Introduction

The human body contains over 200 different types of cells, each of which possess their own individual structures and functions [1]. This variety is mediated and maintained primarily by the unique expression of up to thousands of genes contained within the human genome, which the cell achieves primarily by controlling the rate of gene transcription; this process is referred to as transcriptional regulation [8]. Transcription factors (TFs) are proteins that are the main tools by which transcriptional regulation is accomplished. TFs possess DNA-binding domains that allow them to bind to specific sequences in active regulatory regions of DNA, referred to as enhancer or promoter sequences [15].

There are numerous experimental tools currently available for the detailed study and analysis of transcription factors. One of the most common in-vivo procedures is ChIP-seq, which stands for chromatin immunoprecipitation followed by massively parallel sequencing. Tools such as ChIP-seq identify genome-wide DNA binding sites for transcription factors (along with other DNA interacting proteins such as RNA polymerase), which in turn allows for the computing of corresponding DNA sequence motifs, precise regulatory sites, direct downstream targets of individual transcription factors, and other relevant details for study [10].

Though modern DNA binding region identification procedures such as ChIP-seq have many advantages, they come with a high level of experimental effort and cost that currently stops them from being applied for many different transcription factors from a wide range of human cells [16]. Thus, there is a necessity for bioinformatic methods that can computationally assess transcription factor binding sites in DNA with a high level of accuracy with minimal time and cost. The approach used in this report combines a large amount of experimental data derived from ChIP-seq application on thousands of transcription factors and the assessed active genomic regulatory motifs in various cell types, to build a machine-learning model that can predict transcription factor binding sites on the human genome.

Specifically, AP1 transcription factor was chosen as the target dataset for this model. AP1 is a protein critical for transcription through its role in opening chromatin and maintaining DNA accessibility for other regulatory proteins [2].

However, it has been understood that sequences of active regulatory motifs are not the only predictive factors, as they contain many binding sites that do not end up being bound by transcription factors [9]. Thus, the machine learning model we build in this report also uses DNA shape as a crucial model feature.

Though DNA is usually considered to be a highly constrained and regularly structured molecule, with a phosphate sugar backbone and paired bases arranged in a double helix, the local DNA shape of a region can vary significantly based on the specific position of every base pair within the sequence, as well as the base pairs that surround the region. These pairs are seen to shift, slide, twist and move in other axes relative to one another, while the individual bases within the pair can themselves twist and stagger and shift. In addition, the DNA molecule also contains major and minor grooves, which can widen or narrow depending on the surrounding base pair context [14]. The resulting local DNA shape has been seen to impact the binding of transcription factors (and other DNA-binding proteins) [13].

Supervised machine learning has been a common approach to use in the literature regarding the prediction of transcription factor binding sites [6]. When used in this context, this prediction problem is considered to be a classification problem, where there are two possible options for any given motif sequence: it either corresponds to a TF binding site, or it does not. A number of machine learning classification approaches exist, and brief backgrounds on relevant models is given below.

Logistic Regression is a classification approach often used when the value of a given data variable is categorical and when the classification is binary, as is the case when predicting transcription factor binding sites. It works by fitting a logistic curve to the given dataset, and then computing the probability of a given target value class.

K-Nearest Neighbors (KNN) computes classification by first assessing the class of the nearest K neighbors of a point in the data set, and assigning the given data point the classification of the majority of such neighbours. The value for K is a crucial variable in determining algorithm performance, prediction accuracy and level of bias.

Naive Bayes is a classification algorithm based on applying the Bayes Theorem with the assumption that every feature in the dataset is independent of one another. Though this is a ‘naive’ assumption, this approach tends to perform much faster than other algorithms, and can still be effective in bioinformatics.

Support Vector Machine algorithms work by finding the optimal hyperplane that distinctly classifies the data set, such that it maximizes the distance between differently classified data points. This approach is highly customizable, and can accurately classify a variety of feature spaces by adjusting the kernel function it uses. However, it can suffer from performance issues when a dataset contains a large number of features [11].

Methodology

Several datasets were used concurrently to obtain the AP1 binding site dataset. First, active regulatory regions of the human genome (and thus, possible TF binding sites) were found from the GM12878 cell type, which is a cell line derived from lymphoblasts. This was then run through the HOMER motif analysis algorithm, an approach designed for regulatory element analysis of gene sequences, to obtain a list of all active motifs. This was then compared with a set of genomic coordinates for experimentally derived AP1 binding sites obtained through ChIP-Seq, to obtain positive and negative examples for AP1 binding.

Approximately 161,000 positive binding examples were found for AP1 and 30,000 overall examples were found using Homer. Approximately 6000 of the positive examples overlapped with this. The remaining negative examples were randomized and cut to match the number of positive examples for optimal training. To prepare the data for use in machine learning algorithms, positive and negative examples were randomly divided into 70:30 training and testing groups.

Each example contains a DNA sequence of 14 nucleotides. To transform these categorical values into numeric inputs for machine learning analysis, every nucleotide was encoded using a one-hot scheme.

Nucleotide	**One-Hot Encoding**
A	[1, 0, 0, 0]
C	[0, 1, 0, 0]
G	[0, 0, 1, 0]
T	[0, 0, 0, 1]

Figure 1: One Hot Encoding scheme for each nucleotide.

Additionally, each motif sequence was processed by the DNAshapeR library to produce 5 shape vectors for each motif. Each vector represents the degree to which the DNA sequence conforms to a specific shape at each nucleotide position. The shape classifications were EP, HelT, MGW, ProT, and Roll. These vectors were first considered separately using each classification algorithm to determine their importance in transcription factor binding, then together with the sequence vectors to achieve maximal accuracy.

Several classification approaches were used to predict the transcription factor binding sites, namely Logistic Regression, Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Every classification approach ran on a data set derived by randomly dividing the total dataset into 70% training data and 30% testing data. All of the algorithms and support functions were implemented using the Python scikit-learn library. The following table shows the specified parameters that were used in relevant classification methods.

A radial basis function kernel was chosen for the Support Vector Machine approach, as potentially the sequence and shape feature datasets are not linearly separable. A K-value of 3 was chosen for K-Nearest Neighbours, in line with the literature on transcription factor binding site predictions [7].

Classification Approach	Relevant Parameters
K-Nearest Neighbours	K = 3
Logistic Regression	Default parameters
Support Vector Machines	Radial Basis Function Kernel
Naive Bayes	Gaussian

Figure 2: Parameters used for each learning model.

Two key performance metrics were used to assess the predictive power of each classification method: accuracy and precision. Accuracy was measured by calculating the sum of true positives and true negatives divided by the total size of the dataset, and is a measure of how accurate the model’s predictions are. Precision was measured by calculating dividing the sum of true positives and false positives from the total number of true positives, and is a measure of statistical relevance.

Dummy classifiers from the scikit-learn library were also used to establish a baseline for the dataset. Four such classifiers were used: stratified, most frequent, uniform, and constant. Most frequent classifiers predict labels based on the most frequent label. Stratified classifiers creates random predictions with respect to the training set distribution. Uniform classifiers creates uniform random predictions, and constant always predicts a constant label - for this paper, the label of 1 was chosen [11]. These baselines were established for the dataset incorporating all features (sequence and all five shapes).

Results

	SVM Accuracy	KNN Accuracy	Naive Bayes Accuracy	Logistic Regression Accuracy	Average Accuracy
Sequence	**84.6%**	80.54%	80.59%	84.28%	82.50%
EP	**81.30%**	80.89%	81.30%	81.02%	81.13%
HelT	80.62%	81.30%	81.30%	**84.00%**	81.80%
MGW	**81.30%**	80.77%	81.30%	81.22%	81.15%
ProT	**82.86%**	80.87%	81.30%	81.17%	81.55%
Roll	81.30%	81.20%	81.30%	**84.43%**	82.05%
All Features Together	92.63%	**94.93%**	92.48%	92.83%	93.22%

Figure 3: Accuracy of different machine learning models across different motif features individually and combined. Bolded values are the highest accuracy for each feature.

	SVM Precision	KNN Precision	Naive Bayes Precision	Logistic Regression Precision
Sequence	77.52%	**77.96%**	72.03%	77.56%
EP	72.69%	**78.23%**	72.69%	73.10%
HelT	71.98%	**78.60%**	72.69%	77.70%
MGW	72.69%	**78.58%**	72.69%	72.76%
ProT	76.94%	**77.78%**	72.69%	72.73%
Roll	72.69%	**78.56%**	72.69%	77.32%
All Features Together	87.13%	**91.09%**	87.10%	88.43%

Figure 4: Precision of different machine learning models across different motif features individually and combined. Bolded values are the highest precision for each feature.

All of the shape classifications considered individually performed relatively well in terms of accuracy across tested machine learning models, with each being above 80%. The optimal machine learning model for each individual feature varied, though accuracy was not significantly different across models.

The K-Nearest-Neighbors (KNN) model provided for the best accuracy with all features considered together at 94.9%. The accuracy for all features considered together was higher across all models than any feature considered individually. The SVM provided the greatest number of highest accuracies for features considered individually, at 4/6. All three models aside from KNN performed very similarly for all features considered together, within 1 decimal place.

The KNN model for the combined features dataset additionally provided the best precision at 91.1%. It additionally had the highest precision for every feature considered individually. The other models had similar lower precisions to each other in the same ranking as their accuracies. The precision across all models considering all features together was higher than any precision for a feature considered individually.

	SVM	**KNN**	**Naive Bayes**	**Logistic Regression**
Running Time (s)	18.9	1.47	**0.07**	15.1

Figure 5: In seconds, the amount of time each model takes to train and test on the same set of data considering all features together, as described in the methodology section. Bold text indicates the fastest running model.

The models each trained and tested on the same set of data in under 30 seconds (see figure 5). The Naive Bayes model ran the fastest, with a time of 0.07 seconds. The KNN model ran the fastest out of the viable models with a time of 1.47 seconds.

	Most Frequent	**Stratified**	**Constant**	**Uniform**
All Features Together	49.7%	51.1%	49.7%	48.2%

Figure 6: Accuracy scores of dummy classifier models for the combined feature dataset.

The dummy classifier models each trained and tested on the combined feature dataset. These accuracy scores fell far below the scores generated by all machine learning models considered.

Discussion

In this paper, the prediction of binding sites of AP1 on the human genome was tested using four unique machine learning classification models: K-Nearest (KNN), Support Vector Machines (SVM), Naive Bayes (NB), and Logistic Regression (LR). In addition, local DNA shape analysis data for five shape features was also used: helix twist (HelT), propeller twist (ProT), minor groove width (MGW), electrostatic potential (EP), and roll (Roll). These features were used in the models both independently and together with sequence data. The performance of these models were measured by assessing their accuracy and their precision.

Our results show that, even when solely using local shape information or sequence data , prediction accuracy and precision was high for every classification model used, ranging from 80.8% to 84.4% and 71.9% to 78.6% respectively across all shape features and models. However, when the models were trained using a combination of all shape features and sequence data, the performance of every model significantly increased, with an average model accuracy and precision of 93.2% and 88.4% respectively.

The K-Nearest-Neighbors (KNN) model provided for the best accuracy with all features considered together at 94.9%. The KNN model for the combined features dataset additionally provided the best precision at 91.1%. This may indicate that the KNN approach is optimal for AP1 binding site prediction, and TF binding site prediction in general.

Considering the average model accuracy scores for each isolated feature dataset allows for a potential ranking of feature importance in accurately determining if a transcription factor binds at a particular motif. Based on this metric, the sequence feature had the highest average accuracy score (82.50%) amongst isolated feature datasets, indicating that the linear sequence itself is the most important consideration for AP1 binding sites. This may be partially due to each shape vector being calculated solely based on the sequence and no additional data, meaning each shape by itself does not contain much more information than the sequence by itself. However, combining all of the shape calculations together with the sequence significantly improved the accuracy of all models (*see Fig. 3*), indicating that it does provide the models with significantly more information, thus revealing that shape features play a significant role in transcription factor binding.

HelT and Roll features had higher average accuracy scores compared to the other 3 shape features. This is possibly due to their values being calculated as base pair-steps, while the other shape values are calculated solely using each base pair, leading to the generation of new information [3].

Potential Errors

When considered individually, some shape classifications had the same accuracy when using the same model, such as EP and MGW using an SVM or all of the shape classifications with a Naive Bayes model. This may indicate an underlying error in the training model, or in the feature dataset. However, this is more likely due to the nature of each model and how the shape values for each motif are calculated. Two shapes that are calculated similarly will give similar relative values for each motif, despite the specific values being different. This would result in a classification algorithm such as SVM having the same accuracy, as the optimal hyperplane would divide the groups the same way despite being in a different position. This is also particularly relevant for Naive Bayes, as it assumes that all feature predictive values are independent [11]. This is likely an indication that the Naive Bayes model is not suited for this classification problem.

Further Research

The training models generated in this paper can be further experimented on using datasets derived from other transcription factors. This additional testing can reveal which transcription factor binding sites can be predicted with a high level of accuracy and precision using models trained on the AP1 dataset, suggesting that they have similar binding site locations to AP1 and possible similarities in function and structure. Training new classifier models for additional transcription factors can also reveal if KNN is the best approach for other transcription factor binding site predictions, as it was seen to be for the combined feature dataset of AP1.

One possible method for improving accuracy is incorporating the models investigated in this paper into an ensemble machine learning algorithm. This algorithm would combine weighted decisions from each subordinate algorithm with learned weights to achieve the highest accuracy, using advantages from each algorithm. This can also allow for better confidence estimation of a given classification, as if a large majority of the models used agree with the overall classification, this can be seen as the ensemble approach having a high level of confidence in the label prediction, while if a slim majority of the models agree, then this can be seen as low confidence [4].

In conclusion, a machine learning model, such as the KNN used in this paper, has the ability to reduce the amount of experimental testing needed to confirm transcription factor binding sites.

References

1. Alberts, B., and A. et al. “Molecular Biology of the Cell.” Johnson, 2008.
2. Biddle, Simon C., et al. “Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding.” *Molecular Cell*, 2011.
3. Chiu, Tsu-Pei, et al. “DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding.” *Bioinformatics,* 2016.
4. Ganaie et al. “Ensemble deep learning: A review.” *Cornell University arXiv,* 2021.
5. Johnson, D. S., et al. “Genome-wide mapping of in vivo protein-DNA interactions.” *Science*, 2007.
6. Koo, Peter et al. “Deep learning for inferring transcription factor binding sites.” *Current Opinion in Systems Biology*, 2020.
7. Mahmoud, Maieda. “Prediction of Transcription Factor Binding Sites of SP1 on Human Chromosome1.” *Applied Science,* 2021.
8. Narlikar, Leelavati, and Ivan Ovcharenko. “Identifying regulatory elements in eukaryotic genomes.” *Briefings in functional genomics & proteomics*, 2009.
9. O'Malley, R. C. et al. “Cistrome and epicistrome features shape the regulatory DNA landscape.” *Cell,* 2016.
10. Park, P. J. “ChIP-seq: advantages and challenges of a maturing technology.” *Nat Rev Genet,* 2009.
11. Pedregosa et al. “Scikit-learn: Machine Learning in Python.” JMLR 12, pp. 2825-2830, 2011.
12. Ralston, Amy. “Do Transcription Factors Actually Bind DNA? DNA Footprinting and Gel Shift Assays.” *Nature Education*, 2008.

13. Riechmann, J. L. et al. “Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes.” *Science*, 2000.
14. Rohs, R. et al. “The role of DNA shape in protein–DNA recognition.” *Nature,* 2009.
15. Stormo, G. D., and D. S. Fields. “Specificity, free energy and information content in protein–DNA interactions.” *Trends Biochem Sci*, 1998.
16. Wu, J., et al. “ChIP-chip comes of age for genome-wide functional analysis.” *Cancer Res*, 2006.