Original article

CREWdb: Optimizing Chromatin Readers, Erasers, and Writers Database using Machine Learning-Based Approach

Maya Natesan[1], Mitchell Kong[2], Min Shi[3], Reetika Ghag[1], Shamim A Mollah[1,3*]

[1]Institute for Informatics, Washington University School of Medicine in St. Louis, 600 Euclide Ave. MO 63110

[2]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093

[3]Department of Genetics, Washington University School of Medicine in St. Louis, 600 Euclide Ave. MO 63110

*To whom correspondence should be addressed.

Abstract

Aberration in heterochromatin and euchromatin states contributes to various disease phenotypes. The transcriptional regulation between these two states is significantly governed by post-translational modifications made by three functional types of chromatin regulators: readers, writers, and erasers. Writers introduce a chemical modification to DNA and histone tails, readers bind the modification to histone tails using specialized domains, and erasers remove the modification introduced by writers. Altered regulation of these chromatin regulators plays a key role in complex diseases such as cancer, neurodevelopmental diseases, myocardial diseases, and embryonic development. Due to the reversible nature of chromatin modifications, we have the opportunity to develop therapeutic approaches targeting chromatin regulators. Currently, a limited amount of chromatin regulators have been identified, and a subset of those identified have been ambiguously classified as multiple chromatin regulator types. Thus, we have applied machine learning-based approaches to predict and classify the functionality of chromatin regulator proteins, optimizing the first comprehensive database of chromatin regulators known as *CREWdb*.

Database URL: Source codes are implemented in Python and are freely available for download at: https://github.com/smollah/CREWdb

Introduction

DNA serves as the biological basis for cellular replication, transcription, and translation—processes that govern gene expression in cells. While cells contain identical genetic material, their ability to maintain differential gene expression and unique functionality is due mainly to the regulation of their DNA by epigenetic modifiers. Epigenetic modifications performed by catalyzing enzymes are key players in gene expression patterns during cell differentiation and development [1].

In eukaryotic cells, DNA wraps around proteins known as histones, forming condensed coils. Epigenetic modifiers act on these newly translated histone proteins, influencing chromatin accessibility and gene expression [2]. There are two main types of chromatin remodeling: histone modifications and ATP-dependent chromatin remodeling complexes. The former is a key driver in many known diseases such as acute myeloid leukemia, breast cancer, neurodegenerative disorders, etc [3]. These histone modifications cause changes in DNA accessibility by neutralizing or introducing charges on chromatin, thereby weakening or strengthening nucleosome interactions between histone groups [4]. Stronger interactions condense the DNA and result in transcriptional repression while weaker interactions uncoil the DNA and result in transcriptional activation.

The enzymes responsible for the condensation and uncoiling of chromatin are chromatin regulators, which are characterized as readers, erasers, or writers (REW). Writers introduce functional groups to DNA and histones and include those under histone methyltransferase and histone acetyltransferase domains [5]. Readers contain specialized domains that recognize and interpret these functional groups and include proteins that contain methyl-lysine-recognition motifs such as bromodomains and chromodomains [6]. And, erasers remove the chemical modifications introduced by writers and include those under histone demethylase and histone deacetylase domains. As epigenetic drivers, chromatin regulators play critical roles in the regulation of gene expression. Therefore, a comprehensive understanding of chromatin regulators is crucial for further research and development in treatment therapies for a significant number of diseases including cancers and genetic disorders.

However, current efforts to compile descriptors of epigenetic modifiers are limited. Many studies focus on specific markers and enzymes [7] or tissue-specific cancers [6] rather than all chromatin regulators across all diseases and subtypes. In addition, the available databases that aim to address this problem are currently incomplete, lack coverage, or are inconsistent [8]. Thus, our database, the Chromatin Readers, Erasers, and Writers database (CREWdb) aims to mitigate this problem by providing a comprehensive collection of chromatin regulators and their molecular features across multiple diseases. We initially curated regulators from various known sources from publicly available databases and published studies including CR2Cancer [9], WERAM DB [10], Gillette [6], Nicholson [11], and Hyun [2], and designed a web interface for researchers to gain more information about the functionality and attributes of their chromatin regulator protein of interest. In this study, we aim to further optimize the CREWdb by developing a machine learning pipeline to predict and classify proteins as readers, writers, and erasers given disease-specific features. As a starting point, we focused on predicting chromatin regulator roles in cancer.

This paper will (1) discuss the preliminary development of CREWdb and the web interface (2) outline the proposed machine learning pipeline including data acquisition and preprocessing, feature selection, training and testing (3) demonstrate the machine learning component of the chromatin regulator classification model for the CREW database.

Material and methods

We adopted a well-established data pre-processing and machine learning pipeline to optimize the CREWdb. We integrated curated datasets, performed feature enrichment through feature selection and class rebalancing techniques, and implemented machine learning models to classify proteins as readers, writers, or erasers.

Preliminary Development of CREWdb

Preliminary processing for the CREW database was first performed on several sets of data obtained from established literature sources, including publicly available databases and published research studies. The databases include CR2Cancer, CCLE, WERAM DB, and TCGA, and the studies include Nicholson, Hyun, and Gillette—all of which contain gene names, functional classifications, and validated chromatin reader, writer, and eraser designations. A script mapping was generated to match CREW characterizations to corresponding genes, resulting in 1000 overall recorded CREW-validated proteins. A brief description of CREWdb construction and each data source is described below.

Construction of CREWdb and Web Interface

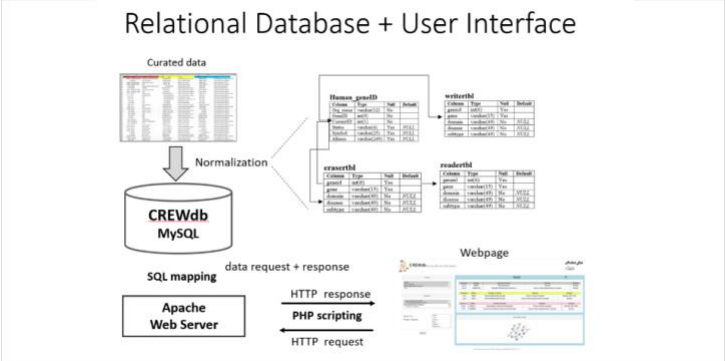

Fig.1: Architecture diagram.

Figure 1 describes the architecture of the CREWdb application. Data was curated from various sources in the form of a paf file containing chromatin regulator proteins and their metadata. The curated data was stored in a relational MySQL database (MariaDB): the CREWdb. The relational database was normalized to maintain the relational constraints (ACID properties) [12]. A web application was built using PhpMyAdmin and hosted on an Apache web server that connects the database and a web-based user interface. This facilitates the dissemination of chromatin regulator information to the end-users of the biomedical communities.

CR2Cancer:

CR2Cancer is a database released in 2017 and contains genomic, transcriptomic, proteomic, clinical, and functional information on over 400 chromatin regulators across various cancer types [9].

The list of 429 chromatin regulators was generated from three papers [13, 7, 8], and among the four-hundred twenty-nine chromatin regulators, one-hundred sixty-seven were mapped to cancer genes in the ONGene, TSGene, and NCG databases. The full name, aliases, chromosome location, and external IDs were annotated. In addition, somatic mutation, RNA-seq, DNA copy number, methylation, RPPA,

and clinical data were derived from TCGA; the site-specific substrates and products were compiled derived from EpiFactors; mutation profiles and gene expression microarray of cancer cell lines from CCLE; and the RNA-Seg and antibody staining data of human normal tissues from HPA.

For each gene, CR2Cancer provides the (i) functional category including protein domain, description, substrates, and products of chromatin regulator specific reactions (ii) mutation category including the mutational landscape of primary tumor tissues (iii) modification category including post-translational modifications from UniProt (iv) RNA expression (v) correlational analysis between RNA expression and semantic copy number alteration (vi) correlational analysis between RNA expression and promoter methylation level (vii) clinical category (viii) proteomic (ix) target category (x) drug and gene-centric networks (xi) and protein-protein interaction as well as other interactions derived from literature sources.

CCLE:
The Cancer Cell Line Encyclopedia, founded in 2008, provides genetic characterizations of human cancer cell lines, providing detailed genomic analyses and visualizations across multiple genes and corresponding cancer cell lines [14]. The CCLE contains sequencing data aligned to the hg19 broad variant reference genome for nearly 1000 cancer cell lines, including RNA seq, Whole Genome Seq, and Whole Exome Seq. The database provides a user-friendly tool to access and download sequencing data across multiple public projects.

WERAM DB:
The Writers, Erasers, Readers of Acetylation and Methylation proteins database provides 584 non-redundant protein data of 8 organisms including *H. sapiens, M. musculus, R. norvegicus, D. melanogaster, C. elegans, A. thaliana, S. pombe* and *S. cerevisiae* from the literature. The data are further classified into 15 families for histone acetylation writers, erasers, and readers and 32 families for histone methylation writers, erasers, and readers, respectively [10].

584 histone regulators were obtained from scientific literature published after 2011. The papers were selected through PubMed Advanced Search Builder and specific filters were applied for histone acetylation and histone methylation. From the extracted histone regulators, protein domain information was obtained from annotations in the UniProt database and functional domain information was obtained from the Pfam database. The database provides a user-friendly interface where users can search by species (animals, plants, fungi) or classification (histone acetylation or histone methylation).

Machine Learning
In order to optimize the preliminary CREW database described above, we collected additional data on biologically relevant features and developed a machine learning pipeline to predict chromatin regulator functionality as shown in Figure 2.

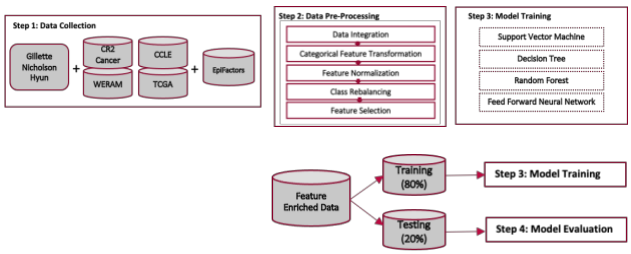

Fig.2: Diagram of overall machine learning pipeline.

Additional data for the machine learning pipeline was extracted from the EpiFactors database. This data curation consisted of 10 unique epigenetic features for 721 genes. The ten selected features included (1) HGNC approved name, (2) Pfam domain, (3) HCNC gene family tag, (4) HCNC gene family designation, (5) functional classification, (6) associated histone modification, (7) associated protein complex, (8) target molecule, (9) target entity, and (10) product. A brief description of the EpiFactors database is included below.

EpiFactors:
EpiFactors is a database released in 2015 which contains information on epigenetic regulators, their complexes, targets, and products across 458 human primary cell samples, for 815 proteins, including 95 histones and protamines [15].

The database includes genes that are specifically defined as epigenetic factors by the founders and include proteins that act as histones, histone variants, or protamines, perform post-translational modifications or can recognize those modifications, change the general structure of chromatin, assist histone folding and assembly, act upon modifications of DNA or RNA, or protein cofactors forming complexes with epigenetic factors [15].

Its data was curated from multiple sources including the Histone Infobase and relevant research papers and reviews. Histones, protamine, their modifying enzymes, DNA methylation enzymes, and chromatin-remodeling proteins were obtained from UniProt through keyword searches. Information on protein complexes was manually extracted from the 'function' field in Uniprot. Each protein documented within Epifactors has been cross-checked with published validations [15].

For each gene, EpiFactors provides (i) HGNC approved symbol (ii) HGNC approved name (iii) UniProt ID (human) (iv) Pfam domains (v) UniProt ID (mouse) (vi) functional classification (vii) histone modification (viii) protein complex (ix) target entity (x) product. The database provides a user-friendly tool to access information on human proteins and complexes involved in epigenetic regulation.

Data Integration
Out of the 721 genes extracted from the Epifactors database, 446 genes were matched to the genes mapped to CREW characterizations. The CREW classifications and 10 epigenetic features were then integrated for the subset of 446 genes, forming the primary dataset for the regulator classification model shown in Figure 3.

	ID (REMOVE)	SYMBOL	REW	HGNC approved name	Pfam domains	HGNC gene family tag	HGNC gene family description	Function	Modification	Protein complex	Target molecule	Target entity	Product
0	10389	SCML2	Reader	sex comb on midleg-like 2 (Drosophila)	DUF3588 PF12140 353-468, MBT PF02820 67-139 17...	SAMD	Sterile alpha motif (SAM) domain containing	Polycomb group (PcG) protein	#	PRC1	#	#	#
1	22955	SCMH1	Reader	sex comb on midleg homolog 1 (Drosophila)	DUF3588 PF12140 355-469, MBT PF02820 62-134 17...	SAMD	Sterile alpha motif (SAM) domain containing	Polycomb group (PcG) protein	#	PRC1	#	#	#
2	5252	PHF1	Reader	PHD finger protein 1	Mtf2_C PF14061 525-566, PHD PF00628 89-142, Pf...	TDRD, PHF	Tudor domain containing, Zinc fingers, PHD-type	Polycomb group (PcG) protein	#	PRC2	#	#	#

Fig.3: Table of features for the integrated training dataset.

Data Transformation
Preliminary data pre-processing included nominal encoding and feature normalization. Categorical feature transformation was performed by mapping categorical values such as domain names and family tags to numerical values. Then, feature normalization was performed by using min-max normalization.

To obtain stable prediction models, all features were linearly normalized by the min-max formula shown in Figure 4.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Fig.4: Equation for min-max normalization.

The variable x_i represents the feature value of the $i^{(th)}$ sample. The *min(x)* and *max(x)* are the minimum and maximum across all feature values under feature *x*. After data transformation steps were performed, class rebalancing was implemented.

Class Distribution Rebalancing

The class distribution of readers (136 samples), writers (160 samples), and erasers (150 samples) were imbalanced in the dataset as shown in Figure 5. To mitigate the risk of inaccurate predictions, we performed a class distribution rebalancing process using a synthetic minority over-sampling technique (SMOTE) [16].

SMOTE is an over-sampling method in which the minority class is over-sampled by creating "synthetic" samples. By taking each minority class sample and introducing synthetic samples along the line segments joining any or all k minority class nearest neighbors, the number of samples in the minority class increases to match the number of samples in the majority class. For the reader and eraser categories, 24 and 55 samples were generated, respectively, to match the number of samples in the writer's category.

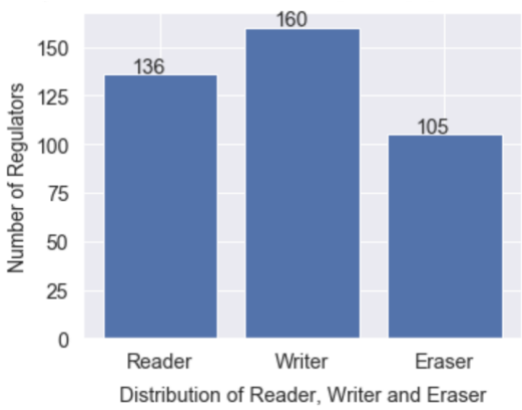

Fig.5: Class distribution of numbers of the readers, writers, and erasers in the studied data for ML.

Feature Selection

In the final pre-processing step, feature selection was performed on the eleven features to select a subset of highly confident features. We adopted a tree-based feature selection technique [17] to compute impurity-based feature importances at a threshold of 0.07. With this feature selection technique, three features were removed. The final dataset contained 525 genes with seven highly confident features shown in Figure 6.

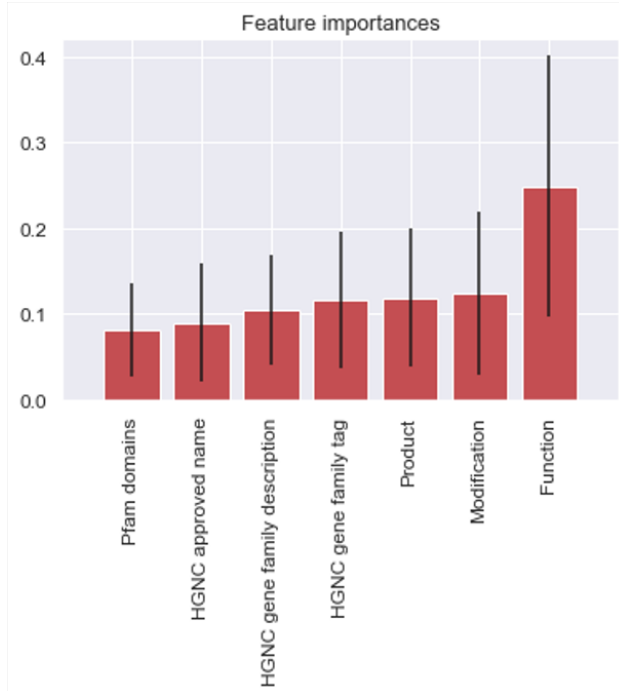

Fig.6: The top seven features (x-axis) with the respective importance weights (y-axis) determined by the tree-based selection method.

Training and Testing

An agnostic approach was adopted for the machine learning pipeline. We chose four different machine models: Decision Tree, Support Vector Machine, Random Forest, and Feed Forward Neural Network. The data was split into training (80%) and testing (20%) datasets.

The Prediction Models

- **Decision Tree** [18]: A non-parametric supervised learning method that aims to predict the value of a target variable by learning simple decision rules inferred from the data features. The model is obtained by recursively partitioning the data space and fitting a simple prediction model within each partition, where the partitioning can be represented graphically as a decision tree. Entropy is used to measure the quality of a split.
- **Support Vector Machine** [19]: A supervised learning machine for two-group classification problems. The main idea for SVM is to map the input feature vectors into some high-dimensional feature space Z through some non-linear mapping kernel function. Then, in this space, a linear decision surface (or hyperplane) is constructed with special properties that ensure the high generalization ability of the model to new data.
- **Random Forest** [20]: A meta-estimator model consisting of a collection of tree-structured (decision tree) classifiers classifier on various sub-samples of the dataset. After a large number of trees is generated, each tree casts a unit vote for determining the most popular class for a given new input sample.
- **Feed Forward Neural Networks** [21]: It is an artificial neural network model consisting of multi-layer perceptron (e.g., fully connected). It takes a feature vector as input and outputs a probability value to indicate which class to be assigned to for the input data. FNN optimizes the log-loss function using stochastic gradient descent.

Results and Evaluation

We evaluated the four models using 10-fold cross-validation and compared their prediction results by calculating F1 scores shown in Table 1. The prediction accuracies were generated through the

respective Sklearn metrics for each model. The F1 scores were calculated through the formula shown in Figure 7.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Figure 7: Equation for F1 score.

Precision is the number of true positives divided by the sum of true positives and false positives and recall is the number of true positives divided by the sum of true positives and false negatives. As shown in Table 1, Support Vector Machine, Decision Tree, Feed Forward Neural Network, and Random Forest had F1 scores of 0.87, 0.83, 0.84, and 0.82, respectively.

Machine Learning Model	Prediction Accuracy	F1 Score
Support Vector Machine	0.879	0.87
Decision Tree	0.825	0.83
Feed Forward Neural Network	0.843	0.84
Random Forest	0.813	0.82

Table 1: Prediction accuracies and F1 Scores of tested models.

A confusion matrix was plotted through Sklearn for each classification model as shown in Figure 8. The matrices summarize the number of true positives, false positives, false negatives, and true negatives among the reader, writer, and eraser predictions generated by each model.

Fig.8: Confusion matrices for the four tested machine learning models.

Overall, we found that a highly confident feature set yields a high model accuracy as shown in Table 1 and the precision-recall curve [22] in Figure 9.

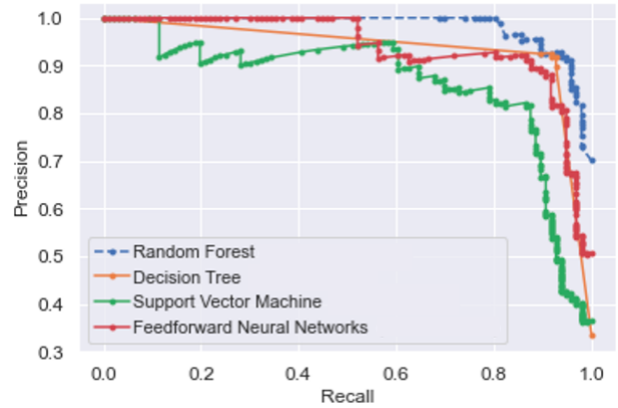

Fig.9: Precision-recall curves for the four tested machine learning models.

Discussion

With the initial data curation of chromatin regulators, we found overlaps between classified chromatin regulators as readers and writers or readers and erasers. Out of the 1274 chromatin regulators extracted, 127 proteins were classified as both readers and writers and 66 proteins were classified as both readers and erasers. This significant overlap reduces the clarity in understanding the epigenetic regulation of diseases such as cancer.

The lack of clarity could be due to the limited amount of available validated annotations. However, we also believe that the ambiguity in chromatin regulator roles is due to the complexity of their biological and molecular interactions. Chromatin regulators have been characterized as enzymes that work as part of a multiprotein complex [3], which explains why a chromatin regulator protein can act as a reader, writer, or eraser depending on its biochemical context.

Generally, knowledge of chromatin regulator functionality is obtained through rigorous wet-lab validations. However, these experiments are often time-consuming and costly. To overcome this barrier, we devised a data-driven approach that extracts biologically relevant features, refines those features, and utilizes them to train machine learning models to classify proteins into their chromatin regulator roles as readers, writers, and erasers.

Considering the machine learning pipeline developed in this study, we specifically trained models on biologically relevant features from the Epifactors database to predict chromatin regulator roles in cancer. The resulting prediction accuracies indicate that a highly confident feature set yields a decent model accuracy. Yet, there are limitations due to the lack of known and validated chromatin regulators, the size of the dataset, and imbalanced class labels. Due to these aspects of the data, the interpretation of prediction accuracies may be ambiguous.

Given the ambiguity in chromatin regulator roles, it is vital to further understand and identify the unique features of chromatin regulators that contribute to their characterizations as readers, writers, and erasers. By deriving and examining the biological significance of each epigenetic feature, we can further identify insightful patterns across extracted data and potential areas of improvement for the machine learning pipeline.

Out of the 7 epigenetic features selected for the machine learning model, *Pfam domain, HGNC gene family description, and HGNC gene family tag* are promising features that could be further expanded on for the development of CREWdb. Pfam domains are derived from a

database of protein families and utilize a hidden Markov model, which searches for alignments with the greatest matches at conserved sites. Many genes classified within the readers, writers, or erasers categories had similar or the same Pfam domains. Readers were consistently associated with ARID, Bromodomain, and JmjC domains; writers were associated with acetyltransferase, DUF, KRAB, MBD, and methyltransferase domains; and erasers were associated with amino oxidases, Arb2, Cupin, F-box, HDAC, SIR2 domains. Given the association between Pfam domains and chromatin regulator functionality, corresponding Pfam domains could be further inspected to understand the classification of readers, writers, and erasers.

In addition to extrapolating existing features, additional epigenetic information on cofactor interactions and histone modifications or disease-specific data could influence the predictability of the model. For example, for predicting chromatin regulator roles in cancer, additional features could include RNA expression derived from RNA-Seq data of primary cancer types, correlations between RNA expression and semantic copy number alteration for specific cancer types, correlations between RNA expression and promoter methylation level for specific cancer types, differential methylation status between tumor and normal samples, proteomics data based on reverse phase protein array for primary tumor tissues and antibody staining level for human normal tissues, and correlations between clinical features such as subtype, overall survival, etc. and RNA expression.

Conclusion

Through our machine learning pipeline, we demonstrate that by utilizing feature enrichment techniques such as feature selection and class rebalancing we can identify highly confident features that lead to high prediction accuracies for chromatin regulator classification. Given the overlap in chromatin regulator roles, we plan to implement a probabilistic model for each chromatin regulator role as reader, writer, or eraser given disease-specific features. This will further enable our understanding of the complex natures of these regulators' multifaceted roles in various disease phenotypes. In the future, we plan to extend our CRWEdb to a comprehensive knowledge base system by introducing text mining techniques to extract knowledge from biomedical articles and by integrating other disease knowledge sources. In doing so, we hope to establish the CREW database as a comprehensive and insightful knowledge source on the functional roles of chromatin regulators across all disease types.

Author Contributions

Original concept: S.M.; experimental design: S.M.; data analysis: M.N, M.K., M.S., S.M.; implementation: M.N., M.K., M.S., R.G., S.M.; data interpretation: M.N, M.S., S.M.; manuscript writing: M.N., R.G., S.M.

References

1. Marmorstein R, Zhou MM. Writers and readers of histone acetylation: structure, mechanism, and inhibition. Cold Spring Harb Perspect Biol. 2014 Jul 1;6(7):a018762. doi: 10.1101/cshperspect.a018762. PMID: 24984779; PMCID: PMC4067988
2. Hyun K, Jeon J, Park K, Kim J. Writing, erasing and reading histone lysine methylations. Exp Mol Med. 2017 Apr 28;49(4):e324. doi: 10.1038/emm.2017.11. PMID: 28450737; PMCID: PMC6130214.
3. Biswas S, Rao CM. Epigenetic tools (The Writers, The Readers and The Erasers) and their implications in cancer therapy. Eur J Pharmacol. 2018 Oct 15;837:8-24. doi: 10.1016/j.ejphar.2018.08.021. Epub 2018 Aug 18. PMID: 30125562.
4. Brown PJ, Müller S. Open access chemical probes for epigenetic targets. Future Med Chem. 2015;7(14):1901-17. doi: 10.4155/fmc.15.127. Epub 2015 Sep 23. PMID: 26397018; PMCID: PMC4673908.
5. Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends Genet. 2016 Jan;32(1):42-56. doi: 10.1016/j.tig.2015.10.007. Epub 2015 Dec 17. PMID: 26704082.
6. Gillette TG, Hill JA. Readers, writers, and erasers: chromatin as the whiteboard of heart disease. Circ Res. 2015 Mar 27;116(7):1245-53. doi: 10.1161/CIRCRESAHA.116.303630. PMID: 25814685; PMCID: PMC4380191.
7. *The cancer genome atlas program*. National Cancer Institute. (n.d.). Retrieved May 17, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga 7
8. Torres IO, Fujimori DG. Functional coupling between writers, erasers and readers of histone and DNA methylation. Curr Opin Struct Biol. 2015 Dec;35:68-75. doi: 10.1016/j.sbi.2015.09.007. Epub 2015 Nov 9. PMID: 26496625; PMCID: PMC4688207.
9. Beibei Ru, Jianlong Sun, Yin Tong, Ching Ngar Wong, Aditi Chandra, Acacia Tsz So Tang, Larry Ka Yue Chow, Wai Lam Wun, Zarina Levitskaya, Jiangwen Zhang, CR2Cancer: a database for chromatin regulators in human cancer, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D918–D924, https://doi.org/10.1093/nar/gkx877
10. Yang Xu, Shuang Zhang, Shaofeng Lin, Yaping Guo, Wankun Deng, Ying Zhang, Yu Xue, WERAM: a database of writers, erasers and readers of histone acetylation and methylation in eukaryotes, *Nucleic Acids Research*, Volume 45, Issue D1, January 2017, Pages D264–D270, https://doi.org/10.1093/nar/gkw1011
11. Nicholson, Thomas B., et al. "Writers, Readers, and Erasers of Epigenetic Marks." *Epigenetic Cancer Therapy*, 2015, pp. 31–66., https://doi.org/10.1016/b978-0-12-800206-3.00003-3.
12. Vossen G. (2009) ACID Properties. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston,MA. https://doi.org/10.1007/978-0-387-39940-9_831
13. Gonzalez-Perez A., Jene-Sanz A., Lopez-Bigas N. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. *Genome Biol.* 2013; 14:r106.
14. *Cancer cell line encyclopedia (CCLE)*. OpenScholar. (n.d.). Retrieved May 17, 2022, from https://sites.broadinstitute.org/ccle/
15. Yulia A. Medvedeva, Andreas Lennartsson, Rezvan Ehsani, Ivan V. Kulakovskiy, Ilya E. Vorontsov, Pouda Panahandeh, Grigory Khimulya, Takeya Kasukawa, The FANTOM Consortium, Finn Drabløs, EpiFactors: a comprehensive database of human epigenetic factors and complexes, *Database*, Volume 2015, 2015, bav067, https://doi.org/10.1093/database/bav067
16. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.
17. https://scikitlearn.org/stable/auto_examples/ensemble/plot_forest_importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py
18. Loh, W.Y., 2011. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), pp.14-23
19. Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), pp.1-27.
20. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.
21. Glorot, X. and Bengio, Y., 2010, March. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.
22. Boyd, Kendrick, Kevin H. Eng, and C. David Page. "Area under the precision-recall curve: point estimates and confidence intervals." In Joint European conference on machine learning and knowledge discovery in databases, pp. 451-466. Springer, Berlin, Heidelberg, 2013.