# Supporting Information

## ProAcePred: prokaryote lysine acetylation sites prediction based on elastic net feature optimization

**Guodong Chen[1], Man Cao[1], Kun Luo[1], Lina Wang[2], Pingping Wen[2], Shaoping Shi[1*]**

[1]Department of Mathematics and Numerical Simulation and High-Performance Computing Laboratory, School of Sciences, Nanchang University, Nanchang, 330031, China.

[2]School of Chemistry, Nanchang University, Nanchang, 330031, China.

*To whom correspondence should be addressed.

## Table of Contents

**Fig. S2.** Comparison of EBGW between acetylation and non-acetylation. The vertical axis represents the log_2 ratio of average EBGW values between acetylation and non-acetylation. The horizontal axis represents the three binary sequences.

**Fig. S3.** Average accessible surface area (AASA) value of residues surrounding acetylation sites and non-acetylation sites (except of center position) among nine species.

**Fig. S4.** The ROC curves and AUC values of 10-fold cross-validations (CV) of the independent test set for nine species.

# 1. Supplementary Experimental Text

**Text S1.Determination of the Sliding Window Sizes**

Since different sliding windows may have distinct prediction performances, optimization of the sliding window sizes is required for selecting features and training models. On the one hand, if the sliding window was too long, a large amount of redundant information would be included. On the other hand, if the sliding window was too short, a lot of valuable information would lose. Thus, we took into account the window size varied from 11 to 41.

In this study, we used the predicted KNN score features accuracy (Acc) as index to evaluate the performance of the sliding windows with different sizes. Support vector machine (SVM) classifier and 10-fold cross-validation were carried out to build model and select feature based on each sliding window size.

For all of the positive and negative sets in nine species, we optimized the final window sizes through training KNN features. In order to reduce the amount of calculation, according to the bacterial morphological classification, eight kinds of bacteria are divided into vibrio and bacillus. Bacillus is composition of E.coli, S.typhimurium, B.subtilis, M.tuberculosis, C.glutamicum, E.amylovora and G.kaustophilus, vibrio only contains of V.parahemolvticus. In this part, K was chosen to be 1.5%, 5.5%, 15%, 25% and 35% of the size of the three training sets. Figure 1 shows the predicted Acc values of each model based on different window size. When the window size is 13, archaea KNN had maximum value of Acc. Therefore, we choose the window size is 13 in the archaea. Similarly, we chose the window size of 21 and 17 for bacteria and vibrio, respectively.

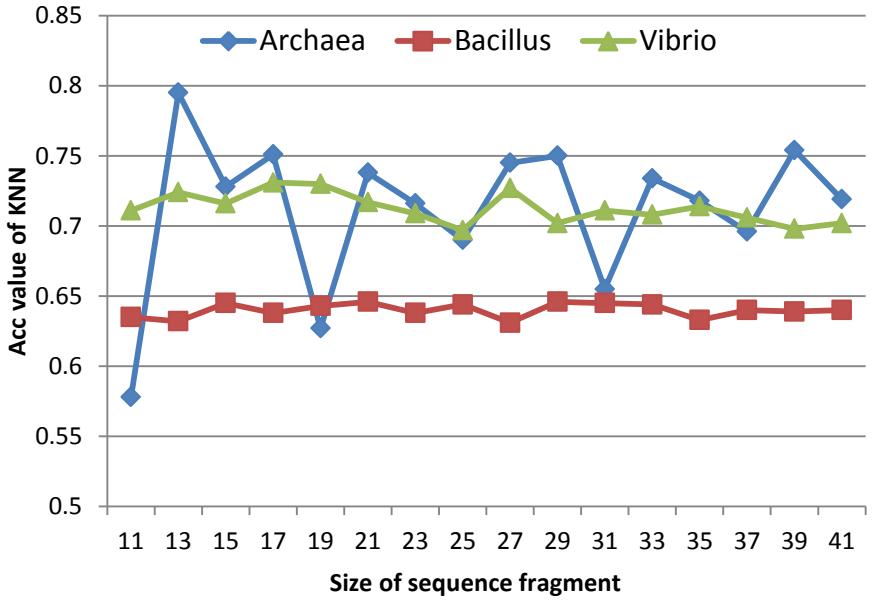



**Figure 1** Acc values of 10-fold cross-validation based on the training dataset and KNN score for sliding window size ranging from 11 to 41.

**Text S2. Sequence-based feature**

**AAC**: The amino acid residues type and position are basic information for a protein sequence, and are widely used in various predicted systems (Chen et al., 2013; Liu et al., 2013). Amino acid composition analysis can efficiently describe the specific state of a given acetylated and non-acetylated in proteome-wide. Therefore, AAC was chosen in the same way as the researchers described (Xu et al., 2015). We through calculated the amino acid residues frequencies in the each of sequence fragments surrounding the query site. There are 20（excluding O）types of amino acids, and thus 20 frequencies are calculated, the sum is equal to 1. Thus, if the length of a protein sequence fragment is L, the dimension of the numeric vector is 20.

**BE**: To transform protein sequences into numeric vectors, we adopted orthogonal binary encoding (Suo et al., 2012). 20 different amino acids together with O are represented by 21 dimensional orthogonal binary vectors, which are ordered as ACDEFGHIKLMNPQRSTVWYO. For example, amino acid A is expressed as 100000000000000000000, amino acid O as 000000000000000000001, and so on. Therefore, if the length of a protein sequence fragment is L, the dimension of the numeric vector is 21*L.

**K-space**: Because the AAC and BE are considered on the basis of a single amino acid residues sequence information, it makes lack of correlation information between amino acids and amino acids. In order to make up for this defect, we took the K-space (Wuyun et al., 2016) to extract correlation information between amino acids. 20 different amino acids together with O had 441 types of amino acid pairs, the 441 type amino acid pairs as follows:

$$\{AA, AC, AD, \cdots, OO\} \quad (1)$$

If an amino acid pairs are separated by other k amino acids in a sequence, its feature vector can be represented as:

$$(N_{AA}, N_{AC}, N_{AD}, \cdots, N_{OO})_{441} \quad (2)$$

Where $N_{ij}$ represents the numbers of corresponding amino acid pairs ij in short peptides, $i, j \in \{A, C, D, \cdots, W, Y, O\}_{21}$, and $K \in \{0,1,2, \cdots, L-1\}$, L is length of protein sequence fragment. In this work, K was chosen to be 0, 1, 2, 3 and 4. So the dimension of the numeric vector is 2205 for a sequence fragment.

**PWAA**: To avoid losing the sequence-order information, we adopted PWAA (Shi et al., 2012) to extract the sequence-order information of amino acid residues around acetyl-lysine sites. Given an amino acid residues $a_i$ $(i = 1,2, \cdots, 21)$, we can express the position information of amino acid $a_i$ in the protein sequence fragment p with

2*m+1 amino acids by the following formula.

$$C_i = \frac{1}{m(m+1)} \sum_{j=-m}^{j=m} x_{i,j} \left( j + \frac{|j|}{m} \right) \qquad (3)$$

Where $m$ denotes the number of upstream residues or downstream residues from the central site in the protein sequence fragment p, $x_{i,j} = 1$ if $a_i$ is the jth position residue in protein sequence fragment p, otherwise $x_{i,j} = 0$ $(j = -m, \cdots, 0, \cdots, m)$. In general, residue $a_i$ is closer to the central site (0 position), the absolute value of $C_i$ is smaller. Therefore, if the length of a protein sequence fragment is L, the dimension of the numeric vector is 21.

**References**

Chen,X. et al. (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. Bioinformatics, 29, 1614-1622.

Liu,B. et al. (2013) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics, 30, 472-479.

Wuyun,Q. et al. (2016) Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set. Plos One, 11, e0155370.

Xu,H.D. et al. (2015) SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics, 31, 3748-3750.

Shi,S.P. et al. (2012) PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. Mol. Biosyst., 8, 1520-1527.

Suo,S.B. et al. (2012) Position-specific analysis and prediction for protein lysine acetylation based on multiple features. Plos One, 7, e49108.

**Text S3. Model Optimization and Evaluation**

SVM is a kind of machine learning algorithm based on statistical learning theory (Noble, 2006), after transforming observed features of positive and negative instances into a vector-based feature space, a ‘maximum margin hyper plane’ that separates the two datasets is created. A radial basis function (RBF) and 10-fold cross-validation were applied to optimize the parameters in the model. Four major parameters of sensitivity (Sn), specificity (Sp), accuracy (Acc) and Mathews Correlation Coefficient (MCC) were chosen to evaluate the prediction performance.

Where,

$$Sn = \frac{TP}{TP + FN}, Sp = \frac{TN}{TN + FP},$$

$$Acc = \frac{Tp + TN}{TP + FP + TN + FN},$$

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

These parameters were defined in terms of the true positive (TP), false negative (FN), true negative (TN), and false positive (FP).

The receiver operating characteristic (ROC) curves were plotted based on Sp and Sn by taking different thresholds, and area under ROC (AUC) was also calculated based on the trapezoidal approximation.

**References**

Noble,W.S. (2006) What is a support vector machine?. Nat. Biotechnol., 24, 1565-1567.

## 2. Supplementary TABLE AND FIGURE

**Table S1. The numbers of acetylated protein and lysine acetylation sites in nine species.**

| Species | E.coli | C.glutamicum | M.tuberculosis | B.subtilis | E.amylovora | S.typhimurium | G.kaustophilus | Archaea | V.parahemolvticus | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein | 1811 | 603 | 649 | 628 | 85 | 190 | 110 | 584 | 656 | 5316 |
| PTM site | 2318 | 1324 | 1108 | 1354 | 124 | 260 | 232 | 654 | 1413 | 8787 |

**Table S2. The numbers of positive and negative samples of training dataset and independent test set among nine species ("positive" represents positive sample; "negative" represents negative sample).**

| Species | Training dataset | | Independent test set | |
|---|---|---|---|---|
| | positive | negative | positive | negative |
| E.coli | 1919 | 12707 | 213 | 1412 |
| C.glutamicum | 1021 | 4333 | 113 | 481 |
| M.tuberculosis | 866 | 3926 | 96 | 436 |
| B.subtilis | 1040 | 5772 | 115 | 641 |
| E.amylovora | 95 | 718 | 10 | 80 |
| S.typhimuricum | 174 | 1467 | 19 | 163 |
| G.kaustophilus | 189 | 1025 | 21 | 114 |
| Archaea | 193 | 1590 | 21 | 176 |
| V.parahemolvticus | 1065 | 5938 | 118 | 659 |

**Table S3. The dimensions of each feature vector and combined all of feature vectors in different species ("Sum" represents dimensions of combined all of feature vectors dimension).**

| | AASA | BE | PWAA | KNN | EBGW | AAC | K-space | Sum |
|---|---|---|---|---|---|---|---|---|
| E.coli | 21 | 441 | 21 | 7 | 15 | 20 | 2205 | 2730 |
| C.glutamicum | 21 | 441 | 21 | 7 | 15 | 20 | 2205 | 2730 |
| M.tuberculosis | 21 | 441 | 21 | 7 | 15 | 20 | 2205 | 2730 |
| B.subtilis | 21 | 441 | 21 | 7 | 15 | 20 | 2205 | 2730 |
| E.amylovora | 21 | 441 | 21 | 5 | 15 | 20 | 2205 | 2728 |
| S.typhimuricum | 21 | 441 | 21 | 5 | 15 | 20 | 2205 | 2728 |
| G.kaustophilus | 21 | 441 | 21 | 5 | 15 | 20 | 2205 | 2728 |
| Archaea | 13 | 273 | 21 | 5 | 15 | 20 | 2205 | 2552 |
| V.parahemolvticus | 17 | 357 | 21 | 5 | 15 | 20 | 2205 | 2640 |

**Table S4. The optimization parameters in eight species models.**

| Species | E.coli | C.glutamicum | M.tuberculosis | B.subtilis | E.amylovora | S.typhimurium | G.kaustophilus | V.parahemolvticus |
|---|---|---|---|---|---|---|---|---|
| $\lambda_2$ | 0.4 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 |
| s | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.3 | 0.3 | 0.1 |

**Table S5. The numbers of seven kinds of feature vectors in optimization model of nine species.**

| | AASA | BE | PWAA | KNN | EBGW | AAC | K_space |
|---|---|---|---|---|---|---|---|
| Archaea | 1 | 24 | 1 | 5 | 2 | 0 | 250 |
| V.parahemolvticus | 4 | 41 | 3 | 5 | 5 | 1 | 244 |
| E.coli | 5 | 62 | 4 | 6 | 4 | 2 | 284 |
| C.glutamicum | 3 | 57 | 4 | 7 | 2 | 2 | 250 |
| M.tuberculosis | 2 | 59 | 3 | 6 | 2 | 0 | 275 |
| B.subtilis | 1 | 76 | 2 | 7 | 3 | 0 | 287 |
| E.amylovora | 3 | 36 | 3 | 4 | 2 | 1 | 106 |
| S.typhimurium | 3 | 54 | 2 | 2 | 0 | 1 | 212 |
| G.kaustophilus | 1 | 49 | 3 | 3 | 1 | 0 | 237 |

**Table S6. Performance evaluation values of Acc, Sn, Sp, MCC and AUC for each of training datasets and independent test datasets among nine species. The "train" represents training datasets; the "test" represents independent test datasets.**

| | | Acc | Sn | Sp | MCC | AUC |
|---|---|---|---|---|---|---|
| E.coli | train | 0.69 | 0.691 | 0.689 | 0.381 | 0.687 |
| | test | 0.899 | 0.887 | 0.911 | 0.798 | 0.887 |
| C.glutamicum | train | 0.8 | 0.807 | 0.794 | 0.602 | 0.807 |
| | test | 0.872 | 0.85 | 0.894 | 0.744 | 0.861 |
| M.tuberculosis | train | 0.834 | 0.83 | 0.838 | 0.671 | 0.83 |
| | test | 0.88 | 0.885 | 0.875 | 0.76 | 0.863 |
| B.subtilis | train | 0.796 | 0.799 | 0.792 | 0.592 | 0.8 |
| | test | 0.952 | 0.957 | 0.948 | 0.904 | 0.942 |
| E.amylovora | train | 0.983 | 0.978 | 0.989 | 0.967 | 0.988 |
| | test | 0.9 | 0.8 | 1 | 0.816 | 0.896 |
| S.typhimurium | train | 0.874 | 0.871 | 0.876 | 0.753 | 0.882 |
| | test | 0.816 | 0.895 | 0.737 | 0.64 | 0.778 |
| G.kaustophilus | train | 0.897 | 0.917 | 0.878 | 0.801 | 0.902 |
| | test | 0.881 | 0.952 | 0.81 | 0.77 | 0.854 |
| Archaea | train | 0.9 | 0.9 | 0.9 | 0.803 | 0.897 |
| | test | 0.81 | 0.81 | 0.81 | 0.619 | 0.855 |
| V.parahemolvtius | train | 0.802 | 0.808 | 0.796 | 0.605 | 0.799 |
| | test | 0.869 | 0.89 | 0.847 | 0.738 | 0.855 |

**Table S7.** Comparison of prediction performance between our method and other tools (Sn: sensitivity, Sp: specificity, Acc: accuracy, MCC: Matthew correlation coefficient).

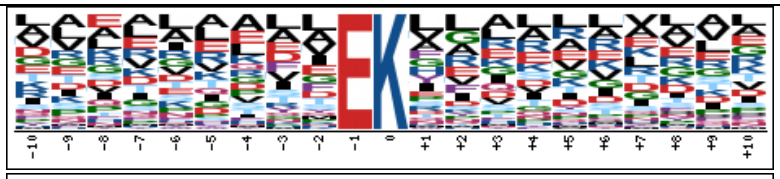
| **Species** | **Predictors** | **Acc** | **Sn** | **Sp** | **MCC** | **AUC** |
|---|---|---|---|---|---|---|
| Archaea | PLMLA | 0.619 | 0.593 | 0.667 | 0.249 | 0.677 |
| | Ensemblepail | 0.548 | 0.556 | 0.542 | 0.096 | 0.539 |
| | LAceP | 0.452 | 0.45 | 0.455 | -0.095 | 0.591 |
| | Phosida | 0.476 | 0.333 | 0.487 | -0.092 | 0.402 |
| | our work | 0.81 | 0.81 | 0.81 | 0.619 | 0.855 |
| V.parahemolvticus | PLMLA | 0.559 | 0.551 | 0.571 | 0.12 | 0.571 |
| | Ensemblepail | 0.581 | 0.598 | 0.568 | 0.164 | 0.59 |
| | LAceP | 0.496 | 0.496 | 0.496 | -0.008 | 0.481 |
| | Phosida | 0.572 | 0.758 | 0.542 | 0.208 | 0.63 |
| | our work | 0.869 | 0.89 | 0.847 | 0.738 | 0.855 |
| E.coli | PLMLA | 0.573 | 0.564 | 0.584 | 0.147 | 0.585 |
| | Ensemblepail | 0.538 | 0.552 | 0.529 | 0.078 | 0.558 |
| | LAceP | 0.498 | 0.497 | 0.498 | -0.005 | 0.501 |
| | SSPKA | 0.516 | 0.549 | 0.51 | 0.044 | 0.474 |
| | Phosida | 0.531 | 0.633 | 0.517 | 0.096 | 0.542 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | our work | 0.899 | 0.887 | 0.911 | 0.798 | 0.887 |
| C.glutamicum | PLMLA | 0.575 | 0.568 | 0.584 | 0.151 | 0.577 |
| | Ensemblepail | 0.491 | 0.485 | 0.494 | -0.019 | 0.441 |
| | LAceP | 0.509 | 0.509 | 0.509 | 0.018 | 0.548 |
| | Phosida | 0.584 | 0.806 | 0.549 | 0.244 | 0.708 |
| | our work | 0.872 | 0.85 | 0.894 | 0.744 | 0.861 |
| M.tuberculosis | PLMLA | 0.547 | 0.537 | 0.565 | 0.098 | 0.528 |
| | Ensemblepail | 0.557 | 0.593 | 0.541 | 0.124 | 0.559 |
| | LAceP | 0.443 | 0.438 | 0.447 | -0.115 | 0.453 |
| | Phosida | 0.583 | 0.808 | 0.548 | 0.244 | 0.679 |
| | our work | 0.88 | 0.885 | 0.875 | 0.76 | 0.863 |
| B.subtilis | PLMLA | 0.6 | 0.58 | 0.632 | 0.206 | 0.608 |
| | Ensemblepail | 0.535 | 0.536 | 0.533 | 0.07 | 0.522 |
| | LAceP | 0.513 | 0.513 | 0.513 | 0.026 | 0.509 |
| | Phosida | 0.6 | 0.78 | 0.561 | 0.261 | 0.662 |
| | our work | 0.952 | 0.957 | 0.948 | 0.904 | 0.942 |
| E.amylovora | PLMLA | 0.5 | 0.5 | 0.5 | 0 | 0.479 |
| | Ensemblepail | 0.65 | 0.667 | 0.636 | 0.302 | 0.657 |
| | LAceP | 0.7 | 0.7 | 0.7 | 0.4 | 0.85 |
| | Phosida | 0.4 | 0.25 | 0.438 | -0.25 | 0.313 |
| | our work | 0.9 | 0.8 | 1 | 0.816 | 0.896 |
| S.typhimurium | PLMLA | 0.553 | 0.55 | 0.556 | 0.105 | 0.55 |
| | Ensemblepail | 0.605 | 0.625 | 0.591 | 0.213 | 0.665 |
| | LAceP | 0.553 | 0.55 | 0.556 | 0.105 | 0.647 |
| | SSPKA | 0.868 | 1 | 0.792 | 0.764 | 0.893 |
| | Phosida | 0.553 | 0.667 | 0.531 | 0.144 | 0.651 |
| | our work | 0.816 | 0.895 | 0.737 | 0.64 | 0.778 |
| G.kaustophilus | PLMLA | 0.452 | 0.464 | 0.429 | -0.101 | 0.457 |
| | Ensemblepail | 0.429 | 0.286 | 0.457 | -0.192 | 0.339 |
| | LAceP | 0.476 | 0.474 | 0.478 | -0.048 | 0.515 |
| | Phosida | 0.524 | 0.6 | 0.514 | 0.074 | 0.605 |
| | our work | 0.881 | 0.952 | 0.81 | 0.77 | 0.854 |

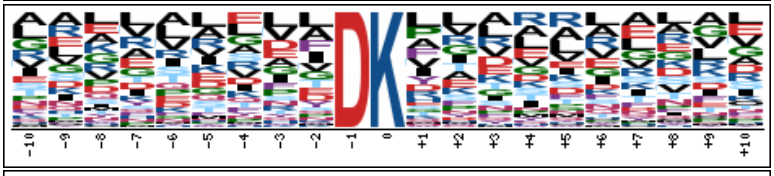
## Prokaryote acetylation sites motif

| # | Motif | Motif Score | Foreground Matches | Foreground Size | Background Matches | Background Size | Fold Increase |
|---|---|---|---|---|---|---|---|
| 1. | .....EK..... | 16.00 | 853 | 6552 | 3620 | 40897 | 1.47 |
| 2. | .....DK..... | 16.00 | 547 | 5699 | 2297 | 37277 | 1.56 |
| 3. | .....K....R..... | 16.00 | 468 | 5152 | 1884 | 34980 | 1.69 |
| 4. | .....K...R..... | 16.00 | 365 | 4684 | 1604 | 33096 | 1.61 |
| 5. | .....KY..... | 16.00 | 296 | 4319 | 650 | 31492 | 3.32 |
| 6. | .....KH..... | 16.00 | 275 | 4023 | 642 | 30842 | 3.28 |
| 7. | .....K.F..... | 16.00 | 252 | 3748 | 1095 | 30200 | 1.85 |
| 8. | .....F.K..... | 16.00 | 241 | 3496 | 922 | 29105 | 2.18 |
| 9. | ...K.....K..... | 16.00 | 206 | 3255 | 606 | 28183 | 2.94 |
| 10. | .....KR..... | 10.36 | 281 | 3049 | 1705 | 27577 | 1.49 |
| 11. | .....K.....R.. | 9.25 | 223 | 2768 | 1363 | 25872 | 1.53 |
| 12. | .....K...K..... | 8.41 | 218 | 2545 | 1400 | 24509 | 1.50 |
| 13. | .....K...K..... | 9.30 | 212 | 2327 | 1361 | 23109 | 1.55 |

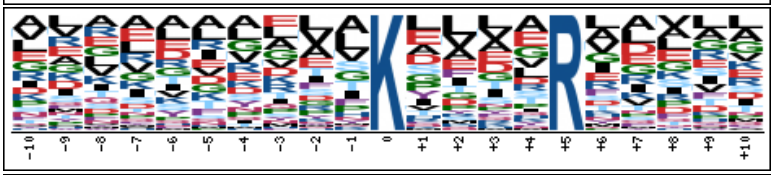
1#



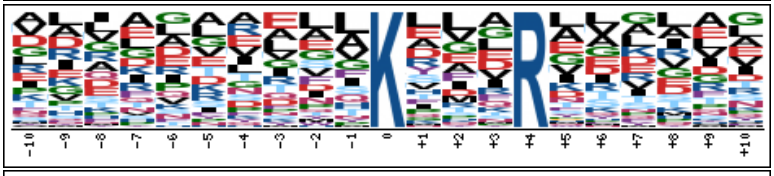

2#



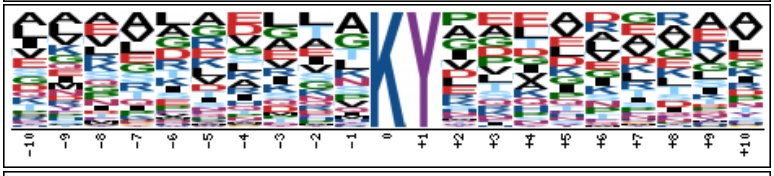

3#



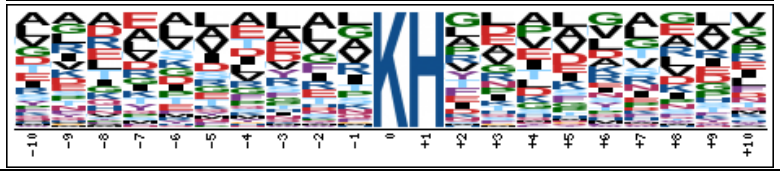

4#

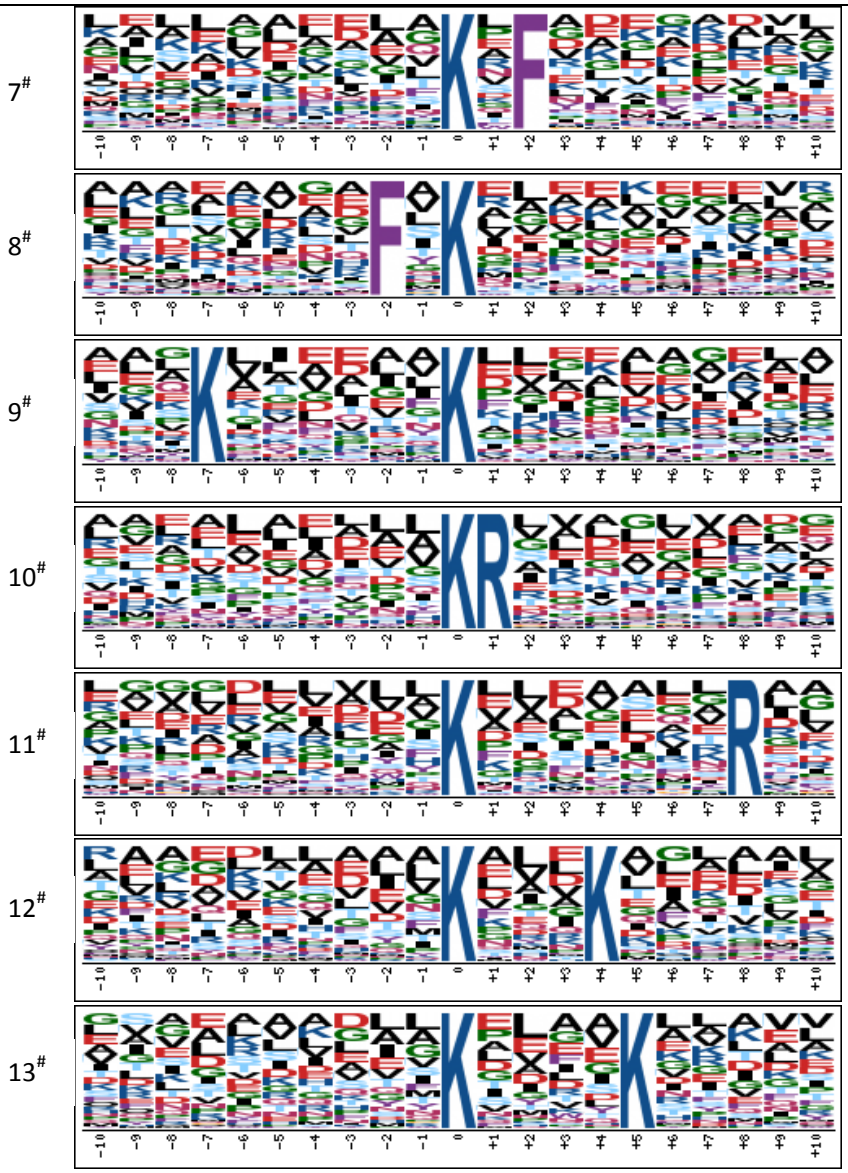



5#




6#






Eukaryote acetylation site motif

| # | Motif | Motif Score | Foreground Matches | Foreground Size | Background Matches | Background Size | Fold Increase |
|---|---|---|---|---|---|---|---|
| 1. | .....AEK..... | 31.65 | 302 | 28053 | 1974 | 346912 | 1.89 |
| 2. | .....GKL..... | 32.00 | 305 | 27751 | 1626 | 344938 | 2.33 |
| 3. | ..K.....GK..... | 32.00 | 280 | 27446 | 1391 | 343312 | 2.52 |
| 4. | .....SKP..... | 30.16 | 218 | 27166 | 1331 | 341921 | 2.06 |
| 5. | ..K.....AK..... | 23.06 | 252 | 26948 | 1717 | 340590 | 1.85 |
| 6. | ...K....A.K..... | 23.43 | 250 | 26696 | 1620 | 338873 | 1.96 |
| 7. | .....GK..... | 16.00 | 2124 | 26446 | 17562 | 337253 | 1.54 |
| 8. | ...V.....EK..... | 22.46 | 220 | 24322 | 1728 | 319691 | 1.67 |
| 9. | .....AK.....K... | 22.23 | 225 | 24102 | 1603 | 317963 | 1.85 |
| 10. | .....AA.K..... | 26.24 | 217 | 23877 | 1501 | 316360 | 1.92 |

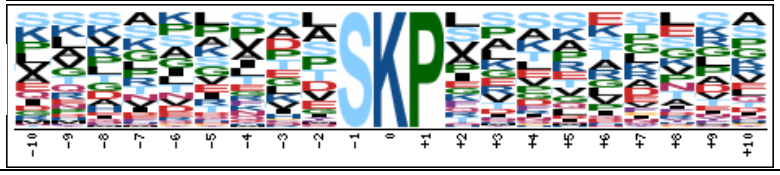
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11. | .....EK..... | 16.00 | 2665 | 23660 | 30259 | 314859 | 1.17 |
| 12. | .....SKL..... | 25.74 | 274 | 20995 | 2089 | 284600 | 1.78 |
| 13. | .....DKL..... | 23.35 | 235 | 20721 | 1372 | 282511 | 2.34 |
| 14. | .....AK..... | 16.00 | 1704 | 20486 | 17523 | 281139 | 1.33 |
| 15. | .....LKP..... | 32.00 | 238 | 18782 | 1484 | 263616 | 2.25 |
| 16. | .....VK..K..... | 23.09 | 200 | 18544 | 1572 | 262132 | 1.80 |
| 17. | .....DK..... | 16.00 | 1547 | 18344 | 13405 | 260560 | 1.64 |
| 18. | .....LK...K..... | 24.35 | 297 | 16797 | 2648 | 247155 | 1.65 |
| 19. | .....SK..... | 16.00 | 1777 | 16500 | 20839 | 244507 | 1.26 |
| 20. | .....VK..... | 16.00 | 1512 | 14723 | 18051 | 223668 | 1.27 |
| 21. | .....LK.....K.... | 22.13 | 220 | 13211 | 2048 | 205617 | 1.67 |
| 22. | .....QK..... | 16.00 | 1407 | 12991 | 17652 | 203569 | 1.25 |
| 23. | .....LK..... | 16.00 | 2062 | 11584 | 27516 | 185917 | 1.20 |
| 24. | .....TK..... | 16.00 | 1342 | 9522 | 15440 | 158401 | 1.45 |
| 25. | .....IK..... | 16.00 | 1262 | 8180 | 15338 | 142961 | 1.44 |
| 26. | .....NK..... | 16.00 | 1089 | 6918 | 12858 | 127623 | 1.56 |
| 27. | .....FK..... | 16.00 | 972 | 5829 | 9688 | 114765 | 1.98 |
| 28. | .....YK..... | 16.00 | 773 | 4857 | 7982 | 105077 | 2.10 |
| 29. | .....PK..... | 16.00 | 957 | 4084 | 17327 | 97095 | 1.31 |
| 30. | .....MK..... | 16.00 | 624 | 3127 | 8367 | 79768 | 1.90 |
| 31. | .....HK..... | 16.00 | 495 | 2503 | 7265 | 71401 | 1.94 |
| 32. | .....CK..... | 16.00 | 397 | 2008 | 5477 | 64136 | 2.32 |
| 33. | .....WK..... | 16.00 | 284 | 1611 | 4035 | 58659 | 2.56 |

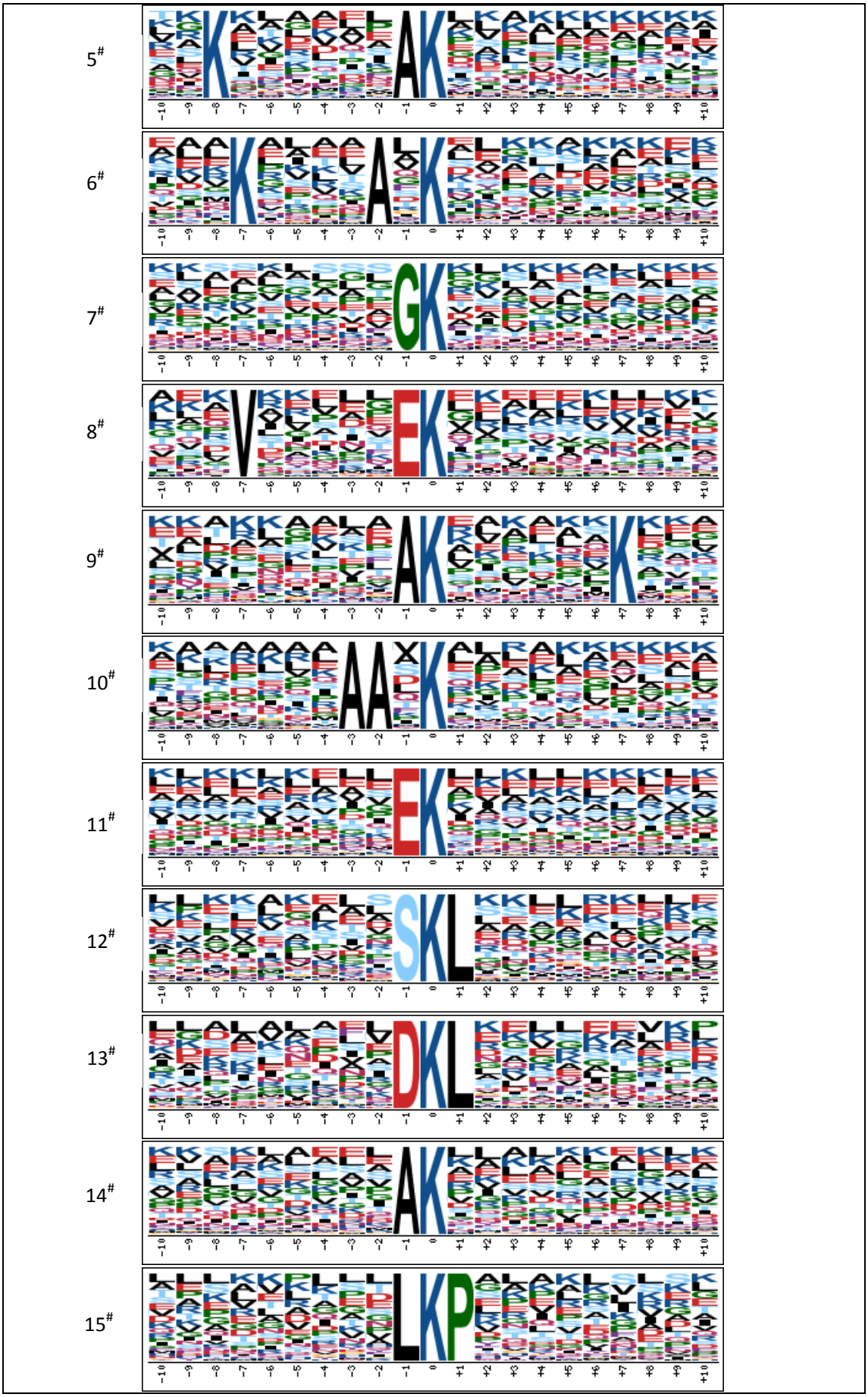
1#

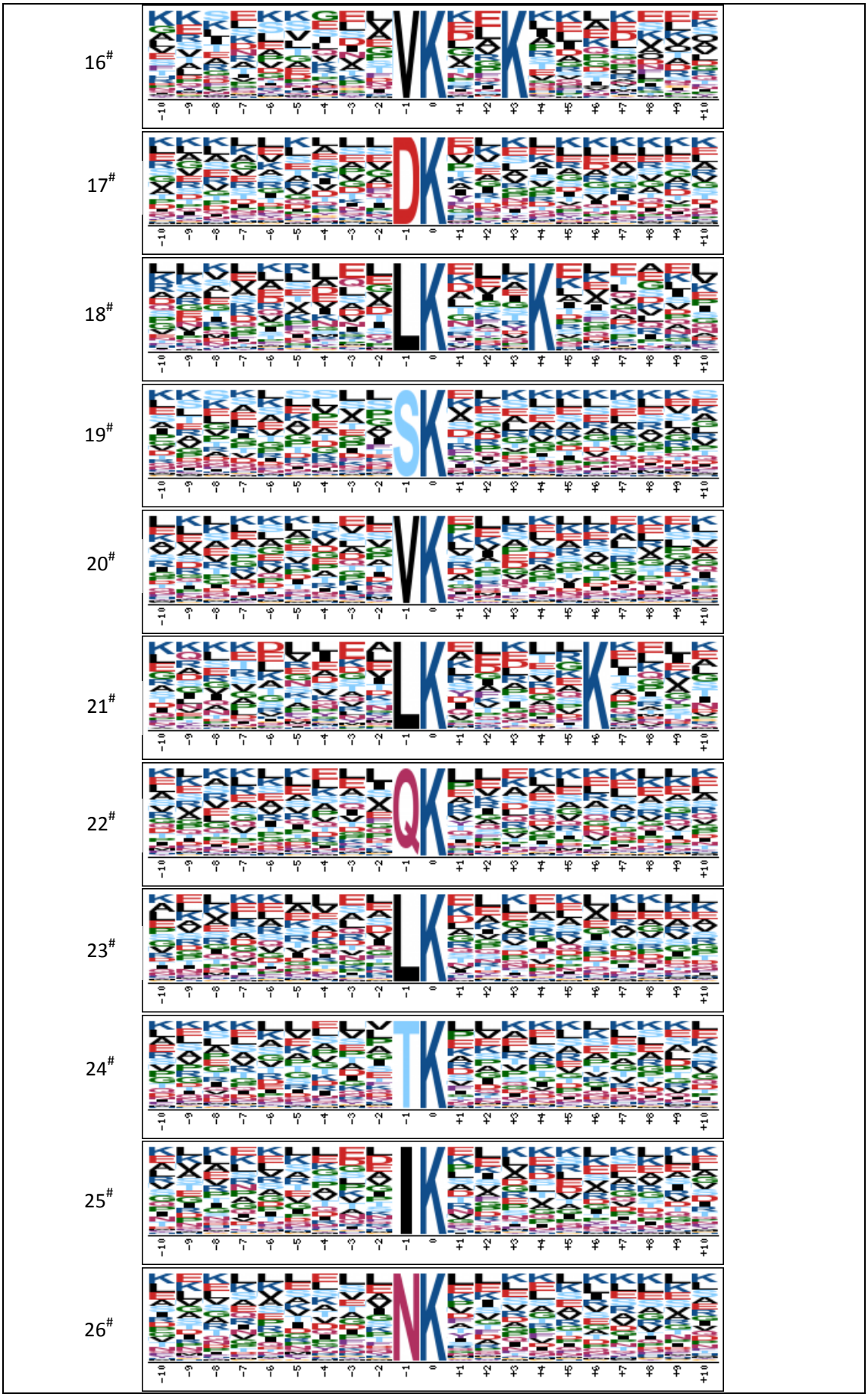

2#


3#


4#




5#
6#
7#
8#
9#
10#
11#
12#
13#
14#
15#


16#
17#
18#
19#
20#
21#
22#
23#
24#
25#
26#

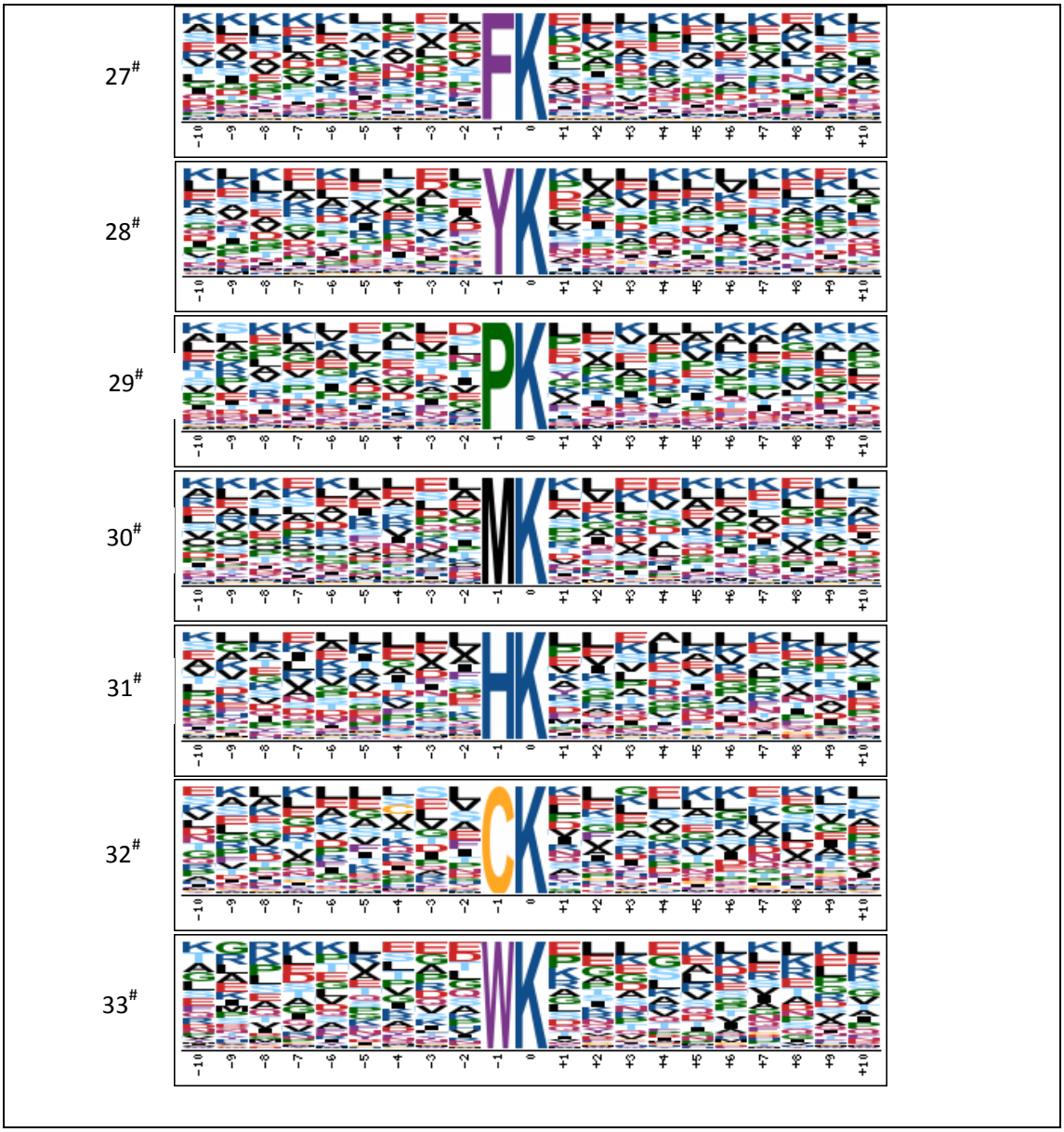



**Fig. S1. Acetylation sites motif of prokaryote and eukaryote.**

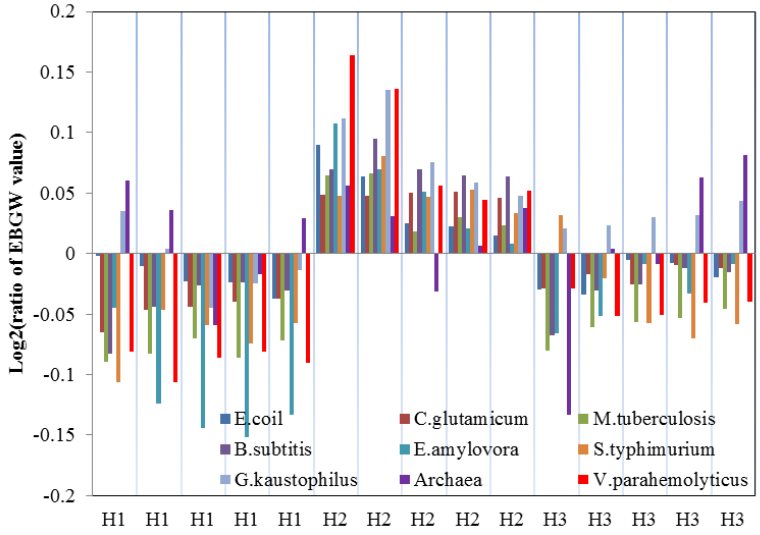



**Fig. S2 Comparison of EBGW between acetylation and non-acetylation. The vertical axis represents the log_2 ratio of average EBGW values between acetylation and non-acetylation. The horizontal axis represents the three binary sequences.**

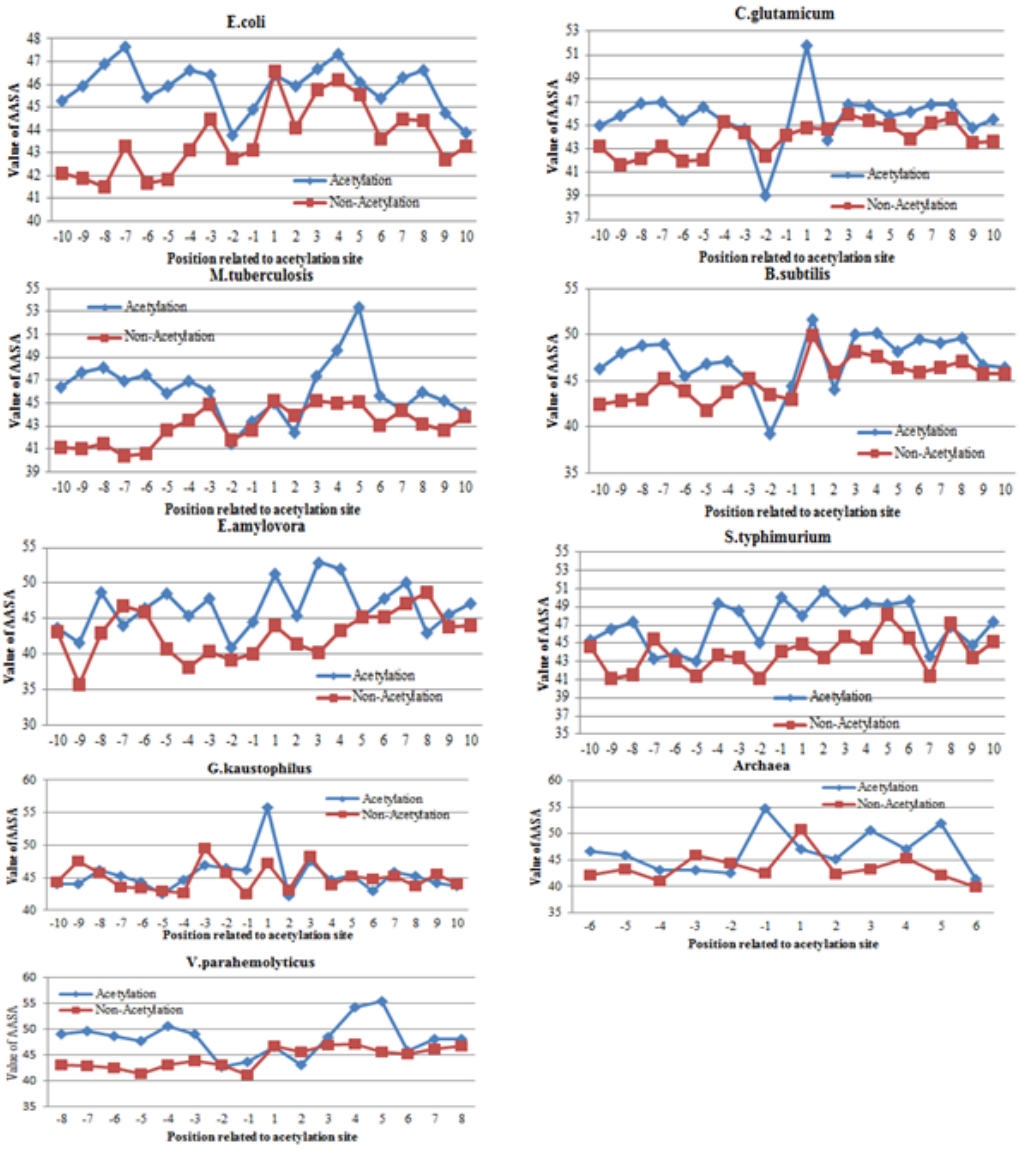



**Fig. S3. Average accessible surface area (AASA) value of residues around acetylation sites and non-acetylation sites (except of center position) among nine species.**

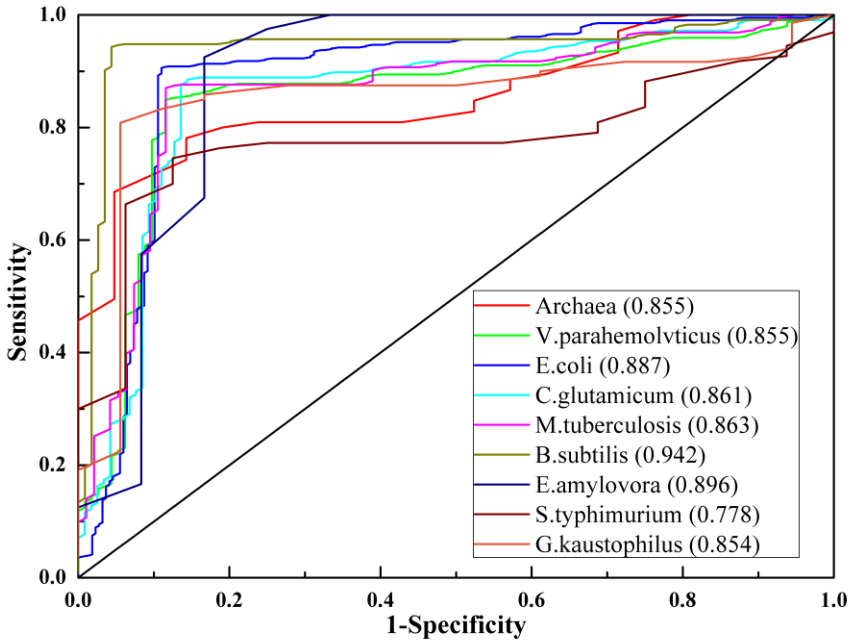



**Fig. S4. The ROC curves and AUC value of 10-fold cross-validations (CV) of the independent test set for nine species.**