Identifying Carcinogenic Multi-hit Combinations using Weighted Set Cover Algorithm

(Extended Abstract, PhD Forum, ICPP-2018)

Sajal Dash[1], Nick Kinney[2], Robin Varghese[2], Harold Garner[2], Wu-chun Feng[1,3], and Ramu Anandakrishnan*[2]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2]Biomedical Sciences, Edward Via College of Osteopathic Medicine, Blacksburg, VA, USA
[3]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

June 7, 2018

Abstract

Different combinations of just two to eight hits (genetic mutations) are estimated to be required for tumorigenesis. We develop a weighted set cover based method to identify set of all these combinations for various cancer types.

Research Objective Disruptions in certain molecular pathways due to combinations of genetic mutations (hits) are known to cause cancer [2, 1, 4, 3]. Although different combinations of just two to eight hits are estimated to be required for carcinogenesis, the specific combinations of mutations responsible for the vast majority of cancers have not been identified. Due to a large number of mutations present in tumor cells, experimentally identifying these combinations is not possible except in very rare cases. Current computational approaches, on the other hand, simply do not search for specific combinations of multiple genetic mutations. Instead, current algorithms search for sets of driver mutations based on mutation frequency and mutational signatures. Individually these driver mutations may increase the risk of cancer; however, they generally do not result in carcinogenesis, without specific additional mutations. We aim to identify all these 2-8 hit combinations for 17 cancer types with enough number of samples by combining algorithm insights into efficient GPU acceleration.

Research Progress Here, we present a fundamentally different approach for identifying mutations most likely to be the cause of cancers: we search for combinations of

*ramu@vt.edu

carcinogenic mutations (multi-hit combinations). By avoiding the convolution of different driver mutations associated with different individual instances of cancer, multi-hit combinations may be able to identify the specific cause for each cancer instance. We mapped the problem of identifying a set of multi-hit combinations to a weighted set cover problem. Although finding an optimal solution to this problem is computationally intractable, there exist algorithms for finding an approximate solution. We use a greedy algorithm to identify sets of multi-hit combinations for seventeen cancer types for which there are at least two hundred matched tumor and blood-derived normal samples in the cancer genome atlas (TCGA). When tested using an independent validation dataset, these combinations are able to differentiate between tumor and normal tissue samples with 91% sensitivity (95% Confidence Interval (CI) = $89 - 92\%$) and 93% specificity (95% CI = $91 - 94\%$), on average for seventeen cancer types. Accuracy was robust to different randomly selected training and test sets. The combinations identified by this method, with experimental validation, can aid in better diagnosis, provide insights into the etiology of various cancer types, and provide a rational basis for designing targeted combination therapies. We have completed initial design of parallel WSC algorithm using GPU accelerators and identified potential research subproblems such as organize unstructured genomic data for regularized memory access.

Current Status in the Ph.D. Program

I have completed my fourth year in the Ph.D. program. I have passed the qualifier exam and preparing for preliminary exam in a couple of months. I am working toward completing my Ph.D. defense by May, 2019. My thesis involves big data processing using high-dimensional geometric algorithms in a streaming and incremental settings. I am combining algorithmic data analysis with scalable HPC programming.

References

[1] Ramu Anandakrishnan. Identifying multi-hit combinations by cancer type to provide a rational basis 1 for targeted combination therapy. *Unpublished Manuscript*, 2017.

[2] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[3] Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, 2015.

[4] Xinan Zhang and Richard Simon. Estimating the number of rate limiting genomic changes for human breast cancer. *Breast cancer research and treatment*, 91(2):121–124, 2005.