LAFEM: A Scoring Model to Evaluate Functional Landscape of Lysine Acetylome

Authors

Jun-Ze Liang, De-Hua Li, Yong-Chun Xiao, Fu-Jin Shi, Tairan Zhong, Qian-Ying Liao, Yang Wang, and Qing-Yu He

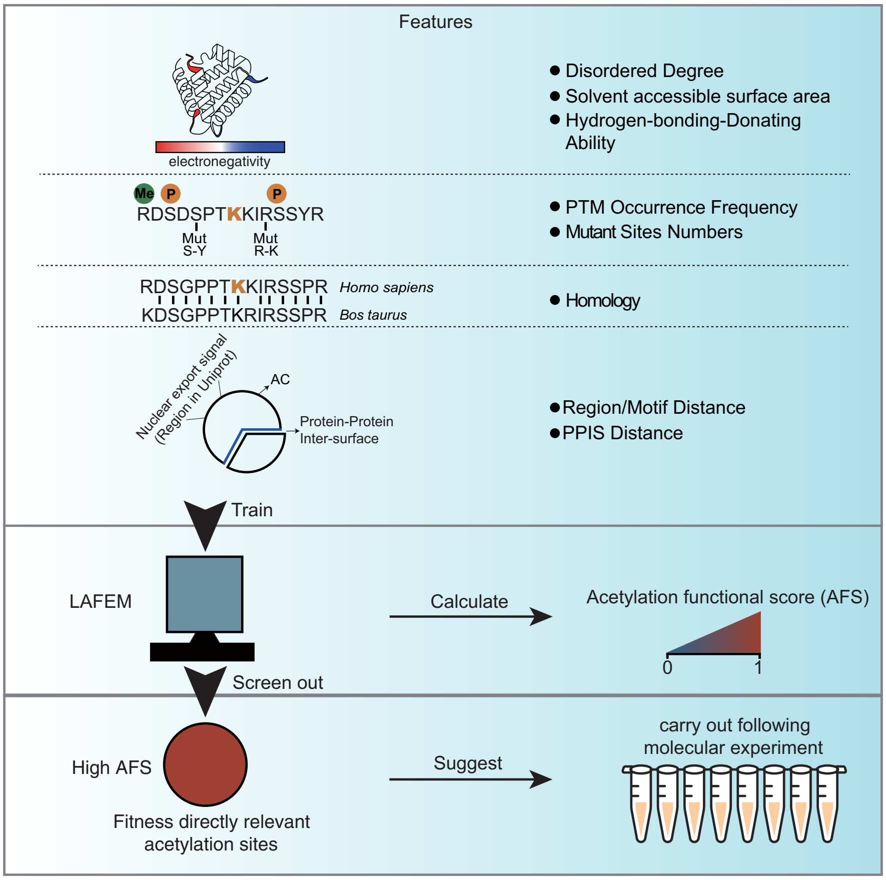
Correspondence

wangyang0507@jnu.edu.cn; tqyhe@email.jnu.edu.cn

In Brief

Not all acetylation sites contribute equally to fitness, which disturbs us to find out the core reason between different biological models. In fact, it is insufficient to evaluate functional acetylation sites only relying on quantification acetylome. Therefore, we developed LAFEM to comprehensively consider the molecular environment of individual acetylation sites. The LAFEM score of acetylation site is beneficial to the following experimental validation and the discover of core biological regularity.

Graphical Abstract



Highlights

- LAFEM is the first prioritization strategy to evaluate functional acetylome.
- LAFEM enriches functional annotations of 15,410 acetylation sites.
- Eight molecular features in LAFEM are related to the function of acetylation site.
- Assessment of acetylation sites by LAFEM provide support for quantitative acetylomics.

2024, Mol Cell Proteomics 23(1), 100700

https://doi.org/10.1016/j.mcpro.2023.100700

LAFEM: A Scoring Model to Evaluate Functional Landscape of Lysine Acetylome

Jun-Ze Liang[1,‡], De-Hua Li[1,‡], Yong-Chun Xiao[2,‡], Fu-Jin Shi[1], Tairan Zhong[1], Qian-Ying Liao[3], Yang Wang[1,*], and Qing-Yu He[1,*]

Protein lysine acetylation is a critical post-translational modification involved in a wide range of biological processes. To date, about 20,000 acetylation sites of *Homo sapiens* were identified through mass spectrometry–based proteomic technology, but more than 95% of them have unclear functional annotations because of the lack of existing prioritization strategy to assess the functional importance of the acetylation sites on large scale. Hence, we established a lysine acetylation functional evaluating model (LAFEM) by considering eight critical features surrounding lysine acetylation site to high-throughput estimate the functional importance of given acetylation sites. This was achieved by selecting one of the random forest models with the best performance in 10-fold cross-validation on undersampled training dataset. The global analysis demonstrated that the molecular environment of acetylation sites with high acetylation functional scores (AFSs) mainly had the features of larger solvent-accessible surface area, stronger hydrogen bonding–donating abilities, near motif and domain, higher homology, and disordered degree. Importantly, LAFEM performed well in validation dataset and acetylome, showing good accuracy to screen out fitness directly relevant acetylation sites and assisting to explain the core reason for the difference between biological models from the perspective of acetylome. We further used cellular experiments to confirm that, in nuclear casein kinase and cyclin-dependent kinase substrate 1, acetyl-K35 with higher AFS was more important than acetyl-K9 with lower AFS in the proliferation of A549 cells. LAFEM provides a prioritization strategy to large scale discover the fitness directly relevant acetylation sites, which constitutes an unprecedented resource for better understanding of functional acetylome.

Lysine acetylation is a pivotal form of protein post-translational modification (PTM) first identified in histones (1). The addition of acetyl group shields the positive charge on the lysine, leading genomic DNA to be looser and active transcription (2, 3). Therefore, protein lysine acetylation is typically associated with a wide range of cellular processes, including cell cycle, cytoskeleton organization, DNA damage repairing, and RNA processing (4, 5). Identification of the novel acetylation sites and characterization of their functions will deepen our grasp of complex life.

Mass spectrometry (MS)–based proteomics is a high-throughput technology to identify PTM sites (6). With advancements of acetylome profiling, more than 20,000 acetylation sites on 5000 human proteins have been identified (7). Up to now, a total of 38,387 acetylation sites from diverse species have been recorded in the PhosphositePlus (PSP) database (8). However, a vast majority of these recorded acetylation sites are lack of annotations about their molecular functions. Since not all acetylation sites are likely to contribute equally to fitness (9–11), functional importance prediction of acetylation sites on large scale is urgently needed. An effective functional predicted strategy not only increase the efficiency of finding out the core reason for the difference between biological models but also assist in following molecular experiments. Currently, quantitative acetylomics remains the primary approach to evaluate the functional importance of acetylation sites. Nevertheless, relying solely on one dimension of expression difference is insufficient to overcome the difficulty in functional prediction. Actually, molecular functions of acetylation sites are also influenced by unmined features of molecular environment surrounding acetylation sites.

The on/off status of a certain acetylation site is highly relied on multiproperties of its surrounding environment (12); these properties are thus meaningful to the establishment of functional evaluating model based on machine learning. Recently, a large-scale prioritization strategy was proposed to estimate the functions of phosphorylation sites by using 59 properties of phosphorylation related to MS evidence, phosphorylation

From the [1]MOE Key Laboratory of Tumor Molecular Biology and State Key Laboratory of Bioactive Molecules and Druggability Assessment, College of Life Science and Technology, Jinan University, Guangzhou, China; [2]Department of Orthopedics, The First Affiliated Hospital of Jinan University, Guangzhou, China; [3]IMEC-DistriNet Research Group, Department of Computer Science, KU Leuven, Leuven, Belgium
[‡]These authors contributed equally to this work.
*For correspondence: Qing-Yu He, tqyhe@email.jnu.edu.cn; Yang Wang, wangyang0507@jnu.edu.cn.

https://doi.org/10.1016/j.mcpro.2023.100700

site regulation, structural environment, and evolutionary conservation (13). Also, several features such as solvent-accessible surface area (SASA), intrinsically disordered regions (IDRs), and hydrogen bonding–donating abilities were reported to highly impact on the function of acetylation. Larger SASA indicates higher possibility to interact with other molecules (14). IDR is an unfolded functional region that is not based on sequence–structure–function paradigm (15). It was found that many regulatory acetylation sites are located in IDRs (12). Hydrogen bonding–accepting ability refers to electronegativity, which is the key to the function of acetylation.

Here, we integrated eight features of lysine acetylation's molecular environment from 2304 acetylation sites to train a scoring model based on random forest (RF) model, termed lysine acetylation functional evaluating model (LAFEM), aiming at scoring the functional importance of acetylation sites. Acetylation functional score (AFS) is provided for evaluating functional importance of acetylation sites by LAFEM, which constitutes an unprecedented resource for better understanding the landscape of functional acetylome.

EXPERIMENTAL PROCEDURES

Labels of Acetylation Sites

In UniProt database, acetylation site with low-throughput experimental evidence was annotated by "Manual assertion based on experiment" or "Manual assertion inferred from combination of experimental and computational evidence." Likewise, in PSP, acetylation site with low-throughput experimental evidence was defined as that at least one piece of literature with low-throughput evidence of the site validated the site. We accordingly labeled all acetylation sites with EXP or non-EXP representing as positive and negative samples for subsequent study. A total of 1280 acetylation sites were labeled with EXP, and 3621 were labeled with non-EXP for training and testing.

SASA Score

AlphaFold2 is a software constructed by neural network–based model (16). It can predict tertiary structure of protein with atomic accuracy in large scale. We downloaded the tertiary structures of protein dataset in *Homo sapiens* from the official website (https://ftp.ebi.ac.uk/pub/databases/alphafold/latest/UP000005640_9606_HUMAN_v4.tar) of AlphaFold2 (DeepMind, Alphabet Inc) Protein Data Bank file was stored as tertiary structure of protein. With Python interfaces, FreeSASA (University Library, University of Bergen) is the software to compute SASA (17). All parameters of FreeSASA were set to default. SASA score was the sum of every amino acid in acetylation site adjacent sequence (ASAS)'s SASA.

Disordered Score

DISOPRED3 is an advanced software to identify IDRs (18). To execute DISOPRED3, BLAST software (version 2.2.26) (19), matrices from the National Center for Biotechnology Information database, and uniref90 file from Universal Protein (UniProt) database were required. After inputting protein sequence file to DISOPRED3, DISOPRED3 would output the diso file containing the possibility of whether the amino acid residue was disordered. The possibility of disordered residue of every amino acid residue at ASAS was summed to evaluate the disorder degree of acetylation sites' surrounding environment. All parameters of DISOPRED3 were set default. Disordered score refers to the average of every amino acid in ASAS's possibility of disordered residue.

Distance from Region/Motif to Acetylation Site, PTM Site, or Mutant Site Locations

Structural components in protein with specific spatial conformations and functions exhibit certain regularities in sequence, which are recorded in the motif (https://www.uniprot.org/help/motif) and region (https://www.uniprot.org/help/region) sections of UniProt. The region section of UniProt is illustrated as follows: It may be used to describe characteristics of a region such as the role of a region in mediating protein–protein interactions or some other biological process, to describe the different regions in multifunctional enzymes or fusion proteins, or to fully describe the characteristics of a region containing repeats. The motif section of UniProt is illustrated as follows: motif is a short (usually not more than 20 amino acids) conserved sequence of biological significance. Specific sequence motifs usually mediate a common function, such as protein binding or targeting a particular subcellular location, in a variety of proteins. The software package rvest was used to download the information of the region/motif's location from UniProt. The distance from acetylation site to region/motif was defined as the shortest distance from acetylation site to any region/motif on proteins ($\log_2$(distance +1) used in the model). If acetylation site was located in region/motif, the distance from acetylation site to region/motif was set 0. All missing values were imputed by the length of protein.

The location of PTM sites and mutant sites was collected from UniProt and PSP databases. PTM sites included common types of PTM, such as phosphorylation, acetylation, methylation, lipidation, glycosylation, and crosslink such as disulfide bond. The mutant sites mainly contained natural but not artificial mutagenesis.

Homologous Score

We downloaded the protein sequence fasta files of six species respectively, that is, *Bos taurus*, *Danio rerio*, *Rattus norvegicus*, *Xenopus laevis*, *Mus musculus*, and *Drosophila melanogaster*. The homologous pairing proteins between *H. sapiens* and the other six species were identified using BLAST+ software (version 2.12.0): We constructed six species databases of proteins through the makeblastdb module of BLAST+. Then, we narrowed down the range for searching homologous proteins through BLAST+. The protein was considered as a candidate homologous protein if its peptides' evalue were less than 0.001 (20). After setting the penalties of gap opening and gap extension to 10 and 4 (21), the homologous protein was the protein with the highest global matching score among candidates. The commands of BLAST+ to construct database and blast were uploaded to Zenodo (a digital library framework powered by the CERN (the European Organization for Nuclear Research) Data Centre and Invenio).

Twenty amino acids before and after acetylation site were used to calculate paring homologous score of acetylation site. Blosum62 matrix was the scoring matrix of calculating homologous score. All missing values were imputed by −174 (−4 * 41 − 10), which was the lowest homologous paring score of 41 amino acids. TimeTree is a public knowledge base of divergence times among organisms (22), currently containing 137,706 species (23). The weight of homologous paring score was defined as the logarithm of divergence time between *H. sapiens* and other species. The isolation time between *H. sapiens* and *B. taurus*, *D. rerio*, *R. norvegicus*, *X. laevis*, *M. musculus*, and *D. melanogaster* was l94.0, 430.7, 87.2, 352.7, 87.2, 693.5 million years, respectively (Supplemental Fig. S1*A*). The weights of after-mentioned species were $log_2(isolation\ time)$. All parameters of TimeTree were set as default. Homologous score meant the weighted sum of homologous pairing scores between ASAS in *H. sapiens* and homologous sequences in other six species.

Hydrogen Bonding–Donating Ability Score

Eighteen parameters of 20 amino acids were used to perform a principal component analysis. One of the principal properties termed PP3 was defined as hydrogen bonding–donating ability, and the PP3 axis scale of per amino acid was defined previously (24). The package Peptide was used to compile diverse properties of amino acids and to calculate PP3 axis scale of peptides. Higher scores indicated stronger hydrogen bonding–donating abilities, as is the case with basic amino acids. Hydrogen bonding–donating ability score was the average of PP3 axis scale scores in ASAS of every amino acid.

Distance from Protein–Protein Intersurface to Acetylation Site

Compared with protein–protein interaction annotations in UniProt database, Interactome has more comprehensive and accurate annotations. Interactome3D renewed the protein–protein interaction annotations in UniProt by constructing virtual protein–protein structure after protein sequence alignment (25). The dataset of protein–protein intersurface (PPIS) location was downloaded from the Interactome3D website. The distance from acetylation site to PPIS was defined as the shortest primary structural distance to PPIS ($\log_2$(distance + 1) used in model). If acetylation site was located in the sequence of PPIS, the distance from acetylation site to PPIS was set as 0. All missing values were imputed by length of protein.

Model Training

We selected five common machine learning models including generalized linear model *via* lasso and elastic-net regularization (GLMNET), K-nearest-neighbor (KNN), RF, decision tree (recursive partitioning and regression tree), and support vector machine (SVM) to compare their performance in this study. GLMNET is a linear model with elastic-net penalty. The response variable is the classifications, and the predicted vectors are the features (26). KNN is the best-known nonparametric method used for classification. The principle is that KNN classifies new samples into the closest classification through feature clustering (27). RF model is an ensemble classifier that trains several decision trees in parallel with bootstrapping followed by aggregation (28). Decision tree is a supervised machine learning technique for inducing a decision tree from training data. In the structures of decision trees, leaves represent classifications, nonleaf nodes are features, and branches are conjunctions of features that cause classification. The goal of the SVM finds a hyperplane in the data space that produces the largest distance from samples of different classes to hyperplane. All machine learning models in this study were constructed by mlr3verse, and the main parameters of these models were set as default.

AFS

AFS is the probability that acetylation site is adjudicated to be directly relevant fitness by a trained RF model. This probability is derived from the average predicted class probability among decision trees in an RF. The class probability of a decision tree is determined by the classification score of the samples within its terminal leaves.

Acetylation Site Quantification

MS/MS data (PDC000233 and PDC000224) from Clinical Proteomic Tumor Analysis Consortium (CPTAC) were processed by MaxQuant (Max-Planck Institute for Biochemistry) (29). Every sample was divided into four fractions. The tolerance of precursor ion and fragment ion was respectively set as default 4.5 and 20 ppm. The minimum peptide length was set as 7, and the maximum peptide mass was set as 4600. The minimum unique peptide and razor peptide were respectively set as 0 and 1. Trypsin/P was designated as the cleavage enzyme, and the digested model was set as specific, allowing for up to two missing cleavages with default numbers of modifications for every peptide. Acetyl (K), oxidation (M), and acetyl (protein N-term) were set as variable modifications, whereas fixed modification contained carbamidomethyl (C). The estimated false discovery rate thresholds of modification site, peptide-spectrum matching, and protein were all set as 0.01. “Match between run” was set as true. Label-free quantification was selected for acetylation quantification. Other parameters were all set as default in MaxQuant (version 1.6.5.0).

Differential Expression Analysis of Acetylation Site Processing

We respectively analyzed the two acetylome datasets by following process: Initially, we filtered out acetylation sites with missing values less than 20%. After taking logarithm, we used quantile regression imputation of left-censored data to impute missing values. Then, we rectified the acetylome through a pool sample in every tandem mass tag quantitative spectrometric experiment. Comparing tissues of cancer and adjacent cancer, we screened out differential expression acetylation sites (DEASs; $|log_2(FC)| \geq 1$ & $p \leq 0.05$) and non-DEASs ($|log_2(FC)| < 1$ & $p > 0.0$) through Mann–Whitney *U* test and fold change.

Cancer Cell Line Encyclopedia

Cancer cell line encyclopedia through quantitative profiling of thousands of proteins by MS across 375 cell lines from diverse lineages to enhance their database from the perspective of protein. In our study, we selected eight lung cancer cell lines including A549, NCI-H1299, NCI-H1975, NCI-H460, NCI-H358, NCI-H3255, PC14, and NCI-H441 to find out the lung cancer with the lowest endogenous expression of nuclear casein kinase and cyclin-dependent kinase substrate 1 (NUCKS1).

Cell Culture and Transfection

Human lung carcinoma cell line A549 was acquired from the American Type Culture Collection. Cells were cultured in Dulbecco's modified Eagle's medium (Life Technologies) with 10% fetal bovine serum in 5% CO_2 at 37 °C. About 1 μg pCDH-plasmid and its NUCKS1-WT, NUCKS1-K35R, and NUCKS1-K9R recombination plasmids were respectively mixed into OPTI-MEM (Life Technologies) with 1 ml/l Lip3000 (Invitrogen), 1 ml/l P3000 (Invitrogen), 1 μg pMD2.G plasmid, and 1 μg psPAX2 plasmid. Forty-eight hours after transfection, cells were expanded and ed out by 2.5 mg/l puromycin.

Colony Formation Assay

A549 cells that overexpressed WT, K9R, K35R, or control were seeded into 12-well plates (1000 cells/well) for 10 days. The cells were fixed by methanol and stained with 1% crystal violet, and the number of colonies was quantified (30).

Western Blotting

Western blotting was performed as previously described (31). Total proteins were isolated by radioimmunoprecipitation assay buffer (Cell Signaling Technologies) and quantified with a bicinchoninic kit (Thermo Fisher Scientific). The isolated proteins were separated by 10% SDS-PAGE prior to semitransfer to polyvinylidene fluoride membranes (Millipore). Next, the membranes were blocked for 1 h in 5% skimmed milk at room temperature and incubated with indicated primary antibodies at 4 °C overnight, followed by incubation with corresponding secondary antibodies for 1.5 h at room temperature. Ultimately, blot bands were visualized *via* ECL reagent (Bio-Rad), and images were obtained using Tanon 5200-Multi (Tanon Science & Technology Co Ltd).

Antibodies against the following proteins were used for the experiments: NUCKS1 was obtained from Cell Signaling Technology (Danvers) and ABclonal; GAPDH was obtained from Proteintech Group. The secondary antibodies including horseradish peroxidase–conjugated goat anti-rabbit antibody and horseradish peroxidase–conjugated goat antimouse antibody were purchased from Bioworld.

RESULTS

Establishment of LAFEM Based on the Molecular Environmental Features of Acetylation Sites in PSP and UniProt

All the data for training and testing utilized in this study were gathered prior to May 2022. Broadly speaking, acetylation sites with low-throughput experimentally validated functions ensure high credibility because these sites' functions were validated by experiments repeatedly. Therefore, we considered that the acetylation sites with low-throughput experimental evidence have high confidence directly relevant to fitness. The significantly different molecular environment among acetylation sites with and without low-throughput experimental evidence was the theoretical basis of LAFEM.

In brief, the establishment of LAFEM was divided into four procedures, including dataset preparation, feature extraction, model construction, testing and validation. In the first procedure, a total of 1280 (244 only from UniProt, 808 only from PSP, and 228 from both two databases) and 3621 (only from UniProt) acetylation sites with or without low-throughput experimental evidence were respectively labeled as positive samples (EXP) and negative samples (non-EXP). Each training dataset in 10-fold cross-validation contained 3259 negative samples (another 362 in internal testing dataset) and 1152 positive samples (another 128 in internal testing dataset). After undersampling in training datasets, equal amounts of 1152 acetylation sites with or without low-throughput experimental evidence were used to train model (Fig. 1*A*; Supplemental Table S1). Remaining 2107 acetylation sites without low-throughput experimental evidence (holdout set) and another 38 acetylation sites with low-throughput experimental evidence in disease file from PSP (external positive samples without training/testing) were combined to form validation dataset (Fig. 1*A*).

Then, we obtained features of acetylation site for its surrounding molecular environment by extracting from databases (PSP, UniProt, Interactome3D, and AlphaFold2) or computing. Totally, eight features including three physicochemical properties (disordered score, SASA score and hydrogen bonding–donating ability score) and five biological properties (region/motif distance, mutant site number, PTM occurrence frequency, intersurface distance, and homologous score) were derived to establish machine learning model (Fig. 1*A*). Using undersampling dataset, RF model with optimal performance in 10-fold cross-validation named LAFEM was selected to prioritize the functional acetylation sites (Fig. 1*A*). Ultimately, the performance of LAFEM was further tested by validation dataset, acetylome from CPTAC, and low-throughput experimental validation (Fig. 1*A*).

Environment Features of Acetylation Sites are Significantly Different Between EXP and Non-EXP

We defined seven amino acids before and after acetylation site as ASAS to investigate surrounding molecular environments of the site. The concrete explanations of the eight features were shown in methods. Initially, we compared three physicochemical properties including SASA score, disordered score, and hydrogen bonding–donating ability score of acetylation sites between EXP and non-EXP groups in training/testing dataset. The averaged SASA score in EXP group was 1506, which was significantly higher than 1145 in non-EXP group (Fig. 2*A*). Moreover, the averaged disordered score in EXP group was 0.502, which was significantly higher than 0.255 in non-EXP group (Fig. 2*B*). Next, the averaged hydrogen bonding–donating ability score in EXP group was 0.195, whereas it was only 0.096 in non-EXP group (Fig. 2*C*). Taken together, the surrounding molecular environment of acetylation sites with low-throughput experimental evidence showed common features in higher disorder degree, larger SASA, and stronger hydrogen bonding–donating ability.

After that, we compared five biological properties including homologous score, mutant site number (excluding artificial mutagenesis), PTM occurrence frequency, region/motif distance, and PPIS distance between EXP and non-EXP groups in training/testing dataset. We found that the averaged homologous score in EXP group was 1.50-fold higher than non-EXP group (Fig. 2*D*). We also investigated the mutant site number and PTM occurrence frequency in ASAS to evaluate functional acetylation site. We found that the average mutant site number was 0.42 in EXP group but only 0.21 in non-EXP group (Fig. 2*G*), and the average PTM occurrence frequency in EXP group was 4.76, higher than that in non-EXP group. The averaged $\log_2$(distance + 1) from acetylation site to region/motif and to intersurface were 2.96 and 3.46 in EXP group, which were lower than those in non-EXP group (Fig. 2, *E* and *F*). These results suggest that acetylation site with low-throughput experimental evidence is closer to region/motif and PPIS, and their surrounding molecular environments have higher homologous score, more mutant site number, and PTM occurrence frequency.

In brief, the molecular environment of fitness directly relevant acetylation sites had significant differences to others. It is meaningful to assess functional acetylation sites through molecular environment of acetylation sites.

Performance Comparison of Various Machine Learning Models and Sampling Methods

To construct the optimal machine learning model, we compared the performance of five machine learning models that encompassed GLMNET, KNN, RF, decision tree (recursive partitioning and regression tree), and SVM tested by 10-

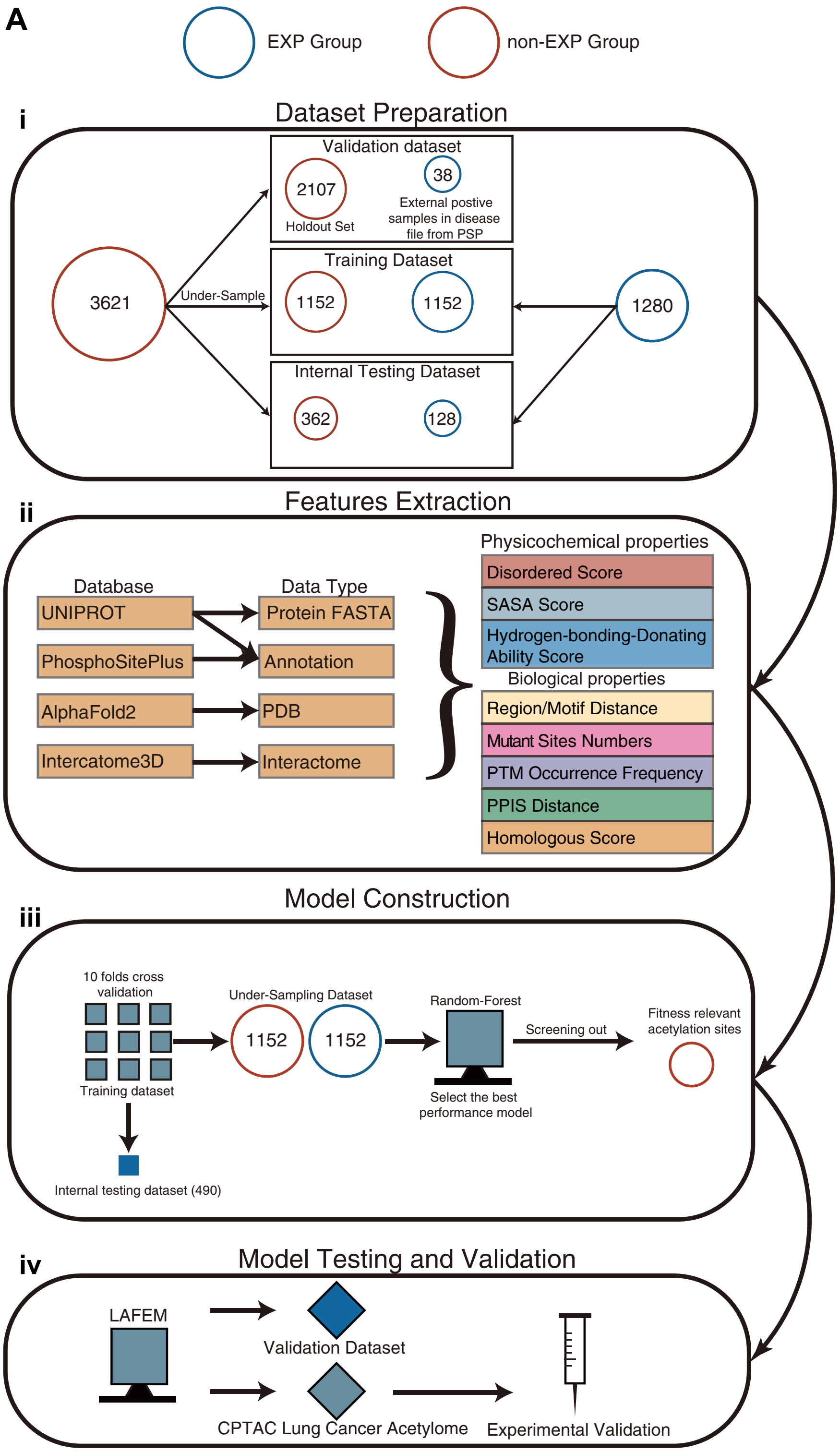


FIG. 1. **Establishment of LAFEM based on the molecular environmental features of acetylation sites in PSP and UniProt.** *A*, the workflow of the proposed methodology for LAFEM establishment: (i) dataset preparation; (ii) feature extraction; (iii) model construction; and (iv) model testing and validation. LAFEM, lysine acetylation functional evaluating model; PSP, PhosphositePlus; UniProt, universal protein.

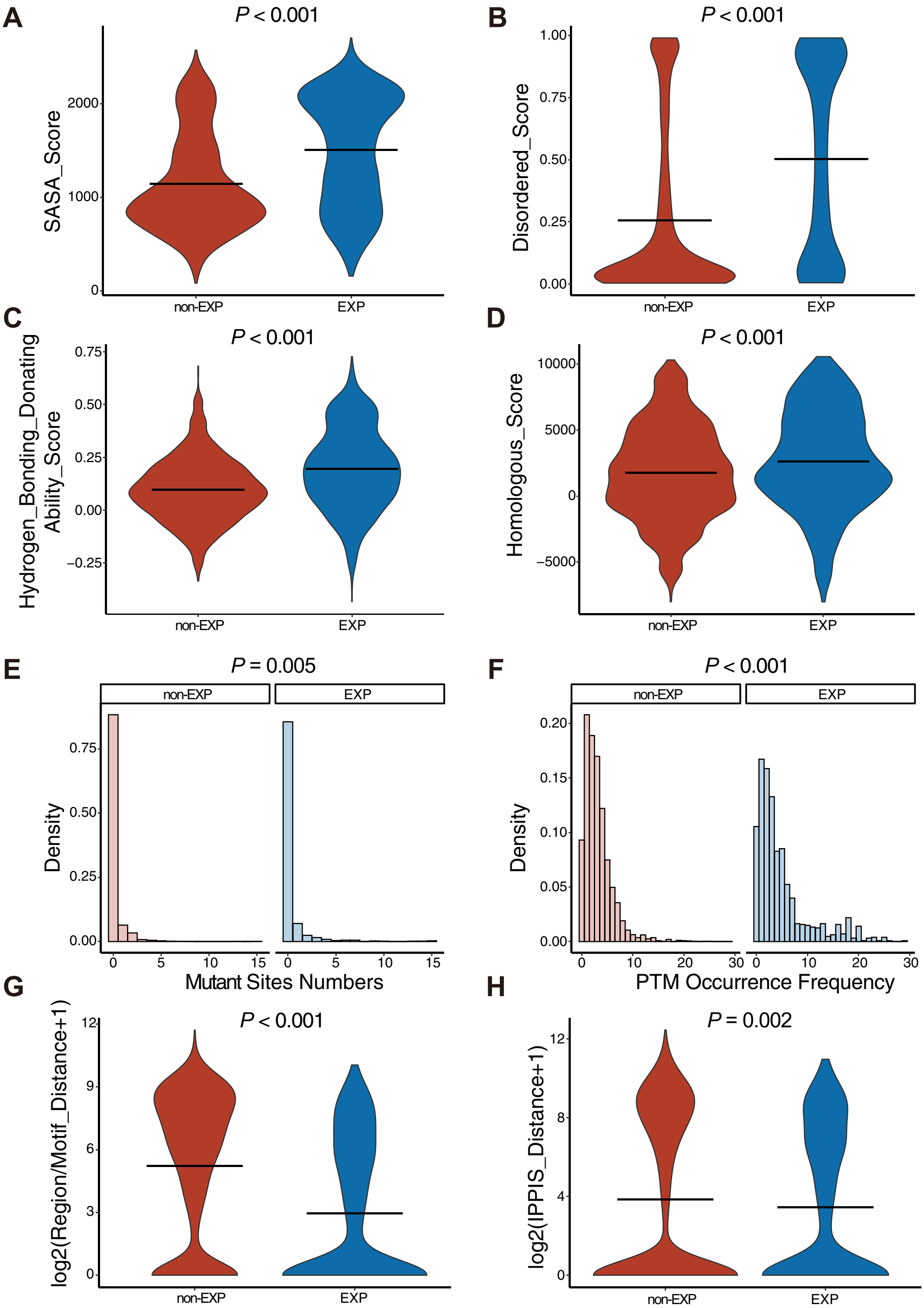


FIG. 2. **The environment features of acetylation sites are significantly different between EXP and non-EXP.** All acetylation sites labeled with EXP or non-EXP are compared for the SASA score (*A*), disordered score (*B*), hydrogen bonding–donating ability score (*C*), homologous score (*D*), mutant site number (*E*), PTM occurrence frequency (*F*), distance from acetylation site to region/motif (*G*), and distance from acetylation site to PPIS (*H*). All *p* values were calculated by two-tailed Mann–Whitney *U* test. PPIS, protein–protein intersurface; PTM, post-translational modification; SASA, solvent-accessible surface area.

fold cross-validation. The results showed that the median Matthews correlation coefficient (MCC) and area under the curve (AUC) of RF model were 0.466 ± 0.043 and 0.809 ± 0.021, respectively, which were the highest among these five models (Fig. 3, *A* and *B*; Supplemental Table S2). Hence, RF model was employed as an ideal machine learning model for training due to its high MCC and AUC.

In our study, we encountered a substantial discrepancy in quantity between the EXP and non-EXP datasets. This imbalance within the data classes may potentially hinder the efficacy of machine learning models. To address this issue, we preferred to select a suitable sampling method by comparing undersampling, oversampling, nonsampling, and synthetic minority oversampling techniques. Given that a significant proportion of acetylation sites remains devoid of functional annotations, we are supposed to avoid our model from missing too many fitness directly relevant acetylation sites. Hence, optimizing positive samples' detection rate under the premise of acceptable specificity emerged as a crucial performance criterion when selecting sampling methods. Upon evaluation using 10-fold cross-validation, models trained from all four datasets each reflecting a distinct sampling method consistently demonstrated high AUC and MCC (Fig. 3*C*; Supplemental Fig. S1*B*; Supplemental Table S3). However, the model trained by undersampling dataset displayed the highest sensitivity and the lowest absolute difference of sensitivity and specificity, indicating that it had the highest positive sample detection rate under the premise of high specificity (Fig. 3*D*; Supplemental Table S3). Taken together, the model trained by undersampling dataset displays higher AUC, MCC, and good positive sample detection rate, suggesting that

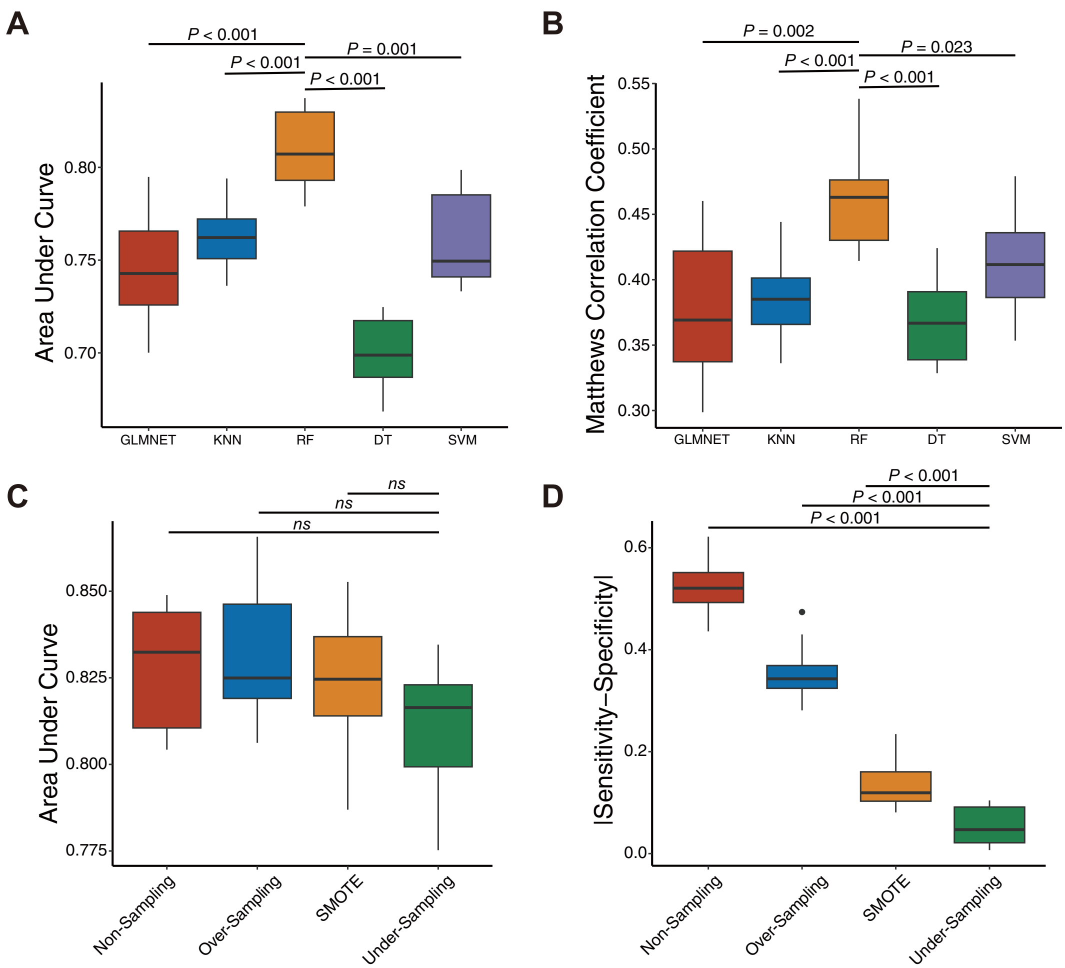


FIG. 3. **Performance comparison of various machine learning models and sampling methods.** *A* and *B*, the performances of five machine learning models including GLMNET, KNN, RF, RPART, and SVM tested by 10-fold crossvalidation were compared for their Matthews correlation coefficient (*A*) and area under the curve (*B*). *C* and *D*, the datasets respectively sampled by four sampling methods including nonsampling, oversampling, SMOTE, and undersampling were used to train RF models. The performances of four sampling methods were compared by area under the curve (*C*) and |specificity–sensitivity| (*D*) of models. *p* values of F3 *A*–*C* were used *right*-tailed Mann–Whitney *U* test, whereas F3 D's *p* value was used *left*-tailed Mann–Whitney *U* test (ns). GLMNET, generalized linear model *via* lasso and elastic-net regularization; KNN, K-nearest-neighbor; ns, not significant; RF, random forest; RPART, recursive partitioning and regression tree; SMOTE, synthetic minority oversampling techniques; SVM, support vector machine.

undersampling is a suitable sampling method for the following study.

Performance Assessment for LAFEM Prediction Based on Validation Dataset

Trained by undersampling dataset, the optimal RF model named LAFEM was used to calculate AFS, achieving 0.849 AUC (Fig. 4*A*) and 0.541 MCC (Fig. 4*C*) in internal testing dataset. AFS represents as the probability of the acetylation site classified into category fitness directly relevant acetylation sites (EXP) by LAFEM. Validation dataset was consisted of holdout set (2107 negative samples without training/testing) and 38 external positive samples in disease file from PSP. Tested by validation dataset, the AUC and MCC of LAFEM

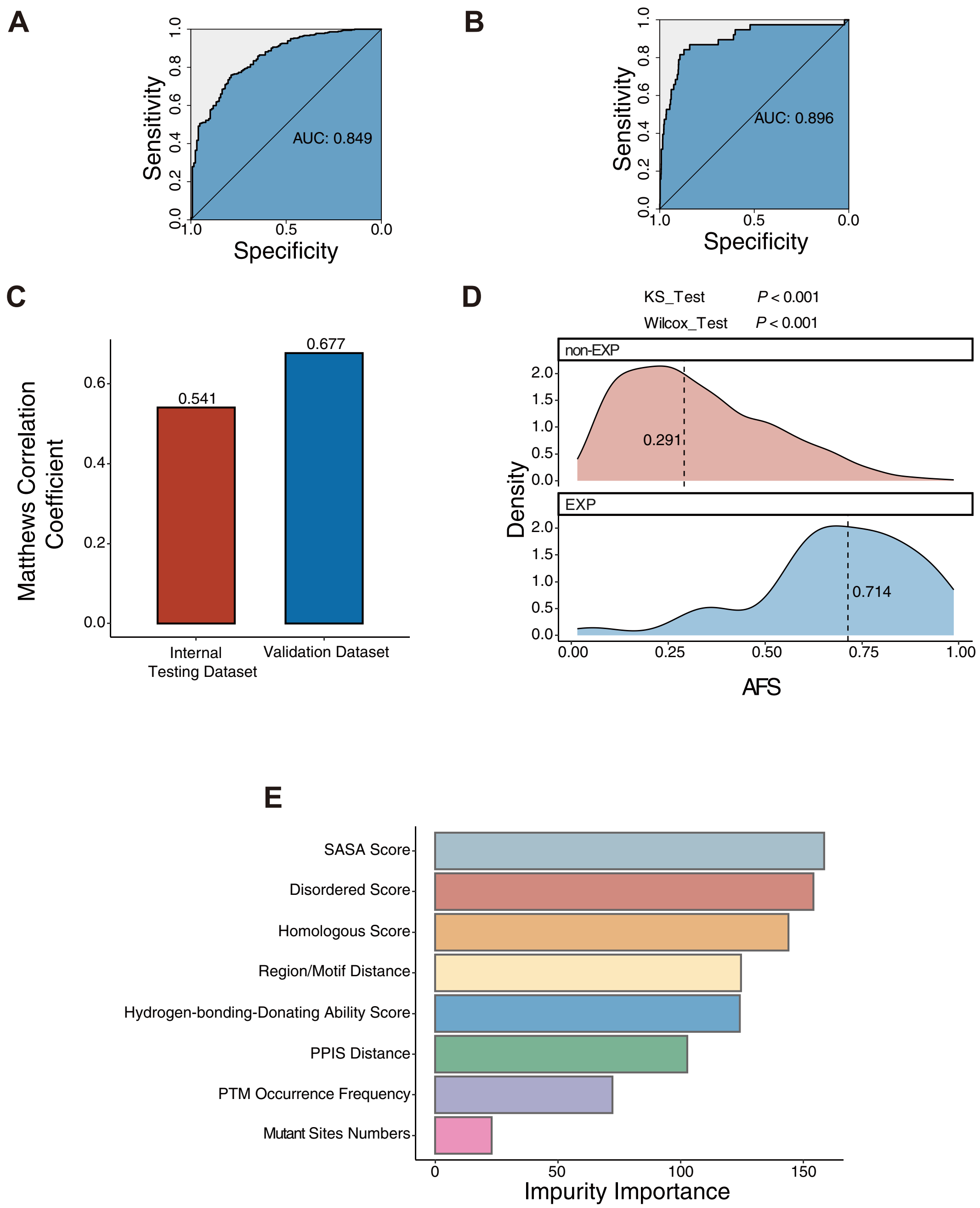


FIG. 4. **Performance assessment for LAFEM prediction based on validation dataset.** *A*, random forest model's receiver operating characteristic curve (ROC) was shown. *B*, the ROC of LAFEM tested by using validation dataset. *C*, the MCCs of internal testing dataset and validation dataset were exhibited; *p* value was performed using Kolmogorov–Smirnov test and two-tailed Mann–Whitney *U* test. *D*, the distribution of AFSs in EXP and non-EXP datasets. *E*, the impurity importance of eight features in LAFEM. AFS, acetylation functional score; LAFEM, lysine acetylation functional evaluating model; MCC, Matthews correlation coefficient.

were 0.896 (Fig. 4*B*) and 0.677 (Fig. 4*C*), higher than AUC and MCC of internal testing dataset. This suggests that the LAFEM is reliable since its good performance in validation dataset. Furthermore, in the validation dataset, the median AFS of 38 external positive samples (EXP) was 0.714 (Fig. 4*D*; Supplemental Table S4), significantly higher than the median AFS of holdout set (non-EXP, 0.291; Fig. 4*D*; Supplemental Table S5). These results further suggest that LAFEM had good accuracy to find the acetylation sites with direct relevance to fitness. After that, we prioritized the contributions of the eight features in LAFEM judged by Gini impurity measurement. Since their impurity importance, disordered score, homologous score, SASA score, distance from acetylation site to region/motif, and hydrogen bonding–donating ability score ranked in the top 5 (Fig. 4*E*), exhibiting that these features played core contribution to LAFEM and are markedly related to the functions of acetylation sites.

Performance of LAFEM on Acetylome in CPTAC

To further ask whether LAFEM can effectively screen out fitness directly relevant acetylation sites, we analyzed the acetylation sites from two CPTAC lung cancer acetylome datasets including PDC000233 (lung squamous cell carcinoma; Supplemental Table S6) and PDC000224 (lung adenocarcinoma; Supplemental Table S7) through LAFEM and differential expression analysis (Supplemental Table S8).

Totally, 12,600 acetylation sites (12,560 of which could be calculated AFS) were identified in the two datasets (Supplemental Table S9). There was insignificant difference in the AFSs between acetylome and the holdout set (Mann–Whitney *U* test $p > 0.05$; Fig. 4*D*; Fig. 5*B*). After differential expression analysis of two datasets, comparing cancer and adjacent cancer tissues, we separated the acetylation sites into four groups, which respectively were DEAS union (DEASU; 87 numbers), DEAS intersection (DEASI; 318 numbers), non-DEASU (559 numbers), and non-DEASI (Fig. 5*A*; 88 numbers). Interestingly, we found that the AFSs of DEASU and DEASI were significantly higher than acetylome, whereas the non-DEASU and DEASI were lower. Moreover, the median AFS from low to high were non-DEASI, non-DEASU, acetylome, DEASU, and DEASI. Therefore, differential analysis and LAFEM demonstrated each other, which could screen out fitness directly relevant acetylation sites (Fig. 5*B*).

However, the constructions of LAFEM and differential expression analysis are based on different methodologies hinting that the two methods may have complementary effects. According to our result, the median abundance of acetylation sites in two groups of acetylome had a low correlation coefficient to AFS. It validated that LAFEM and differential expression analysis assess functional acetylation sites through different methods, further implying that the complementary effect between LAFEM and differential expression analysis of quantification acetylome (Supplemental Fig. S3*B*). Hence, we considered that the combined usage of LAFEM and differential expression analysis can enhance the confidence to screen out fitness directly relevant acetylation sites in acetylome.

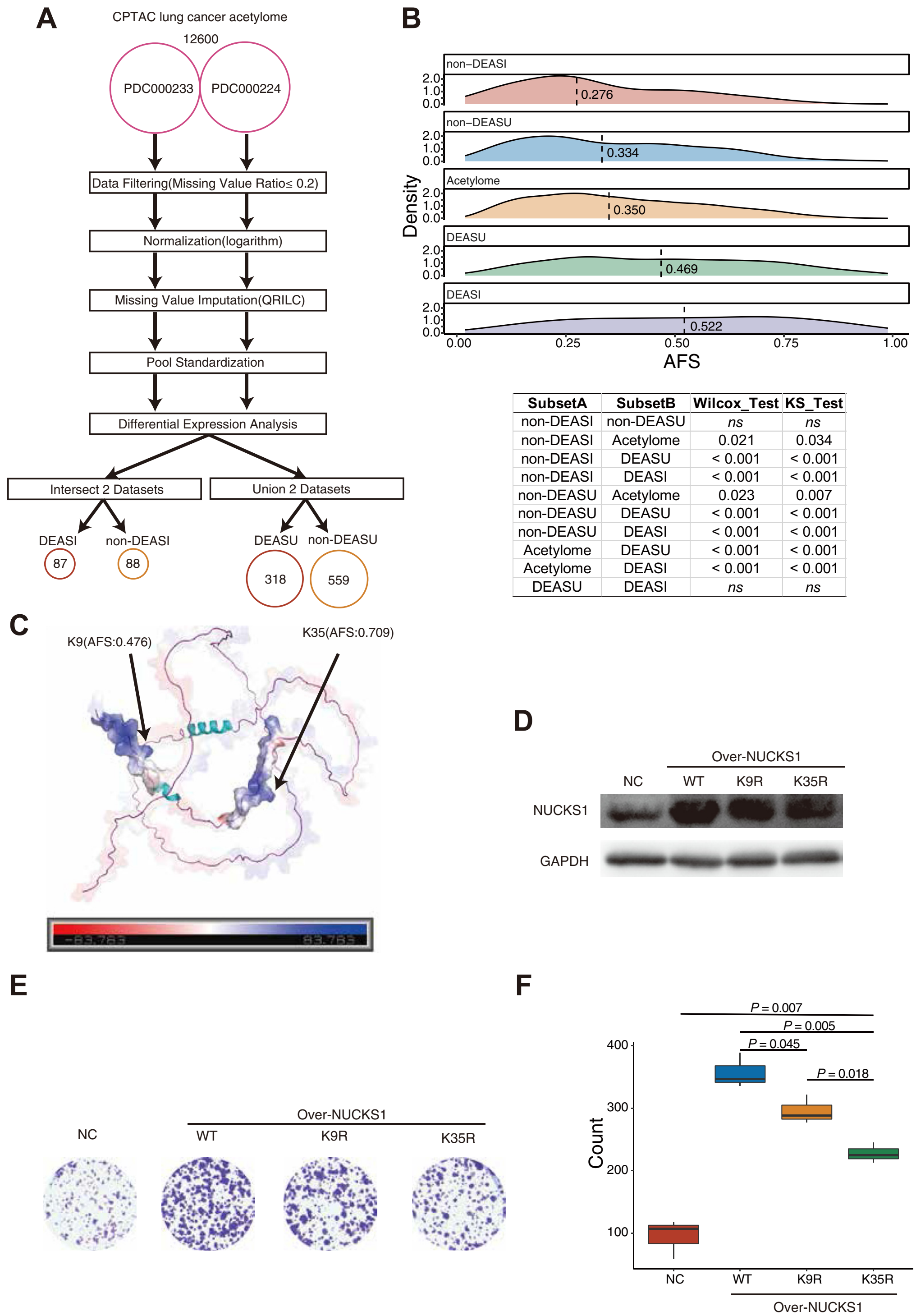
We then exemplified this approach by comparing the functional importance of two differentially overexpressed acetylation sites in lung squamous cell carcinoma and lung adenocarcinoma with different AFSs. The AFS of NUCKS1 ac-K35 was 0.709 and ac-K9 was 0.476 (Fig. 5*C*). Obviously, the ASAS of NUCKS1-K35 had significantly higher disordered degree score than K9 since there was an α-helix adjacent to K9 (Fig. 5*C*). In addition, the ASAS of NUCKS1-K35 had stronger hydrogen bond–donating ability than K9 (Fig. 5*C*). These differences of molecular environment were accountable to the higher AFS of ac-K35 than ac-K9 in NUCKS1.

To test the role of ac-K9 and ac-K35 of NUCKS1 in cancer proliferation, K9 and K35 were substituted as arginine (R), respectively. A549 cells with the lowest endogenous expression of NUCKS1 among several common lung cancer cell lines were overexpressed with WT of NUCKS1, K9R, or K35R, respectively (Fig. 5*D* and Supplemental Fig. S3*A*). According to the colony formation assay, we found that overexpression of NUCKS1-WT significantly promoted cell growth, whereas K9R and K35R attenuated this effect. Strikingly, K35R displayed stronger inhibitory effect on colony formation than K9R (Supplemental Table S10; Fig. 5, *E* and *F*), suggesting that ac-K35 at NUCKS1 with higher AFS is more important than ac-K9 in the proliferation of A549. These evidences fully demonstrate the accuracy of LAFEM to filter out fitness directly relevant acetylation sites.

DISCUSSION

We have developed LAFEM, the pioneering prioritization strategy designed to large-scale screen out fitness directly relevant acetylation sites. In this study, the acetylation sites annotated with and without low-throughput experimental evidence were labeled as positive and negative samples, respectively. For each ASAS, eight features that reflect the molecular environment surrounding acetylation sites were subjected to train. LAFEM offered a high-confidence list of 15,410 acetylation sites as ranked by AFS (Supplemental Table S11). The prioritization strategy constitutes an unprecedented resource to further develop new methods necessary to large-scale discover fitness directly relevant acetylation sites.

Advancements in MS have facilitated the large-scale identification of PTM sites. However, lack of functional annotations of these sites handicaps their low-throughput experimental validation. Moreover, relying solely on one dimension of expression difference is insufficient to screen out fitness directly relevant PTM sites in PTM proteome. Therefore, it is essential to develop a prioritization strategy to evaluate the functional importance of PTM site through attributes of PTM site molecular surrounding environment unrelated to



SubsetA	SubsetB	Wilcox_Test	KS_Test
non-DEASI	non-DEASU	*ns*	*ns*
non-DEASI	Acetylome	0.021	0.034
non-DEASI	DEASU	< 0.001	< 0.001
non-DEASI	DEASI	< 0.001	< 0.001
non-DEASU	Acetylome	0.023	0.007
non-DEASU	DEASU	< 0.001	< 0.001
non-DEASU	DEASI	< 0.001	< 0.001
Acetylome	DEASU	< 0.001	< 0.001
Acetylome	DEASI	< 0.001	< 0.001
DEASU	DEASI	*ns*	*ns*

FIG. 5. **Performance of LAFEM on acetylome in CPTAC.** *A*, the workflow of the generation processes of DEASI, non-DEASI, DEASU, non-DEASU, and identified acetylation sites in original acetylome. *B*, the AFSs' distributions of DEASI, DEASU, non-DEASI, non-DEASU, and original acetylome were shown in histograms; *p* value was performed using Kolmogorov–Smirnov test and two-tailed Mann–Whitney *U* test (ns). *C*, structure of NUCKS1 was shown by PyMOL (Protein Data Bank: AF-Q9H1E3-F1-model_v2.pdb). The color of protein surface from *red* to *blue* represents electric potential from low to high. At the perspective of secondary structure, random coil and helix were respectively stained as *purple* and *turquoise*. *D*, Western blot of NUCKS1 expression in A549 overexpressed with WT-NUCSK1, K9R-NUCKS1, and K35R-NUCKS1. *E*, the colony formation ability of A549 overexpressed with WT-NUCSK1, K9R-NUCKS1, and K35R-NUCKS1. *F*, the colony number was statistically plotted; *p* value was performed using Student's *t* test. AFS, acetylation functional score; CPTAC, Clinical Proteomic Tumor Analysis Consortium; DEASI, differential expression acetylation site intersection; DEASU, differential expression acetylation site union; LAFEM, lysine acetylation functional evaluating model; ns, not significant; NUCKS1, nuclear casein kinase and cyclin-dependent kinase substrate.

expression. This will aid in identifying fitness directly relevant PTM sites and uncovering the core reasons behind differences in biological models at the aspect of PTM. It was reported that the physicochemical and biological features of PTM sites determine PTM sites' surrounding environments, and these features show comparable importance with the quantitative information from MS (12). Incorporating these features of phosphosites into a machine learning model enhanced the functional interpretation of identified phosphosites according to Ochoa's article (13).

Hence, it is necessary to construct a machine learning model to evaluate functional acetylation sites. Selecting an appropriate machine learning model and sampling method is crucial for the construction of LAFEM. The RF model generally outperforms other classifiers in terms of stability and efficiency because of its inherent advantages. Notably, among the five machine learning models we tested, the RF model displayed the best results, evidenced by its higher AUC and MCC values. Because of the absence of functional annotations for the vast majority of acetylation sites, we are supposed to avoid our model from missing too many fitness directly relevant acetylation sites. Therefore, it is essential to maximize the positive sample detection rate while maintaining acceptable specificity. We ascertained that the model trained by the undersampling dataset had high AUC, MCC, and the most balanced sensitivity and specificity, implying that undersampling was the most suitable sampling method in this study. All in all, we finally choose RF model to be the classifier and undersampling to sample.

Large SASA or high disordered degree may significantly influence protein–protein interactions (32). In our study, we found that the surrounding environment of known functional acetylation sites displayed larger SASA and higher disordered degree. Hydrogen bonding–donating ability reflects the electronegativity of acetylation site's environment. Addition of acetyl group changes the electronegativity of the site's environment, leading to the alteration of protein conformation. In this work, we discovered that the functional acetylation sites displayed stronger hydrogen bonding–donating ability. Homology is the most common indicator to estimate the functional conservation of site or sequence for proteins (33). More crossspecies conservation implies higher importance in regulating biological processes during evolution (34). This elucidates the significant contribution of the homology feature to LAFEM.

While enzyme–substrate interactions are crucial for identifying acetylation functions, unlike phosphorylation, there are fewer studies on acetylation and deacetylation enzymes with their substrates and a lack of applicable software. The study of Schölz *et al.* (35) identified the acetylomes of 20 deacetylation inhibitors, which is a remarkable work. However, the specificity of these deacetylation inhibitors is debatable. Moreover, we attempted to analyze acetylase-regulated motifs through MEME through this study. Nevertheless, only 11 acetylation sites met the predicted regulation motif for histone deacetylase family among 683 candidate acetylation sites (fold change > 1.5 at least one histone deacetylase inhibitor treatment). In addition, three met the predicted regulation motif for SIRT family among 464 candidate sites. The low coverage suggested that acetylase substrate–related feature is not appropriate for large-scale evaluation of functional importance of acetylome (Supplemental Fig. S4, *A* and *B*). The other two features that are not modeled are PTM distance and mutant site distance, and these two features of detailed descriptions are shown in Supplemental Table S12.

Importantly, we demonstrated that LAFEM performed well on the validation dataset and acetylome, which determined that LAFEM has good accuracy to screen out fitness directly relevant acetylation sites. Meanwhile, the low correlation coefficient between AFS and abundance of acetylation sites in acetylome showed that LAFEM assessed functional acetylation sites not based on quantification acetylome (Supplemental Fig. S3*B*). Therefore, the two methods are complementary when evaluating functional acetylation sites. Their combined usage can evaluate functional acetylation sites more comprehensively, and it may improve the confidence to filter out fitness directly relevant acetylation sites for further investigation.

At present, insufficient attention has been paid to the study of negative dataset, coupled with the lack of detailed functional annotations in most acetylation sites, making it difficult to conduct large-scale research on these acetylation sites. Though there are acetylation sites not directly relevant to fitness in negative dataset, the reason caused the absence of evidence for acetylation site's function is complex, so it is hard to define them precisely. We recommend that preferentially studying the acetylation fitness directly relevant acetylation sites screened out by LAFEM, but it does not deny that other acetylation sites are nonfunctional or without studying significance.

Regrettably, LAFEM still has some limitations as shown in this study, even though LAFEM performed well in screening out fitness directly relevant acetylation sites. First, it may bring bias in a certain when studying for functional acetylation sites that are hard to detect. Second, it was trained by acetylation sites of human; therefore, we cannot assure its accuracy in other species. Then, it evaluates the functional importance of acetylation sites, but it cannot determine the condition under which the acetylation site performs its function. Hence, we suggest that LAFEM should be used in conjunction with other functional screening methods like differential expression analysis or functional hypothesis of acetylation sites. In addition, LAFEM is not suitable for studying acetylation sites on isoforms because Interactome3D does not consider the interactions of isoforms, and UniProt database has lack of information on isoforms. We recommended establishing a method to not rely on a database to study acetylation on isoforms when more functional and nonfunctional acetylation sites are validated.

Not all acetylation sites are likely to contribute equally to fitness, which handicaps the discovery of core acetylation sites (11). LAFEM is the first prioritization strategy not relying on quantification acetylome to solve this difficulty as it can effectively screen out fitness directly relevant acetylation sites. The good performance of LAFEM also sets an example for constructing large-scale prioritization strategies for other PTMs. AFSs of 15,410 acetylation sites (Supplemental Table S11) further enrich their functional annotations, which provides the convenience for researchers to carry out their molecular experiment and find out the core reason of difference between biological models at the perspective of acetylome.

DATA AVAILABILITY

The LAFEM application, the commands of BLAST+ to construct database and BLAST, SASA score of 23,391 proteins, quantitative acetylome results (PDC000233 and PDC000224) of MaxQuant, DEAS and non-DEAS, respectively, in the two acetylomes, diso files (generated by DISOPRED3 to calculate disordered score) of all proteins mentioned in our study, Protein Data Bank files downloaded from AlphaFold2 and table of PPIS downloaded from Interactome3D were uploaded to Zenodo (10.5281/zenodo.10132537).

Supplemental data—This article contains supplemental data.

Funding and additional information—This work was supported by the National Key Research and Development Program of China (no. 2022YFA1304604 and 2020YFE0202200), the National Natural Science Foundation of China (no. 82103208), the Guangdong Basic and Applied Basic Research Foundation (2023A1515030115).

Author contributions—J.-Z. L. and D.-H. L. conceptualization; J.-Z. L., D.-H. L., T. Z., and Q.-Y. L. methodology; J.-Z. L., D.-H. L., and Y.-C. X. validation; J.-Z. L., D.-H. L., T. Z., and Q.-Y. L. formal analysis; J.-Z. L. and F.-J. S. investigation; Q.-Y. H. and Y. W. resources; J.-Z. L. and F.-J. S. data curation; J.-Z. L. writing–original draft; Q.-Y. H. and Y. W. writing–review & editing; Q.-Y. H., Y. W., J.-Z. L., and D.-H. L. visualization; Q.-Y. H. and Y. W. supervision; Q.-Y. H. funding acquisition.

Conflict of interest—The authors declare no competing interests.

Abbreviation—The abbreviations used are: AFS, acetylation functional score; ASAS, acetylation site adjacent sequence; AUC, area under the curve; CPTAC, Clinical Proteomic Tumor Analysis Consortium; DEAS, differential expression acetylation site; DEASI, DEAS intersection; DEASU, DEAS union; GLMNET, generalized linear model via lasso and elastic-net regularization; IDR, intrinsically disordered region; KNN, K-nearest-neighbor; LAFEM, lysine acetylation functional evaluating model; MCC, Matthews correlation coefficient; NUCKS1, nuclear casein kinase and cyclin-dependent kinase substrate 1; MS, mass spectrometry; PPIS, protein–protein intersurface; PSP, PhosphositePlus; PTM, post-translational modification; RF, random forest; SASA, solvent-accessible surface area; SVM, support vector machine; UniProt, universal protein.

Received February 28, 2023, and in revised form, November 18, 2023 Published, MCPRO Papers in Press, December 15, 2023, https://doi.org/10.1016/j.mcpro.2023.100700

REFERENCES

1. Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (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
2. Narita, T., Weinert, B. T., and Choudhary, C. (2019) Functions and mechanisms of non-histone protein acetylation. *Nat. Rev. Mol. Cell Biol.* **20**, 156–174
3. Wang, Y., Zhang, J., Li, B., and He, Q. Y. (2019) Advances of proteomics in novel PTM discovery: applications in cancer Therapy. *Small Methods* **3**, 1900041
4. Tessarz, P., and Kouzarides, T. (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat. Rev. Mol. Cell Biol.* **15**, 703–708
5. Wang, R., and Wang, G. (2019) Protein modification and autophagy activation. *Adv. Exp. Med. Biol.* **1206**, 237–259
6. Weinert, B. T., Iesmantavicius, V., Moustafa, T., Scholz, C., Wagner, S. A., Magnes, C., *et al*. (2014) Acetylation dynamics and stoichiometry in Saccharomyces cerevisiae. *Mol. Syst. Biol.* **10**, 716
7. Weinert, B. T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B. K., Scholz, C., *et al*. (2018) Time-resolved analysis reveals Rapid dynamics and Broad Scope of the CBP/p300 acetylome. *Cell* **174**, 231–244.e212
8. Hornbeck, P. V., Kornhauser, J. M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., *et al*. (2019) 15 years of PhosphoSitePlus(R): integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.* **47**, D433–D441
9. Beltrao, P., Bork, P., Krogan, N. J., and van Noort, V. (2013) Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* **9**, 714
10. Kanshin, E., Bergeron-Sandoval, L. P., Isik, S. S., Thibault, P., and Michnick, S. W. (2015) A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep.* **10**, 1202–1214
11. Landry, C. R., Levy, E. D., and Michnick, S. W. (2009) Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197
12. Bludau, I., Willems, S., Zeng, W. F., Strauss, M. T., Hansen, F. M., Tanzer, M. C., *et al*. (2022) The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **20**, e3001636
13. Ochoa, D., Jarnuczak, A. F., Vieitez, C., Gehre, M., Soucheray, M., Mateus, A., *et al*. (2020) The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373
14. Lee, T. Y., Hsu, J. B., Lin, F. M., Chang, W. C., Hsu, P. C., and Huang, H. D. (2010) N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *J. Comput. Chem.* **31**, 2759–2771
15. Kurotani, A., and Sakurai, T. (2015) In Silico analysis of correlations between protein disorder and post-translational modifications in Algae. *Int. J. Mol. Sci.* **16**, 19812–19835
16. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., *et al*. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
17. Mitternacht, S. (2016) FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res.* **5**, 189
18. Jones, D. T., and Cozzetto, D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863
19. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421

20. Pearson, W. R. (2013) An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics* **Chapter 3**, 3.1.1–3.1.8
21. Lifschitz, S., Haeusler, E. H., Catanho, M., Miranda, A. B., Armas, E. M., Heine, A., *et al*. (2022) Bio-strings: a relational database data-type for dealing with large biosequences. *BioTech (Basel)* **11**, 31
22. Hedges, S. B., Dudley, J., and Kumar, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972
23. Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., *et al*. (2022) TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174
24. Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R., and Clementi, S. (2004) Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemometr.* **18**, 146–155
25. Mosca, R., Ceol, A., and Aloy, P. (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53
26. Friedman, J., Hastie, T., and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22
27. Chanal, D., Yousfi Steiner, N., Petrone, R., Chamagne, D., and Péra, M.-C. (2022) Online Diagnosis of PEM Fuel cell by Fuzzy C-means clustering. In: Cabeza, L. F., ed. *Encyclopedia of Energy Storage*, Elsevier, Oxford: 359–393
28. Misra, S., and Li, H. (2020) Noninvasive fracture characterization based on the classification of sonic wave travel times *Machine Learning for Subsurface Characterization*. Elsevier, Amsterdam, NH: 243–287
29. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
30. Zhang, J., Zhou, Y., Li, N., Liu, W. T., Liang, J. Z., Sun, Y., *et al*. (2020) Curcumol overcomes TRAIL resistance of non-small cell lung cancer by targeting NRH: Quinone oxidoreductase 2 (NQO2). *Adv. Sci. (Weinh)* **7**, 2002306
31. Wang, Y., Zhang, J., Li, Y. J., Yu, N. N., Liu, W. T., Liang, J. Z., *et al*. (2021) MEST promotes lung cancer invasion and metastasis by interacting with VCP to activate NF-kappaB signaling. *J. Exp. Clin. Cancer Res.* **40**, 301
32. Chakravarty, D., Janin, J., Robert, C. H., and Chakrabarti, P. (2015) Changes in protein structure at the interface accompanying complex formation. *IUCrJ* **2**, 643–652
33. Fuertes, M. A., Rodrigo, J. R., and Alonso, C. (2019) Conserved critical evolutionary gene structures in orthologs. *J. Mol. Evol.* **87**, 93–105
34. Wang, J., Wu, G., Chen, L., and Zhang, W. (2013) Cross-species transcriptional network analysis reveals conservation and variation in response to metal stress in cyanobacteria. *BMC Genomics* **14**, 112
35. Scholz, C., Weinert, B. T., Wagner, S. A., Beli, P., Miyake, Y., Qi, J., *et al*. (2015) Acetylation site specificities of lysine deacetylase inhibitors in human cells. *Nat. Biotechnol.* **33**, 415–423