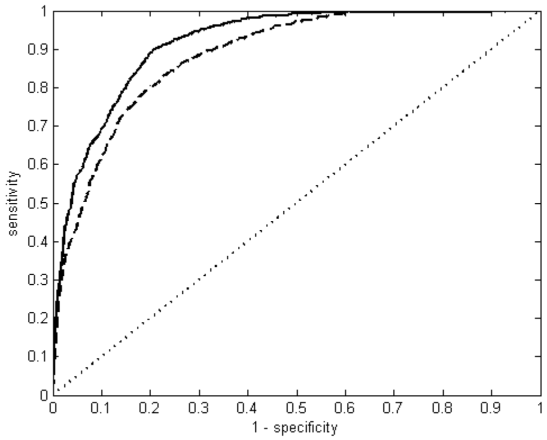

Figure 3. Receiver operating characteristic curves for (i) support vector machine with Gaussian kernel and $\sigma = 0.01$, - solid curve; (ii) support vector machine with polynomial kernel and $p = 2$, - dashed curve; (iii) random model – dotted line. The area under the curve for the best model is $AUC = 86.8\%$.

expressions for identifying candidate acetylation sites. Each candidate site was subsequently verified manually, and the author statements from the corresponding publications are stored. The acetylation sites found using this methodology were used for the analysis of sequence, physicochemical, structural and evolutionary properties of acetylation sites. This analysis was conducted by comparing them to the "background" sites, i.e. lysines from the same set of proteins that were not identified to be acetylated. Prominent properties of acetylation neighborhoods were high positive charge, primarily the content of Lys and Arg, the presence of Gly, and also presence of His or Tyr at position +1. Interestingly, the presence of Gly has also been identified as a signature of methylated residues.

While the findings given above are consistent with those from the literature, our analysis adds new information that lysine acetylation is highly correlated with protein intrinsic disorder. This analysis has been made by means of predictors. Interestingly, out of several disorder prediction models developed over the years by the Dunker-Obradovic group and collaborators, the role of VSL2B was far more important than any other model. VSL2 is the newest and most accurate predictor and its usefulness in protein acetylation provides support that improving the accuracy of disorder prediction can have positive consequences on prediction of residue-based function. Without the use of the VSL2 predictions as a feature, the prediction accuracy drops by about 10 percentage points.

The prediction accuracy achieved by our final models has been achieved after limited experimentation with training parameters and could possibly be improved, even for the present dataset. However, at this stage, we were more concerned with developing robust models that would have good generalization properties in a real life setting.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help of William J. Sullivan for communicating to us the newest data regarding acetylation sites. Sourav Roy helped us at the initial stage of data collection. Support was provided by NIH Grant 1 R01 LM007688 awarded to AKD and Zoran Obradovic, and also by the Indiana Genomics Intitiative (INGEN), funded in part by the Lilly Endowment.

REFERENCES

1. Phillips, D.M. (1963) The presence of acetyl groups of histones. *Biochem J*. **87**, 258-263.
2. Vidali, G., E.L. Gershey, and V.G. Allfrey (1968) Chemical studies of histone acetylation. The distribution of epsilon-N-acetyllysine in calf thymus histones. *J Biol Chem*. **243**(24), 6361-6366.
3. Imhof, A., X.J. Yang, V.V. Ogryzko, Y. Nakatani, A.P. Wolffe, and H. Ge (1997) Acetylation of general transcription factors by histone acetyltransferases. *Curr Biol*. **7**(9), 689-692.
4. MacRae, T.H. (1997) Tubulin post-translational modifications--enzymes and their mechanisms of action. *Eur J Biochem*. **244**(2), 265-278.
5. Berger, S.L. (1999) Gene activation by histone and factor acetyltransferases. *Curr Opin Cell Biol*. **11**(3), 336-341.
6. Bannister, A.J., E.A. Miska, D. Gorlich, and T. Kouzarides (2000) Acetylation of importin-alpha nuclear import factors by CBP/p300. *Curr Biol*. **10**(8), 467-470.
7. Roth, S.Y., J.M. Denu, and C.D. Allis (2001) Histone acetyltransferases. *Annu Rev Biochem*. **70**, 81-120.
8. Sterner, R., G. Vidali, and V.G. Allfrey (1979) Studies of acetylation and deacetylation in high mobility group proteins. Identification of the sites of acetylation in HMG-1. *J Biol Chem*. **254**(22), 11577-11583.
9. Kouzarides, T. (2000) Acetylation: a regulatory modification to rival phosphorylation? *Embo J*. **19**(6), 1176-1179.
10. Yang, X.J. (2004) Lysine acetylation and the bromodomain: a new partnership for signaling. *Bioessays*. **26**(10), 1076-1087.
11. Yang, X.J. (2004) The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res*. **32**(3), 959-976.
12. Allfrey, V.G., R. Faulkner, and A.E. Mirsky (1964) Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci U S A*. **51**, 786-794.
13. Sterner, D.E. and S.L. Berger (2000) Acetylation of histones and transcription-related factors. *Microbiol Mol Biol Rev*. **64**(2), 435-459.
14. Borrow, J., V.P. Stanton, Jr., J.M. Andresen, R. Becher, F.G. Behm, R.S. Chaganti, C.I. Civin, C. Disteche, I. Dube, A.M. Frischauf, D. Horsman, F. Mitelman, S. Volinia, A.E. Watmore, and D.E. Housman (1996) The translocation t(8;16)(p11;p13) of acute myeloid leukaemia fuses a putative acetyltransferase to the CREB-binding protein. *Nat Genet*. **14**(1), 33-41.
15. Reifsnyder, C., J. Lowell, A. Clarke, and L. Pillus (1996) Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases. *Nat Genet*. **14**(1), 42-49.
16. Taunton, J., C.A. Hassig, and S.L. Schreiber (1996) A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science*. **272**(5260), 408-411.
17. Grozinger, C.M., C.A. Hassig, and S.L. Schreiber (1999) Three proteins define a class of human histone deacetylases related to yeast Hda1p. *Proc Natl Acad Sci U S A*. **96**(9), 4868-4873.
18. Gray, S.G. and T.J. Ekstrom (2001) The human histone deacetylase family. *Exp Cell Res*. **262**(2), 75-83.
19. Grozinger, C.M. and S.L. Schreiber (2002) Deacetylase enzymes: biological functions and the use of small-molecule inhibitors. *Chem Biol*. **9**(1), 3-16.
20. Verdin, E., F. Dequiedt, and H.G. Kasler (2003) Class II histone

deacetylases: versatile regulators. *Trends Genet*. **19**(5), 286-293.

21. Walsh, C.T., Posttranslational modification of proteins: expanding nature's inventory. 2006, Englewood, CO: Roberts and Company Publishers.
22. Kim, S.C., R. Sprung, Y. Chen, Y. Xu, H. Ball, J. Pei, T. Cheng, Y. Kho, H. Xiao, L. Xiao, N.V. Grishin, M. White, X.J. Yang, and Y. Zhao (2006) Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Mol Cell*. **23**(4), 607-618.
23. Plewczynski, D., A. Tkacz, L.S. Wyrwicz, and L. Rychlewski (2005) AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics*. **21**(10), 2525-2527.
24. Li, A., Y. Xue, C. Jin, M. Wang, and X. Yao (2006) Prediction of N(epsilon)-acetylation on internal lysines implemented in Bayesian Discriminant Method. *Biochem Biophys Res Commun*. **350**(4), 818-824.
25. Dyson, H.J. and P.E. Wright (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. **6**(3), 197-208.
26. Fink, A.L. (2005) Natively unfolded proteins. *Curr Opin Struct Biol*. **15**(1), 35-41.
27. Dunker, A.K., J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, and Z. Obradovic (2001) Intrinsically disordered protein. *J. Mol. Graph. Model*. **19**(1), 26-59.
28. Romero, P., Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, and A.K. Dunker (2001) Sequence complexity of disordered protein. *Proteins*. **42**, 38-48.
29. Vucetic, S., C.J. Brown, A.K. Dunker, and Z. Obradovic (2003) Flavors of protein disorder. *Proteins*. **52**, 573-584.
30. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, C.J. Brown, and A.K. Dunker (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*. **53**(S6), 566-572.
31. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, and A.K. Dunker (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*. **61 Suppl 7**, 176-182.
32. Peng, K., P. Radivojac, S. Vucetic, A.K. Dunker, and Z. Obradovic (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. **7**(1), 208.
33. Daughdrill, G.W., G.J. Pielak, V.N. Uversky, M.S. Cortese, and A.K. Dunker, *Natively disordered protein*, in *Protein Folding Handbook*, J. Buchner and T. Kiefhaber, Editors. 2005, Wiley-VCH: Verlag GmbH & Co. KGaA: Weinheim. p. 271-353.
34. Iakoucheva, L.M., P. Radivojac, C.J. Brown, T.R. O'Connor, J.G. Sikes, Z. Obradovic, and A.K. Dunker (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*. **32**(3), 1037-1049.
35. Daily, K.M., P. Radivojac, and A.K. Dunker. (2005) Intrinsic disorder and protein modifications: building an SVM predictor for methylation. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005* 475-481 San Diego, California, U.S.A.
36. Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*. **33 Database Issue**, D154-159.
37. Peri, S., J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T.K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H.N. Shivashankar, B.P. Rashmi, M.A. Ramya, Z. Zhao, K.N. Chandrika, N. Padma, H.C. Harsha, A.J. Yatish, M.P. Kavitha, M. Menezes, D.R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S.K. Anand, V. Madavan, A. Joseph, G.W. Wong, W.P. Schiemann, S.N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G.C. Blobe, C.V. Dang, J.G. Garcia, J. Pevsner, O.N. Jensen, P. Roepstorff, K.S. Deshpande, A.M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*. **13**(10), 2363-2371.
38. Wootton, J.C. and S. Federhen (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*. **266**, 554-571.
39. Daróczy, Z. (1970) Generalized information functions. *Information and Control*. **16**, 36-51.
40. Vihinen, M., E. Torkkila, and P. Riikonen (1994) Accuracy of protein flexibility predictions. *Proteins*. **19**, 141-149.
41. Radivojac, P., Z. Obradovic, D.K. Smith, G. Zhu, S. Vucetic, C.J. Brown, J.D. Lawson, and A.K. Dunker (2004) Protein flexibility and intrinsic disorder. *Protein Science*. **13**(1), 71-80.
42. Eisenberg, D., R.M. Weiss, and T.C. Terwilliger (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U. S. A*. **81**, 140-144.
43. Uversky, V.N. (2002) What does it mean to be natively unfolded? *Eur. J. Biochem*. **269**(1), 2-12.
44. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*. **25**, 3389-3402.
45. Joachims, T., Learning to classify text using support vector machines: methods, theory, and algorithms. 2002: Kluwer Academic Publishers.
46. Peng, K., S. Vucetic, B. Han, X. Xie, and Z. Obradovic. (2003) Exploiting unlabeled data for improving accuracy of predictive data mining. *Third IEEE Int'l Conf. on Data Mining* 267-274 Melbourne, FL.
47. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**(1), 25-29.
48. Vacic, V., L.M. Iakoucheva, and P. Radivojac (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*. **22**(12), 1536-1537.
49. Gu, W. and R.G. Roeder (1997) Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell*. **90**(4), 595-606.
50. Hung, H.L., J. Lau, A.Y. Kim, M.J. Weiss, and G.A. Blobel (1999) CREB-Binding Protein Acetylates Hematopoietic Transcription Factor GATA-1 at Functionally Important Sites. *Molecular and Cellular Biology*. **19**(5), 3496-3505.
51. Fu, M., M. Rao, C. Wang, T. Sakamaki, J. Wang, D. Di Vizio, X. Zhang, C. Albanese, S. Balk, and C. Chang (2003) Acetylation of Androgen Receptor Enhances Coactivator Binding and Promotes Prostate Cancer Cell Growth. *Molecular and Cellular Biology*. **23**(23), 8563-8575.
52. Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, and C.J. Sigrist (2006) The PROSITE database. *Nucleic Acids Res*. **34**(Database issue), D227-230.
53. Dunker, A.K., C.J. Brown, J.D. Lawson, L.M. Iakoucheva, and Z. Obradovic (2002) Intrinsic disorder and protein function. *Biochemistry*. **41**(21), 6573-6582.