A Symbolic Regression Approach to Hepatocellular Carcinoma Diagnosis Using Hypermethylated CpG Islands in Circulating Cell-Free DNA

Rushank Goyal
Betsos
rgoyal@betsos.org

Abstract

Purpose Hepatocellular carcinoma is the most common primary liver cancer, accounting for 90% of cases, and a major cause of death worldwide. Despite this, alpha-fetoprotein tests are the only blood-based diagnostic tools available, and their use is limited by their low sensitivity. DNA methylation changes, which have been implicated in a majority of cancers, offer an alternative method of diagnosis through measuring such changes in circulating cell-free DNA present in blood plasma.

Method A genetic programming-based symbolic regression approach was applied to gain the benefits of machine learning while avoiding the opacity drawbacks of "black box" models. The data included plasma samples from 36 patients with hepatocellular carcinoma as well as a control group of 55 that contained patients with and without cirrhosis. A 75-25 train-test splitting was done before training.

Results The symbolic regression methodology developed an equation utilizing the methylation levels of three biomarkers, with an accuracy of 91.3%, a sensitivity of 100%, and a specificity of 87.5% on the test data. All three biomarkers are differentially methylated in cancerous and non-cancerous samples. The performance matches prior research while providing the added benefits of transparency.

Conclusion Circulating cell-free DNA presents opportunities for minimally invasive early diagnosis of hepatocellular carcinoma, and utilizing transparent machine learning approaches like symbolic regression can allow accurate diagnosis by combining biological and mathematical principles. Future validation of the model obtained here on a larger and more diverse dataset can reveal the potential for such approaches in cancer diagnosis and open the way for further research.

Keywords Hepatocellular carcinoma, Symbolic regression, Circulating cell-free DNA

1 Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer, with approximately 782,000 new cases and 746,000 deaths (Ferlay et al., 2014). It is the second largest cause of cancer deaths in East Asia and sub-Saharan Africa as well as being the fastest rising cause of cancer-related death in the United States (Rawla, Sunkara, Muralidharan, & Raj, 2018). Risk factors such as chronic Hepatitis B (CHB), chronic Hepatitis C (CHC), nonalcoholic steatohepatitis (NASH), and aflatoxin exposure contribute to the development of hepatocelluar carcinoma, with CHB and CHC alone accounting for 75% of cases (Chen, 2018; Baecker, Liu, La Vecchia, & Zhang, 2018; Ahmed et al., 2019; Wang & Gribskov, 2019). With alpha-fetoprotein (AFP) tests being the only blood-based diagnostic tool currently available for HCC – the utility of which is hampered by its low sensitivity – there is an unmet need for an effective plasma test to enable early diagnosis (Xu et al., 2017).

CG dinucleotides, also known as CpG pairs, are under-represented in the human genome (21% of expected) due to the formation of methylcytosine through attachment of methyl groups to cytosine in CpG pairs; the methylcytosine spontaneously deaminates to thymine (Illingworth & Bird, 2009). There are, however, interspersions of largely nonmethylated sequences called CpG islands (CGIs) that possess large numbers of GC and CG nucleotides (Deaton & Bird, 2011). Hypo- or hyper-methylation of these CGIs acts as a prevalent molecular signature for most cancers, including HCC (Wen et al., 2015). In a number of studies, such methylation changes were detected years before other signs of cancerous development (Shi, Wang, & Caldwell, 2007).

Circulating cell-free DNA (cfDNA) is a term describing extracellular DNA found in body fluids like blood, sputum, urine, *et cetera* (Sun, An, & Yang, 2019). Its importance and promising outlook as a non-invasive marker for cancer has been recognized, utilizing genetic, methylation, and, to a lesser extent, quantitative analyses (Aarthy, Mani, Velusami, Sundarsingh, & Rajkumar, 2015; Kustanovich, Schwartz, Peretz, & Grinshpun, 2019). For HCC in particular, methylation analysis of cfDNA has yielded multiple diagnostic biomarkers, including but not limited to p15, p16, GSTP1 and RASSF1A (Ng, Di Costanzo, Terracciano, & Piscuoglio, 2018).

Symolic regression is a machine learning (ML) technique that attempts to create a mathematical expression to explain the target variable utilizing a range of basic mathematical functions, and, specifically, genetic programming-based symbolic regression (GPSR) is an implementation of SR that uses genetic programming to search the massive SR solution spaces effectively (Wilstrup & Kasak, 2021). GPSR has been able to perform well in deriving expressions for real-world applications in fields as disparate as biology, robotics, physics and finance (Orzechowski, La Cava, & Moore, 2018). Evidence suggests that it outperforms other ML techniques on smaller datasets, which is useful in clinical and biological contexts where data is often limited (Wilstrup & Kasak, 2021). Its transparency compared to other "black box" machine learning models also makes it more suited to application in clinical settings (Quinn, Jacobs, Senadeera, Le, & Coghlan, 2021; Price, 2018).

2 Materials and Methods

The data used for the analysis are indexed in the NCBI Gene Expression Omnibus with the accession number GSE63775. The data consist of plasma samples from 36 HCC and 55 control subjects, out of which 17 had cirrhosis (Wen et al., 2015). Values for each CGI in the samples are measured in methylated alleles per million mapped reads (MePM). Stratified random train-test splitting was performed in a 75-25 ratio (random seed of 1) to obtain the training and testing sets respectively.

Python 3.7.12 was used, with `pandas` and `numpy` for preprocessing, `matplotlib` and `seaborn` for plotting and visualization, `scikit-learn` for train-test splitting and model evaluation, and `gplearn` for training the symbolic classifier model. Note that a symbolic classifier works by first developing a symbolic regressor and then passing the output through a logistic function to produce a prediction value, which corresponds to 0 (negative) if less than 0.5, and 1 (positive) otherwise. The symbolic classifier was trained using `population_size`=2000 and `parsimony_coefficient`=0.01, with a random seed of 1 to ensure reproducibility (Andrew L. Beam, 2020).

3 Results

Figure 1 shows the output expression in a tree form. Equation 1 shows the final model as a standard mathematical equation, where chr_5, chr_{21} and chr_{19} represent the levels of the three CGIs identified as biomarkers by GPSR: chr5:92923487-92924497, chr21:40757602-40757900 and chr19:41531804-41532051, respectively. x is the output, and can be put through the logistic function to obtain the probability (p) of hepatocellular carcinoma, as shown in Equation 2. A probability less than 0.5 indicates that the absence of HCC is more likely, and vice versa.

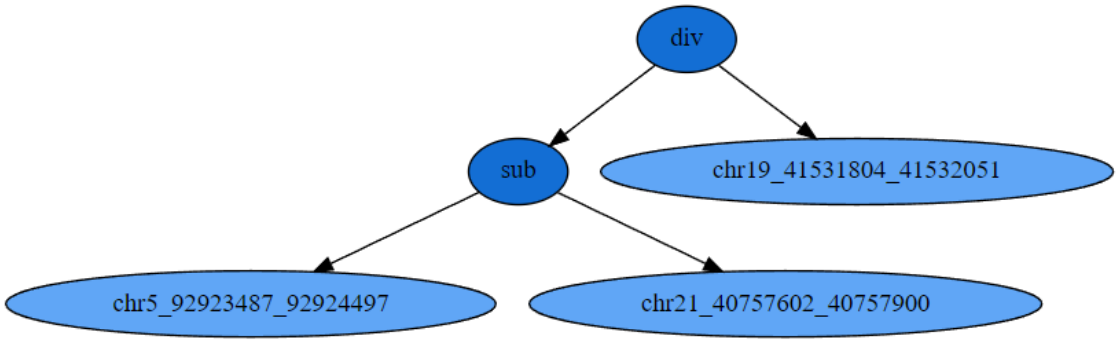

Figure 1: Output Equation as Tree

$$x = \frac{chr_5 - chr_{21}}{chr_{19}} \tag{1}$$

$$p = \frac{1}{1 + e^{-x}} \tag{2}$$

The results when utilizing this equation for classification on the testing dataset are shown in Figure 2 in the form of a confusion matrix. The accuracy comes out to be 91.3%, and the sensitivity and specificity are 100% and 87.5%, respectively.

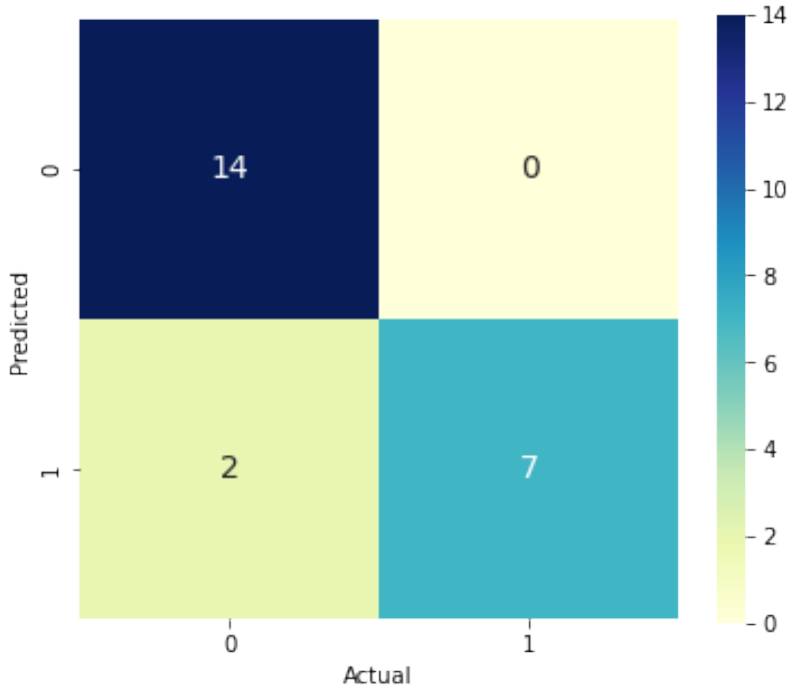

Figure 2: Confusion Matrix on Test Data

4 Discussion

Prior ML work in this area has attained similar performances, though it lacks the explainability of symbolic regression (Filho, Lacerda, & Pappa, 2020; Khalid, Tuszynski, Szlek, Jachowicz, & Mendyk, 2015). Utilizing a combination of random forest, the least absolute shrinkage and selection operator (LASSO) and logistic regression, Xu et al. (2017) obtained a sensitivity of 83.3% and a specificity of 90.5% on their validation set. Zhang et al. (2020) were able to get a sensitivity and specificity of 91.93% and 100% respectively using 11 biomarkers identified by a support vector machine (SVM) model. Another study also used an SVM to achieve a precision (positive predictive value) of 96% and a recall (sensitivity) of 86% (Gonçalves, Reis, Pereira-Leal, & Cardoso, 2021).

Multiple studies have identified blood plasma biomarkers without applying ML. Lin, Chen, Tang, and Wang (2005) showed that around 77% of HCC patients possess p16 methylation and 41% possess DAPK methylation. 55.7% of patients have a methylated CCND2 gene (Tsutsui et al., 2010). RASSF1A hypermethylation was present in 42.5% of patients, correlating with

tumor size (Yeo et al., 2005). Ji et al. (2014) identified MT1M and MT1G as biomarkers with high specificities of 93.5% and 87.1% respectively, though their sensitivities were low, at 48.8% and 70.2% respectively.

Compared to simple biomarker analysis, machine learning techniques display better performance, which is expected due to their more complex and opaque nature. The symbolic regression approach described in this paper, however, attains similar sensitivity and specificity to conventional ML approaches, is much more transparent and allows for mathematical reasoning of the results in a biological context, which can be a source of useful further research (Narayanan, Cruz Bournazou, Guillén Gosálbez, & Butté, 2022; Cardoso et al., 2020).

Since the data in this study contained cfDNA – which can be obtained from blood plasma – rather than analyzing liver tissue, the results present possibilities for minimally invasive diagnosis of HCC. Indeed, cfDNA holds potential for improved early cancer diagnosis in general (Stewart et al., 2018). Further analysis of the results on a larger and more diverse validation set as well as biological analyses of the three-biomarker signature are important, but the results obtained here show initial promise.

5 Data Availability Statement

The data are available in the National Center for Biotechnology Information's Gene Expression Omnibus (NCBI GEO) under the accession number GSE63775.

6 Declaration of Conflicts of Interest

No funding was received for conducting this study. The author has no relevant financial or non-financial interests to disclose.

7 Ethical Declarations

This research study was conducted retrospectively using human subject data made available by prior research. Ethical approval was not required, as confirmed by the institutional review board (IRB) of the Sri Aurobindo Institute of Medical Sciences in Indore, India.

References

Aarthy, R., Mani, S., Velusami, S., Sundarsingh, S., & Rajkumar, T. (2015, Dec). Role of Circulating Cell-Free DNA in Cancers. *Molecular Diagnosis & Therapy*, *19*(6), 339–350. doi: doi:10.1007/s40291-015-0167-y

Ahmed, O., Liu, L., Gayed, A., Baadh, A., Patel, M., Tasse, J., ... Arslan, B. (2019, Jan). The Changing Face of Hepatocellular Carcinoma: Forecasting Prevalence of Nonalcoholic Steatohepatitis and Hepatitis C Cirrhosis. *Journal of Clinical and Experimental Hepatology*, *9*(1), 50–55. doi: doi:10.1016/j.jceh.2018.02.006

Andrew L. Beam, P. (2020, Jan). Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*, *323*(4), 305–306. doi: doi:10.1001/jama.2019.20866

Baecker, A., Liu, X., La Vecchia, C., & Zhang, Z.-F. (2018, May). Worldwide Incident Hepatocellular Carcinoma Cases Attributable to Major Risk Factors. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)*, *27*(3), 205. doi: doi:10.1097/CEJ.0000000000000428

Cardoso, P., Branco, V. V., Borges, P. A. V., Carvalho, J. C., Rigal, F., Gabriel, R., ... Correia, L. (2020). Automated Discovery of Relationships, Models, and Principles in Ecology. *Frontiers in Ecology and Evolution*, *0*. doi: doi:10.3389/fevo.2020.530135

Chen, C.-J. (2018, Apr). Global elimination of viral hepatitis and hepatocellular carcinoma: opportunities and challenges. *Gut*, *67*(4), 595–598. doi: doi:10.1136/gutjnl-2017-315407

Deaton, A. M., & Bird, A. (2011, May). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022. doi: doi:10.1101/gad.2037511

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... Bray, F. (2014, Oct). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*(5), E359–E386. doi: doi:10.1002/ijc.29210

Filho, R. M., Lacerda, A., & Pappa, G. L. (2020, Jul). Explaining Symbolic Regression Predictions. In *2020 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1–8). IEEE. doi: doi:10.1109/CEC48606.2020.9185683

Gonçalves, E., Reis, M., Pereira-Leal, J. B., & Cardoso, J. (2021, Jun). DNA methylation fingerprint for the diagnosis and monitoring of hepatocellular carcinoma from tissue and liquid biopsies. *medRxiv*, 2021.06.01.21258144. Retrieved from `https://doi.org/10.1101/2021.06.01.21258144`

Illingworth, R. S., & Bird, A. P. (2009, Jun). CpG islands – 'A rough guide'. *FEBS Letters*, *583*(11), 1713–1720. doi: doi:10.1016/j.febslet.2009.04.012

Ji, X.-F., Fan, Y.-C., Gao, S., Yang, Y., Zhang, J.-J., & Wang, K. (2014, Apr). MT1M and MT1G promoter methylation as biomarkers for hepatocellular carcinoma. *World Journal of Gastroenterology : WJG*, *20*(16), 4723. doi: doi:10.3748/wjg.v20.i16.4723

Khalid, M. H., Tuszynski, P. K., Szlek, J., Jachowicz, R., & Mendyk, A. (2015, Dec). From Black-Box to Transparent Computational Intelligence Models: A Pharmaceutical Case Study. In *2015 13th International Conference on Frontiers of Information Technology (FIT)* (pp. 114–118). IEEE. doi: doi:10.1109/FIT.2015.30

Kustanovich, A., Schwartz, R., Peretz, T., & Grinshpun, A. (2019, Aug). Life and death of circulating cell-free DNA. *Cancer Biology & Therapy*, *20*(8), 1057–1067. doi: doi:10.1080/15384047.2019.1598759

Lin, Q., Chen, L.-b., Tang, Y.-m., & Wang, J. (2005, Dec). Promoter hypermethylation of p16 gene and DAPK gene in sera from hepatocellular carcinoma (HCC) patients. *Chinese Journal of Cancer Research*, *17*(4), 250–254. doi: doi:10.1007/s11670-005-0020-7

Narayanan, H., Cruz Bournazou, M. N., Guillén Gosálbez, G., & Butté, A. (2022, Feb). Functional-Hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. *Chemical Engineering Journal*, *430*, 133032. doi: doi:10.1016/j.cej.2021.133032

Ng, C. K. Y., Di Costanzo, G. G., Terracciano, L. M., & Piscuoglio, S. (2018). Circulating Cell-Free DNA in Hepatocellular Carcinoma: Current Insights and Outlook. *Frontiers in Medicine*, *0*. doi: doi:10.3389/fmed.2018.00078

Orzechowski, P., La Cava, W., & Moore, J. H. (2018, Jul). Where are we now? a large benchmark study of recent symbolic regression methods. In *GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 1183–1190). New York, NY, USA: Association for Computing Machinery. doi: doi:10.1145/3205455.3205539

Price, W. N. (2018, Dec). Big data and black-box medical algorithms. *Science Translational Medicine*, *10*(471), eaao5333. doi: doi:10.1126/scitranslmed.aao5333

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., & Coghlan, S. (2021, Aug). The three ghosts of medical AI: Can the black-box present deliver? *Artificial Intelligence in Medicine*, *102158.*. doi: doi:10.1016/j.artmed.2021.102158

Rawla, P., Sunkara, T., Muralidharan, P., & Raj, J. P. (2018). Update in global trends and aetiology of hepatocellular carcinoma. *Contemporary Oncology*, *22*(3), 141. doi: doi:10.5114/wo.2018.78941

Shi, H., Wang, M. X., & Caldwell, C. W. (2007, Sep). CpG islands: their potential as biomarkers for cancer. *Expert Review of Molecular Diagnostics*, *7*(5), 519–531. doi: doi:10.1586/14737159.7.5.519

Stewart, C. M., Kothari, P. D., Mouliere, F., Mair, R., Somnay, S., Benayed, R., ... Tsui, D. W. Y. (2018, Apr). The value of cell-free DNA for molecular pathology. *Journal of pathology*, *244*(5), 616. doi: doi:10.1002/path.5048

Sun, Y., An, K., & Yang, C. (2019, Jul). Circulating Cell-Free DNA. In *Liquid Biopsy*. IntechOpen. doi: doi:10.5772/intechopen.80730

Tsutsui, M., Iizuka, N., Moribe, T., Miura, T., Kimura, N., Tamatsukuri, S., ... Oka, M. (2010, Apr). Methylated cyclin D2 gene circulating in the blood as a prognosis predictor of hepatocellular carcinoma. *Clinica Chimica Acta*, *411*(7), 516–520. doi: doi:10.1016/j.cca.2010.01.004

Wang, S., & Gribskov, M. (2019, Jun). Transcriptome analysis identifies metallothionein as biomarkers to predict recurrence in hepatocellular cacinoma. *Molecular Genetics & Genomic Medicine*, *7*(6), e693. doi: doi:10.1002/mgg3.693

Wen, L., Li, J., Guo, H., Liu, X., Zheng, S., Zhang, D., ... Peng, J. (2015, Nov). Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients - Cell Research. *Cell Research*, *25*, 1250–1264. doi: doi:10.1038/cr.2015.126

Wilstrup, C., & Kasak, J. (2021, Mar). Symbolic regression outperforms other models for small data sets. *ArXiv e-prints*. Retrieved from `https://arxiv.org/abs/2103.15147v3`

Xu, R.-h., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., ... Zhang, K. (2017, Nov). Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma - Nature Materials. *Nature Materials*, *16*, 1155–1161. doi: doi:10.1038/nmat4997

Yeo, W., Wong, N., Wong, W.-L., Lai, P. B. S., Zhong, S., & Johnson, P. J. (2005, Apr). High frequency of promoter hypermethylation of RASSF1A in tumor and plasma of patients with hepatocellular carcinoma. *Liver International*, *25*(2), 266–272. doi: doi:10.1111/j.1478-3231.2005.01084.x

Zhang, Z.-M., Tan, J.-X., Wang, F., Dao, F.-Y., Zhang, Z.-Y., & Lin, H. (2020). Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. *Frontiers in Bioengineering and Biotechnology*, *0*. doi: doi:10.3389/fbioe.2020.00254