JOURNAL OF
EMERGING INVESTIGATORS

Article

Can the nucleotide content of a DNA sequence predict the sequence accessibility?

Shreyes Balachandran[1], Diwakar Balachandran[2]
[1]St. John's School, Houston Texas
[2]The University of Texas MD Anderson Cancer Center, Houston, Texas

SUMMARY

Sequence accessibility is an important factor affecting gene expression. Sequence accessibility or openness impacts the likelihood that a gene is transcribed and translated into a protein and performs functions and manifests traits. The DNA, which carries the genes, is packaged as chromatin. There are two types of chromatin, heterochromatin and euchromatin. Heterochromatin tends to be inaccessible and thus is often not expressed. In contrast, euchromatin is more accessible and is expressed. Accessibility of a gene depends on the type of chromatin it is in, and with increased accessibility, there is a greater likelihood of gene transcription and expression. There are many potential factors that affect the accessibility of a gene. In this study, our hypothesis was that the content of nucleotides in a genetic sequence predicts its accessibility. Using a machine learning linear regression model, we studied the relationship between nucleotide content and accessibility. DNA sequences are made up of four nucleotides. We compared the quantity of each of these four nucleotides, adenosine, thymine, guanine, and cytosine either as single nucleotide or in specific combinations of two nucleotides with sequence accessibility using the K562 cell line. Of all the combinations tried, we discovered that the cytosine-guanine combination content had the highest positive correlation with accessibility, and therefore with gene expression. This correlation allows us to better predict which genetic sequences will be more frequently expressed based solely on the nucleotide content and sequence. Predicting gene expression through machine learning algorithms promises to catalyze our ability to understand the structure and function of specific gene sequences.

INTRODUCTION

The human genome consists of approximately 30,000 genes and three billion base pairs (combination of two nucleotides on the two strands of the DNA double helix) (1). The human genome project sequenced the entire human genome by splicing it into parts and using bacterial artificial chromosomes (1). However, only a fraction of nearly 30,000 genes are known to be expressed (1). The regulation of gene expression is dependent on multiple factors including sequence accessibility (which impacts mutations which cause diseases), inheritance patterns, and personalized health and treatment solutions (2). Because sequence accessibility is highly correlated to gene expression, understanding the underlying factors which influence sequence accessibility and therefore expression is critical (3).

The genome contains the DNA that codes for proteins, which perform nearly all biological tasks in the cell. The four nucleotides that compose DNA are adenosine (A), thymine (T), guanine (G), and cytosine (C). These nucleotides usually occur as base pairs with A pairing with T and G pairing with C in the double helix DNA structure. DNA exists as chromatin, which is condensed DNA, making up the entirety of the genome (4). Histones are proteins around which DNA is wrapped to pack the DNA, as their positive charge attracts negatively charged DNA (4). The coiled DNA and histone nucleoprotein complex is referred to as chromatin (3). The basic unit of the chromatin is a nucleosome which consists of nearly 147 base pairs of DNA wrapped around an octamer of histone proteins (4). Chromatin can be tightly bound or loosely bound. Tightly bound chromatin, or heterochromatin, composes the majority of chromatin in humans. Typically, the genes in heterochromatin are less accessible because of the tight binding and are less expressed. In contrast, loosely bound chromatin, known as euchromatin, contain genes that are more accessible and more likely to be expressed (5).

The process of gene expression involves the transcription of DNA. The process of transcription is the copying DNA to RNA. RNA polymerase and transcription factors (proteins) bind to certain sites on the chromatin, which must open—a process called chromatin remodeling (6). There are regulatory regions for the chromatin called enhancers and promoters within the chromatin. Once the transcription factor binds to the promoter site, the genes regulated by this promoter are transcribed into mRNA. The mRNA is later translated to proteins, which accomplish cellular functions (7). Accessible genes are more likely to be transcribed, and therefore more likely to be transcribed into proteins and express their intended function. Generally, accessible regions of chromatin are permissive for transcription factor binding and are therefore hotspots for regulation of gene expression; conversely, genomic regions that are highly occupied by histone proteins are not permissive for transcription factor binding and are less likely to be active regulatory regions. This impacts cellular function and expression of traits. As a result of its greater accessibility, euchromatin is more likely to be expressed (8).

Prediction of accessibility is important for many reasons such as identifying which genetic sequences will be more represented in the cells, which sequences may be more prone to cause disease, and which sequences are targets for gene editing by CRISPR (9). Determining the genomic localization of chromatin-bound proteins is therefore essential in determining functional roles, sequence motifs important for factor binding, and regulatory networks controlling gene expression (10). The measurement of sequence accessibility is determined by several techniques including ChIP (chromatin immunoprecipitation), Formaldehyde-assisted Isolation of Regulatory Elements (FAIRE) sequencing, RNA-sequencing (RNA-seq), DNase hypersensitivity site sequencing, and Assay for Transposase Accessible Chromatin (ATAC-seq) (10, 11). These techniques helped define the interplay between DNA sequence characteristics, histone occupancy, and transcription factor binding in the regulation of gene expression (12). Nucleotide content is one such characteristic of the DNA sequence that impacts gene expression which determines this complex interplay of factors and is the focus of this study.

The purpose of this study was to assess correlation of nucleotide content with sequence accessibility and thereby gene expression using a machine learning linear regression model for prediction. We predict that the machine learning model will show a correlation between nucleotide content and gene accessibility. Machine learning requires training of a dataset to validate the algorithm used. We employed a trained data set to create a model to formulate the correlation between the input variable (nucleotide content) and the dependent variable (sequence accessibility). We did this for nucleotide content which were comprised of a single or a pair of nucleotides. We then calculated the correlation coefficient (R-value) to determine which nucleotide sequences had the closest relationship with sequence accessibility. Our results demonstrated a correlation between the GC sequence and accessibility. These results demonstrate the utility of using machine learning models to predict gene expression in both health and disease.

RESULTS

Our machine learning model was based on the use of a well describe cell line to both train and then test the model to examine the ability of nucleotide content to predict gene expression. The K562 cell line (chronic myelogenous leukemia cell line), which is a continuous cell line, and accessibility data was derived using the DNase sequencing (DNase-seq) technique to fragment the DNA. This process is described in detail in the methods section. A training set consisting of all the chromosomes, except Chromosome 1, was used for training the machine learning model to perform a linear regression comparing the nucleotide content to the sequence accessibility. This was subsequently validated on the test set, the selected DNA sequences used to test our model, which consisted of Chromosome 1.

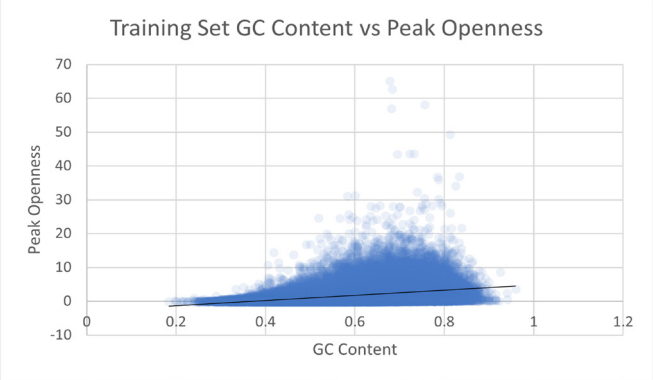

Figure 1: A scatterplot between the GC content of each peak and the accessibility of each peak over the training set. The trained linear regression model is in black (r = 0.397; MSE =3.837; Predicted openness = 7.600 * GC content – 2.772). Openness/accessibility is measured as a fraction of available DNA sites.

There were a total of 106,629 nucleotides for the training set. There were a total of 12,452 nucleotides for the test set which corresponded to Chromosome 1. The machine learning model for linear regression trained on the training set showed a positive slope between the G, C content and the accessibility of the peak, which suggests that a higher G, C content is conducive to accessibility (**Figure 1**). We then computed the Pearson correlation coefficient r between the nucleotide content and accessibility in the training set. This was then tested on the test set (**Figure 2**). When the model was tested, it yielded different correlations for each nucleotide group to which it was applied.

The sequence CG (combined) correlated the most with the accessibility and openness of the chromatin (**Table 3**). The R-value for the CG sequence was positive 0.469 and showed the highest R-value for all the analyses and for the combination double base pairs. The next highest R-value was positive 0.429 and was for the triple base pair combination CGC. The R-values for the other nucleotides are shown in the following tables. Using this method to correlate the type of base nucleotide to accessibility and gene expression showed a moderate correlation for the CG content of the chromatin

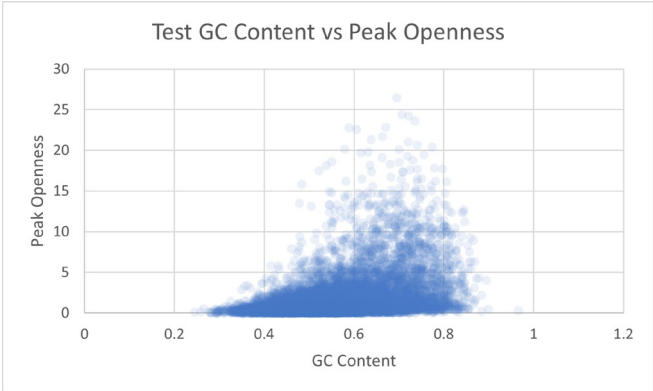

Figure 2: A scatterplot between the test set GC content of each peak and the accessibility of each peak over the test set (r = 0.417; MSE = 5.119).

Individualized	G and C	A and T	T and C	A and G
Slope	7.00	-9.520	0.088	-0.088
Intercept	-2.772	6.006	1.707	1.794
R-value	0.417	-0.417	0.0028	0.0028

Table 1

Combined (G)	GA	GC	GG	GT
Slope	-7.449	26.396	12.624	-17.071
Intercept	2.207	-0.227	0.619	2.576
R-value	-0.077	0.392	0.255	-0.147

Table 2

Combined (C)	CA	CC	CG	CT
Slope	-20.351	13.204	29.126	-11.148
Intercept	3.212	0.556	0.752	2.630
R-value	-0.213	0.271	**0.469**	-0.126

Table 3

Combined (A)	AA	AC	AG	AT
Slope	-16.407	-15.142	-11.530	-29.837
Intercept	2.643	2.482	2.649	2.951
R-value	-0.235	-0.130	-0.130	-0.296

Table 4

Combined (T)	TA	TC	TG	TT
Slope	-28.220	-8.423	-20.527	-16.471
Intercept	2.694	2.272	3.232	2.662
R-value	-0.265	-0.087	-0.217	-0.241

Table 5

Combined (CG)	CGA	CGC	CGG	CGT
Slope	100.996	55.089	57.329	91.100
Intercept	1.287	1.072	1.047	1.302
R-value	0.266	0.429	0.406	0.235

Table 6

Tables 1-6: Correlation between nucleotide content and gene accessibility. R-values are correlation coefficients.

(**Table 3**). Although the correlation was moderate, this was the highest association for all single nucleotides and double nucleotides sequences.

These findings showed a moderate correlation between the CG nucleotide content and the sequence accessibility both in the training set and the test set (**Figures 1,2**). This was optimized to reduce the mean standard error (MSE) The correlation for other single nucleotides (A, T, G, C) and for double nucleotide combinations such as AT, TA were less correlated with sequence accessibility (**Tables 2-5**). Given the higher R-value with CG content, three nucleotide sequences were run in the predictive model for CGA, CGT, CGG, CGC and the R-values for these 3 nucleotide models were less than with the CG content correlation alone (**Table 6**).

DISCUSSION

We examined sequence accessibility and compared it with the nucleotide sequence of the gene to further understand if the preponderance and sequence of certain nucleotides is associated with increased accessibility. We created a machine learning model of linear regression between these two variables. We trained the model on a training set and then subsequently on a test set. The results between the two sets were similar indicating the high validity of this model and confirming the moderate correlation between the CG sequence and sequence accessibility. One possible reason for this is because DNA methylation occurs more frequently in areas where cytosine and guanine predominate (13).

Our study adds to the growing literature that uses machine learning linear regression model to evaluate nucleotide content and sequence and correlates this with the sequence accessibility (2). There are a few other studies on machine learning and sequence accessibility which address nucleotide content as a factor in determining gene expression (14-16). Most recently, Zrimec et al. describes a machine learning deep neural network model to determine the relationship between DNA protein binding site and gene expression (16).

Gene expression has been studied extensively. There are several factors that control gene expression. These include both intrinsic factors such as sequence accessibility and extrinsic factors such as availability of nutrients (17, 18). Gene expression is divided into transcription (conversion of the DNA to mRNA) and translation (conversion of the mRNA into protein). For initiation of transcription, several steps must occur, including the attaching of the RNA polymerase and mediator complex to the promoter portion of the DNA sequence, a small DNA segment which is just upstream of the gene. Gene and DNA accessibility can affect the gene expression by controlling the accessibility of the DNA strand to transcription factors and mediator and RNA polymerase complex (16).

We calibrated the training set with GC (independent) content, exclusively, so calibrating it for different nucleotides may have reduced the MSE values we found, limiting the ability to generalize our findings. Furthermore, we did not isolate the promoter sequence for the genes and assess their nucleotide content. This may have independently impacted sequence accessibility. Lastly, we used only the K562 cell line as the large training set to study the predictive ability of the machine learning model. Future studies may employ other cell lines to see if the machine learning model can be improved. Future studies to predict the correlation of nucleotide content and sequence accessibility need to be performed including to determine if the nucleotide content of the specific promoter sequence for the genes correlates with sequence accessibility (19). Similarly, the enhancer sequence for genes could be isolated and correlation between the nucleotide content of the enhancer sequence and sequence accessibility could be assessed. In our study, nucleotide content was not found to directly affect gene expression, it was found to affect gene accessibility. In addition we recognize that while gene accessibility is correlated to gene expression, it is not precisely causal Finally, it is important to consider that nucleotides are not equally present in this cell line, which could cause further error (20).

Our predictive machine learning model was able to find

a correlation between nucleotide content and sequence accessibility and expression. This novel tool of machine learning will continue to allow for further exploration of this and other factors of gene expression. Further studies with machine learning can explore the complex interplay of multiple factors which influence gene expression.

MATERIALS AND METHODS

The data used for this research was obtained from the ENCODE (Encyclopedia of DNA Elements) Consortium. ENCODE is an international collaborative effort which seeks to identify and build a comprehensive list of functional elements of the human genome (21). We used data from the K562 cell line which is a continuous cell line (20). Accessibility data was derived from a sample which used the deoxyribonuclease DNase-seq technique, which was used to fragment the DNA. In DNase-seq, the population of cells is treated with the DNase enzyme. The DNase cuts the DNA in the chromatin into smaller fragments by attaching to the hypersensitive regions of the DNA. This occurs to a far greater extent where the DNA is not too tightly bound by histones. The regions which were tightly bound by histones contribute only to large fragments (11). Regions which are particularly accessible are called peaks, and the accessibility varies across the peaks. We downloaded the accessible peaks in K562 from the ENCODE project. From the experiment with ID ENCSR000EOT, we downloaded the IDR-thresholded peaks with file ID ENCFF185XRG. Using the human genome alignment hg38, we extracted the underlying DNA sequence of each called peak using bedtools getfasta (22). We use the signal value column of the peak file as the accessibility of each peak.

We used the DNase-seq peaks and computed the nucleotide sequence in each peak. We then split these peaks into two sets: the training set, which consists of peaks from all chromosomes except chromosome 1, and the test set, which consists of peaks in chromosome 1. We trained the machine learning model using the training set. On this machine learning regression model, we used the nucleotide content of each underlying peak as our input feature and the peak accessibility as our output target. The predictive model trained on the training set shows a positive slope between the GC content and the peak, which suggests that a higher GC content is conducive to sequence accessibility. We then computed the Pearson correlation between the peak's GC content and accessibility in the training set. To calculate the GC content of each peak, we took the DNA sequence of the peak, searched for all G and C nucleotides, and found the ratio of that amount to the total amount of nucleotides in the DNA sequence of the peak.

The following variables were determined in our model: x is the nucleotide content and y is the sequence accessibility. To train our linear regression model, we computed the slope and y-intercept of the training set model using these equations 1 and 2.

$$\text{Slope} = (n * \sum xy - \sum x * \sum y)/(n*\sum x^2-(\sum x)^2) \qquad \textit{Eqn. 1}$$

$$\text{Intercept} = (\sum y*\sum x^2 - \sum x\sum xy)/(n*\sum x^2-(\sum x)^2) \qquad \textit{Eqn. 2}$$

To compute the coefficient of correlation, R-value, between GC content and peak accessibility, we used equation 3.

$$r = (n*\sum xy-\sum x*\sum y)/\text{sqrt}((n*\sum x^2 - (\sum x)^2)*(n*\sum y^2-(\sum y)^2)) \qquad \textit{Eqn. 3}$$

The equations were used to create a predictive equation and we calculated the R-value for the training set. Then, we found predicted *Y* values for the test set by using the *X* values into our training set equation, and compared them to the actual values of *Y* using equation 4 for MSE (mean squared error) with *n* being the number of peaks; *x*, the value of the GC content; *y*, the true accessibility; and $\hat{y}$, the predicted accessibility.

$$\text{MSE} = (1/n)*\sum(y-\hat{y})^2 \qquad \textit{Eqn. 4}$$

Additionally, we found equations and R-values for the test set using the equations 1–3 to see how well they could be independently predicted. The R-value and MSE determined the effectiveness of a correlation. We used DNAse seq along with linear regression as the method to assess sequence accessibility in this study. Linear regression was used to understand the relationship between the variables and the R-value was also evaluated. Linear regression using machine learning remains a novel tool in assessing the impact of nucleotide content of genes. The combination of the nucleotides is further classified as individualized or combined. Individualized refers to looking for each provided nucleotide separately, whereas, combined, is the opposite and means looking for their occurrences together in the order written. All possible double nucleotides (all combinations) were checked against the accessibility of the genes. The R-value of the different nucleotides sequences was tabulated and correlation with sequence accessibility was measured.

Received: November 18, 2021
Accepted: August 18, 2022
Published: March 10, 2023

REFERENCES

1. Andersson, R. and A. Sandelin. "Determinants of Enhancer and Promoter Activities of Regulatory Elements." *Nat Rev Genet*, vol. 21, no. 2, 2020, pp. 71-87, doi:10.1038/s41576-019-0173-8.
2. Bannister, A. J. and T. Kouzarides. "Regulation of Chromatin by Histone Modifications." *Cell Res*, vol. 21, no. 3, 2011, pp. 381-395, doi:10.1038/cr.2011.22.
3. Chereji, R. V. et al. "Accessibility of Promoter DNA Is Not the Primary Determinant of Chromatin-Mediated Gene Regulation." *Genome Res*, vol. 29, no. 12, 2019, pp. 1985-1995, doi:10.1101/gr.249326.119.
4. Eraslan, G. et al. "Deep Learning: New Computational Modelling Techniques for Genomics." *Nat Rev Genet*, vol. 20, no. 7, 2019, pp. 389-403, doi:10.1038/s41576-019-0122-6.

5. Johnston, S. J. and J. S. Carroll. "Transcription Factors and Chromatin Proteins as Therapeutic Targets in Cancer." *Biochim Biophys Acta*, vol. 1855, no. 2, 2015, pp. 183-192, doi:10.1016/j.bbcan.2015.02.002.
6. Kelley, D. R. et al. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks." *Genome Res*, vol. 26, no. 7, 2016, pp. 990-999, doi:10.1101/gr.200535.115.
7. Klein, D. C. and S. J. Hainer. "Genomic Methods in Profiling DNA Accessibility and Factor Localization." *Chromosome Res*, vol. 28, no. 1, 2020, pp. 69-85, doi:10.1007/s10577-019-09619-9.
8. Klemm, S. L. et al. "Chromatin Accessibility and the Regulatory Epigenome." *Nat Rev Genet*, vol. 20, no. 4, 2019, pp. 207-220, doi:10.1038/s41576-018-0089-8.
9. Larson, M. H. et al. "Crispr Interference (Crispri) for Sequence-Specific Control of Gene Expression." *Nat Protoc*, vol. 8, no. 11, 2013, pp. 2180-2196, doi:10.1038/nprot.2013.132.
10. Moraes, F. and A. Góes. "A Decade of Human Genome Project Conclusion: Scientific Diffusion About Our Genome Knowledge." *Biochem Mol Biol Educ*, vol. 44, no. 3, 2016, pp. 215-223, doi:10.1002/bmb.20952.
11. Parson, W. et al. "Cancer Cell Line Identification by Short Tandem Repeat Profiling: Power and Limitations." *Faseb j*, vol. 19, no. 3, 2005, pp. 434-436, doi:10.1096/fj.04-3062fje.
12. Politz, J. C. R. et al. "The Redundancy of the Mammalian Heterochromatic Compartment." *Curr Opin Genet Dev*, vol. 37, 2016, pp. 1-8, doi:10.1016/j.gde.2015.10.007.
13. Quinlan, A. R. and I. M. Hall. "Bedtools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics*, vol. 26, no. 6, 2010, pp. 841-842, doi:10.1093/bioinformatics/btq033.
14. Rao, Y. S. et al. "Impact of Gc Content on Gene Expression Pattern in Chicken." *Genet Sel Evol*, vol. 45, no. 1, 2013, p. 9, doi:10.1186/1297-9686-45-9.
15. Rodriguez, J. et al. "Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity." *Cell*, vol. 176, no. 1-2, 2019, pp. 213-226.e218, doi:10.1016/j.cell.2018.11.026.
16. Rood, J. E. and A. Regev. "The Legacy of the Human Genome Project." *Science*, vol. 373, no. 6562, 2021, pp. 1442-1443, doi:10.1126/science.abl5403.
17. Sahu, R. K. et al. "The Mechanisms of Action of Chromatin Remodelers and Implications in Development and Disease." *Biochem Pharmacol*, vol. 180, 2020, p. 114200, doi:10.1016/j.bcp.2020.114200.
18. Shashikant, T. and C. A. Ettensohn. "Genome-Wide Analysis of Chromatin Accessibility Using Atac-Seq." *Methods Cell Biol*, vol. 151, 2019, pp. 219-235, doi:10.1016/bs.mcb.2018.11.002.
19. Singh, K. P. et al. "Mechanisms and Measurement of Changes in Gene Expression." *Biol Res Nurs*, vol. 20, no. 4, 2018, pp. 369-382, doi:10.1177/1099800418772161.
20. Teichmann, S. A. et al. "Uncovering the Interplay between DNA Sequence Preferences of Transcription Factors and Nucleosomes." *Cell Cycle*, vol. 11, no. 24, 2012, pp. 4487-4488, doi:10.4161/cc.22666.
21. Venkatesh, S. and J. L. Workman. "Histone Exchange, Chromatin Structure and the Regulation of Transcription." *Nat Rev Mol Cell Biol*, vol. 16, no. 3, 2015, pp. 178-189, doi:10.1038/nrm3941.
22. Zrimec, J. et al. "Learning the Regulatory Code of Gene Expression." *Front Mol Biosci*, vol. 8, 2021, p. 673363, doi:10.3389/fmolb.2021.673363.

Copyright: © 2023 Shreyes Balachandran, Diwakar Balachandran. All JEI articles are distributed under the attribution non-commercial, no derivative license (http://creativecommons.org/licenses/by-nc-nd/3.0/). This means that anyone is free to share, copy and distribute an unaltered article for non-commercial purposes provided the original author and source is credited.