Journal of Research High School

Volume 2023 (1) 70 - 75
ISSN 2688-3651 [online]

In Silico Drug Target Identification via Methylation Factors With Implementation in TP53 Liver Cancer Gene

Sahithi Pogula[1] *

[1]Hopkinton High School, Hopkinton, MA USA

Received July 6, 2022; Revised January 13, 2023; Accepted, January 24, 2023

Abstract

The limitations of drug discovery are infamous, with a single drug development setting back an institution millions of dollars and decades of time with only a 0.1% rate of success. Novel identification of target leads for drugs through currently unused epigenetic measures reduces these hurdles, incredibly expediting the drug pipeline. Here we identified two novel approaches for reducing liver cancer. In the first approach, this project focuses on reducing the hypermethylation of 5-methylcytosine (5mC) by modulating the 5-hydroxymethylcytosine (5hmC) biological pathway. Using pattern search and similarity indexes, Algorithm-1 can identify potential positions of pathway induction by comparing aberrant sequences with normal sequences and using a CpG island pattern trigger. Algorithm-1 identified the TP53 liver cancer gene at positions 33-35. In the second approach, this project focuses on inhibiting spliceosome factors that are responsible for 5mC production. Using machine learning methods and pattern search, such spliceosome factors are identified. Algorithm-2 identified the Prp9 protein. Both the proposed solutions provide new insight into a cure for Liver Cancer never previously explored through traditional methods of drug discovery.

Keywords: Epigenomics, Liver Cancer, DNA Methylation, Spliceosomes, Drug

1. Introduction

Drug discovery in the present day is defined by an array of hurdles not limited to an extensive and inefficiently utilized amount of funds, years and even decades of research, as well as incredibly low Phase 1 and clinical trial approval rates. Characterized by traditional methods of research and thousands of diseases gone untreated, present-day drug discovery is something truly negatively impacting society. Among the diseases included in this category is Hepatocellular Carcinoma or Liver Cancer (Fan, et al., 2018). Combining the facts that HCC is the third-deadliest cancer in the world and that there is no present cure on the market (only peripheral solutions deterring side effects have been previously identified) makes a deadly prediction of the future where the prevalence of HCC is only predicted to grow, coming as an economic, social, and moral burden to society. Epigenetics may possibly be the long-awaited solution to this highly sought-after problem.

Epigenetics refers to the study of heritable phenotype changes that do not involve alterations in the DNA sequence. Unlike genetic modifications, implications on epigenetics provide a novel tactic of gene expression modulation free of non-targeting side effects. DNA methylation, a naturally occurring epigenetic process, is closely correlated to embryonic development (Liu, et al., 2016), regulation of gene expression, X-chromosome inactivation, genomic imprinting, and genomic stability (Chen, et al., 2011). When located in the gene promoter, DNA

* Corresponding Author
sahithi.pogula@gmail.com

Advisor: Kristen Murphy
kmmurphy@hopkinton.k12.ma.us

methylation also acts to repress gene transcription. DNA methylation is catalyzed by a family of DNA methyltransferases (Dnmts) that transfer a methyl group from S-adenyl methionine (SAM) to the fifth carbon of a cytosine residue to form 5mC. Oftentimes, concentrated methylation sites are characterized as CpG sites (gene sites containing many C's and G's) (Chen, et al., 2011).

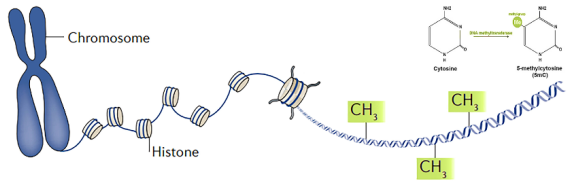

Figure 1. DNA methylation system naturally occurring via the utilization of the DNA methyltransferase (Wu, et al., 2017)

This natural process is made quite relevant to HCC studies due to its involvement in tumor progression. The development of HCC, in the current day, is marked by prevalent etiological conditions involving lifestyle as well as the environment. Recent studies, however, are coming to the surface stating otherwise (Hlady, et al. 2019). Aberrations in DNA methylation are now being directly correlated to molecular lesions of cancer cells as well as start points of tumor progression (Dreval, et al., 2019). The very regulated and maintained internal processes of DNA methylation was proven to be more radical and abnormal in HCC cases (Fan, et al., 2018). Hypermethylation was identified as a result.

Knowing the negative implications of such aberrant methylation, it is critical to identify these points of abnormality and revert, via demethylation, the hypermethylated cases of such points. The objective of this research was to accurately and efficiently identify loci of aberrant methylations for use in future applications of epigenetic modulation. Along with this, major motivations for this research included exploring new frontiers of epigenetic modulation, one being spliceosome factors and their inner entanglements to the onset of various carcinomas. Anticipated results for this research include studying the correlation between these various factors and methylation as well as building algorithms and models relevant to the topic. A primary end goal of this research was to identify loci of aberrant methylation and specific spliceosome factors by using a computational model on Liver Cancer, particularly the TP52 oncogene

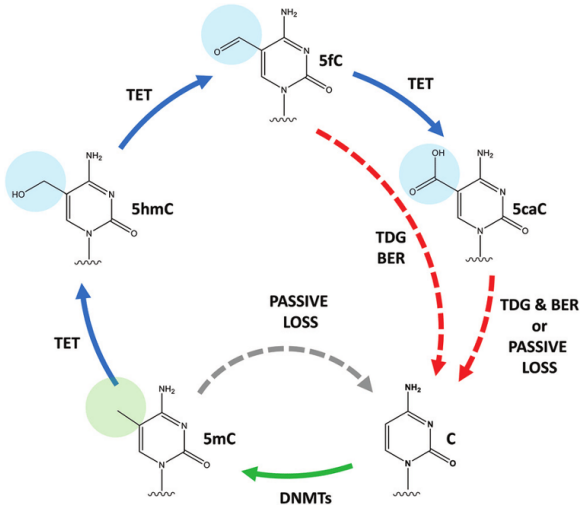

Figure 2. Multiple 5hmC DNA demethylation pathways. A methylated cytosine first starts its journey at a 5hmC (shown in the bottom left-hand corner). It is then oxidized by various TET family proteins to reach the state of 5caC. Through the BER process, it is then finally converted to an unmodified cytosine (shown in the bottom right-hand corner). (Chen, et al., 2011)

1.1 5hmC Pathway Demethylation Process

The 5-hydroxymethylcytosine (5hmC) pathway provides a viable solution for reducing such hypermethylation. As a whole, the 5hmC serves as a major role player in the roots of cancer development, not limited to simply Liver Cancer. In various types of carcinoma, aberrant methylation can be traced to be an underlying cause. Utilizing the 5hmC pathway, this aberrancy can be resolved. The exact points serving as tumor progressors oftentimes are characterized by extra methylation missing in a healthy case (Shi, et al., 2017). If these points are identified, the 5hmC pathway can be triggered artificially to revert methylation. The methyl moiety of 5mC is lost either passively during DNA replication or actively through enzymatic DNA demethylation. This project focuses on the latter. In a

process of oxidation reactions, conducted Ten-Eleven Translocation family proteins (TETs), the 5hmC is converted into a 5-carboxyl cytosine (5caC) (Dong, et al., 2012). Subsequently, the DNA base-excision repair (BER) pathway can also remove the methylated cytosine by filling in an unmodified cytosine. Thus, aberrancies can be reverted utilizing this pathway.

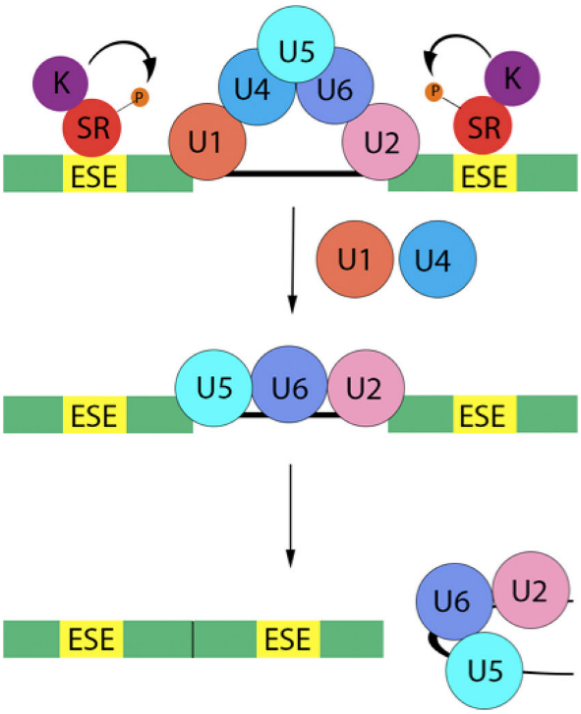

Figure 3. Alternative splicing mechanism as shown by spliceosome complex. Exonic splicing enhancers (ESEs) are present at the pre-mRNA splicing stage. The SR protein binds to ESE at the pre-mRNA splicing, but also are involved in mRNA export, genome stabilization, nonsense-mediated decay, and translation. (Wang, et al., 2015)

1.2 Spliceosome Factors

Modulating spliceosome behavior could also provide an effective solution to reducing the HCC-causing hypermethylation. This concept relies heavily on DNMT1-interacting RNAs (DiRs). DiRs are, essentially, the "blockers" between the DNMT1 and the gene. When a DiR is present, the DNMT1 is blocked from methylating at that site. In this way, DiRs regulate hypermethylation from occurring on genes. DiR use is not limited solely to Liver Cancer, but also expands broadly to encompass the general methylation status of every gene in the body. Their behavior is most distinctively linked (as well as scientifically proven) to the onset of HCC and Liver Cancer. The production of DiRs is ensured through the spliceosome and the various spliceosome factors it contains to conduct alternative splicing (Wang, et al., 2015). When different spliceosome factors are used, splicing as well occurs differently—alternative splicing. In abnormal cases, DiRs are not spliced due to the spliceosome factors present. In this case, due to the repression of DiRs, the DNMT1 has unregulated access to the gene, resulting in hypermethylation. Once again, such hypermethylation serves as a lesion point to tumor progression. If the exact spliceosome factor responsible for DiR repression is identified, inhibition of it could revert the aberrant state of the gene. Splicing pattern analysis and matching individual factors to disease cases could reveal the etiological factor.

2. Materials and Methods

2.1 Phase 1

This phase comprises point-reversion of hypermethylation utilizing the 5-hmC pathway. Point reversion refers to reversing the methylation status of one singular base pair (in this case C) from a methylated to unmodified state. Through this phase, an algorithm is developed which is able to identify hypermethylation directly from the gene in mere minutes. Results then can be used as navigational points for 5-hmC pathway triggering.

To start off, various methylated and unmethylated gene sequences were collected from the public databases GenBank and KEGG. Previously identified sites of methylation were also retrieved from the application iMethyl. This application was built by many scientists that worked towards identifying methylations on healthy genes for more than a decade. Program enabling libraries were then set up—NumPy, Pandas, TensorFlow. A Naïve Brute Force algorithm was then used to sort through sequences. This method relies on a rather primitive tactic of more trial and error. Although it is quite time-taking, it guarantees accurate results without the

chance of any outside bias. In a taken methylated sequence, if a "substring" of it was identified to contain the pattern "CGCG", the sequence is labeled. This "CGCG" pattern is characteristic of a CpG site where methylation is likely to present. In the sequence, if more than 3 substrings are identified to contain this pattern, methylation is evidently identified. Using Hamming Distance, a string similarity index, this methylated sequence is checked against an unmethylated sequence to detect erroneous methylation identification. The methylation sites identified are then checked against iMethyl's methylation sites, using Hamming Distance, to identify hypermethylation.

2.2 Phase 2

This phase comprises spliceosome factor inhibition to prevent DiR repression. Through this phase, an algorithm is developed which is able to identify the spliceosome factor responsible for the specific splicing pattern that occurs in the case of DiR repression. Results can then be used to artificially inhibit the specific etiological factor identified.

Pre-mRNA and mRNA were first collected from the public database GenBank. Spliceosome edit patterns were also compiled from various NCBI publishments. After ANN enabling programs, such as TensorFlow, were set up, basic tree nodes for a decision tree were built. The resulting decision tree was then trained multiple times using the data till cases of perfect fitting occurred. The Greedy Algorithm and Gini Index were used to quantify these accuracy rates. A Greedy Algorithm is a problem-solving heuristic that is able to find solutions to issues by inputting scenarios and evaluating them. This is important in accuracy rate calculations as it can serve as a comparison point for reconfirming solutions. The decision tree was then configured to output 3 top pattern matches for the splicing pattern seen in the pre-mRNAs, mRNAs, and the splicing factors. A Naïve Brute Force algorithm was then built to narrow down on the pattern match of the 3 top identified factors and the edit pattern. The end output of this algorithm is a splicing factor that is most likely to have acted upon the pre-mRNA to convert it to the mRNA used.

2.3 Phase 3

Through this phase, the algorithms developed in the previous phases were used on the liver-cancer-associated TP53 gene. The Phase 1 algorithm identified hypermethylation at positions 33-35 on the TP53 gene. The Phase 2 algorithm found Prp5, Prp9, and Prp11 in the U2 unit of the spliceosome with the most pattern match to the TP53 gene. The Prp9 factor was then narrowed down as most probably responsible for DiR inhibition.

3. Results

In order to quantify error for Phase 1, the error distance from the methylation site identified and the methylation site present was calculated using the Jaccard Index. Trends are given below

3.1 Phase 1 algorithm summary of tests

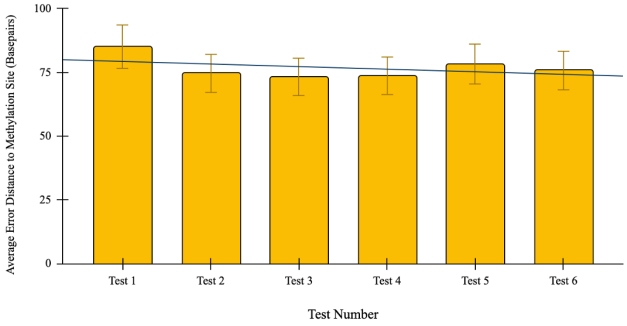

Figure 4. Average Error Distance to Methylation Site (set of 1,000 Baseparis)(units) vs. Test Number

As seen in the Phase 1 algorithm summary table, results stay relatively constant throughout out various tests and error distance remains low over the taken interval. These error distance values were used to calculate an 87.5% accuracy rate for the Phase 1 algorithm.

To test accuracy for the Phase 2 algorithm, the Greedy algorithm was used to assign values to how accurately a spliceosome factor was assigned to a given splicing pattern. Trends, given below, vary over tests due to the underfitting-overfitting aspect of machine learning algorithms.

3.2 Phase 2 algorithm summary of tests

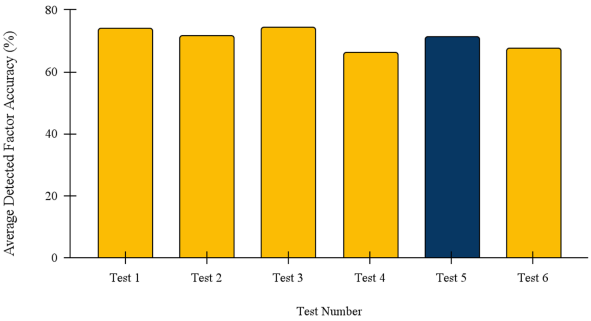

Figure 5. Average Detected Factor Accuracy (%) vs. Test Number

The Phase 2 algorithm summary table shows variance over tests, but a point of maximum accuracy and lowest error variances at Test 5 (highlighted in blue). A total algorithm accuracy rate of 74.3% was calculated from the above values.

100% accuracy was found for all individual Hamming Distance, Greedy, Naïve Brute Force, and Jaccard Index algorithms used throughout all phases. This accuracy rate refers to the baseline working of the algorithms. When a training set of data was used to initially test these algorithms before implementing real-world data, the algorithms produced accurate results matching exactly what was expected of them.

4. Discussion

Considering both algorithms developed in Phase 1 and Phase 2 relatively accurate (in comparison to current algorithms present on drug target identification), two variant propositions to reverting the aberrant case of HCC are founded. First, to induce the 5hmC pathway at positions 33-35. Second, to inhibit the Prp9 spliceosome factor as an acting protein in the spliceosome. Both tactics are predicted to reduce hypermethylation. Due to the fact that neither Phase 1 nor Phase 2 algorithms are predicted to have a 100% accuracy rate, these proposed solutions cannot be guaranteed. In this way, this research serves as more of a basis for novel HCC drug target identification rather than as a cemented solution. Despite this, it provides a tactic of *in silico* drug target identification never previously explored. While methylation and spliceosome factors have been previously studied, it had been work done over decades, only limited to a taken case study (Xu, et al., 2017). Such efficient case-specific analysis has never been done before. While modern-day drug target computational analyses are only able to narrow down to the top 50-100 leads, this research provides an accurate way to narrow down to 2-5 leads. Such narrowing down saves millions of dollars and years of time spent on drug research.

5. Conclusion

Through this research project, 2 specific algorithms were developed to revert the hypermethylation present in HCC cases. The Phase 1 algorithm focuses on identifying points of hypermethylation in a given sequence to trigger point-reversion utilizing the 5hmC demethylation pathway. The Phase 2 algorithm identifies spliceosome factors responsible for DiR inhibition, a characteristic of hypermethylation, using machine learning. Accuracy rates for the given phases came out to 87.5% and 74.3%, respectively. Algorithms developed in both phases were then applied to the TP53 gene to identify 2 specific propositions to reverting hypermethylation in HCC.

Acknowledgment

Annalisa DiRuscio, Ph.D., MD - Initial research on the role of DiRs.

References

Chen, Z. X., and Riggs, A. D. (2011). DNA methylation and demethylation in mammals. *J. Biol. Chem.* 286, 18347–18353. doi:10.1074/jbc.R110.205286.

Dong, E., et al. (2012). Upregulation of TET1 and downregulation of APOBEC3A and APOBEC3C in the parietal cortex of psychotic patients. *Transl. Psychiatry* 2, e159. doi: 10.1038/tp.2012.86

Dreval, K., et al. (2019). Gene Expression and DNA Methylation Alterations During Non-alcoholic Steatohepatitis-Associated Liver Carcinogenesis. Frontiers in Genetic. doi: 10.3389/fgene.2019.00486

Fan, G., et al. (2018). DNA methylation biomarkers

for hepatocellular carcinoma. *PubMed Central.* doi:10.1186/s12935-018-0629-5

Hlady, R., et al. (2019). Genome-wide discovery and validation of diagnostic DNA methylation-based biomarkers for hepatocellular cancer detection in circulating cell free DN. *PubMed Central.* 9(24), 7239–7250. doi:10.7150/thno.35573

Liu, J., et al. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nature Communications, 7*(866), 1–7.

Shi, D., et al. (2017). New Insights into 5hmC DNA Modification: Generation, Distribution and Function. *Frontier Genetics*. doi: 10.3389/fgene.2017.00100

Wang, Y., et al. (2015). Mechanism of alternative splicing and its regulation (Review). B*iomedical Reports*, 3, 152-158. doi: 10.3892/br.2014.407

Wu, X., & Zhang, Y. (2017). TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nature Reviews Genetics*, 18(9), 517–534.

Xu, Rh., et al. (2017) Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature Mater*, 16, 1155–1161. doi: 10.1038/nmat4997