Research Statement – Mark D.M. Leiserson

As a computer scientist and interdisciplinary researcher working in the field of computational biology, I am inspired by the potential benefits of computational approaches to biological research and driven by my passion for solving difficult computational challenges. In that vein, the majority of my research has been in cancer genomics, which provides diverse and difficult computational problems with important applications. The cost of DNA sequencing is declining rapidly such that measuring a tumor's mutations is sometimes cheaper than drug regimens, providing the opportunity for personalized cancer therapy tailored to a tumor's mutation profile. However, a major challenge to developing new personalized therapies for cancer is the size and complexity of the datasets: The number of hypotheses far outstrips the resources to test them experimentally. Thus, computational approaches that automatically generate and prioritize hypotheses are an urgent priority. These approaches need to be computationally efficient – scaling to increasingly large cancer genomics datasets – and statistically rigorous in making predictions.

The main problem I have focused on in my research is distinguishing the few causal mutations responsible for cancer from the numerous random mutations in a cohort of tumors. The ability to identify these key mutations has many benefits, including for identifying new drug targets and personalizing treatment. I have developed computational approaches for this problem – utilizing techniques from graph theory, statistics, and combinatorial optimization – and applied these approaches in multiple collaborations with cancer biologists.

Looking forward, I am well positioned to continue to make an impact in cancer genomics. Already, advances in cancer research powered by DNA sequencing have led to new research questions and computational challenges. The expertise I have developed in cancer genomics has prepared me to continue and expand on my current research to face these challenges. Furthermore, new biotechnology and initiatives promise to generate datasets that can be used to address key open biological questions. As an interdisciplinary researcher with experience applying my methods to real datasets, I look forward to to collaborating with biologists to address computational challenges presented by these new cutting edge technologies.

1 Current Research

My current research is focused on developing automated methods for identifying the mutations responsible for cancer, known as *driver mutations* (Figure 1).[1] This is a difficult problem for two main reasons. First, driver mutations are relatively rare. A typical tumor has from three to eight driver mutations, and dozens to thousands of passenger mutations that have no consequence for cancer. Second, cancer is a heterogeneous disease, such that different combinations of driver mutations cause cancer in different patients. Thus, some driver mutations may be observed in only a few patients, even in large tumor cohorts.

My work addresses these challenges by searching for combinations of driver mutations. This is motivated by a common explanation for mutational heterogeneity across tumors: Mutations target groups of interacting genes called pathways that perform critical cellular functions, and each pathway can be perturbed in numerous ways. Thus, the problem becomes identifying the driver mutations in genetic pathways. I have made two types of contributions

[1] Please see http://maxleiserson.com for more information about my current research.

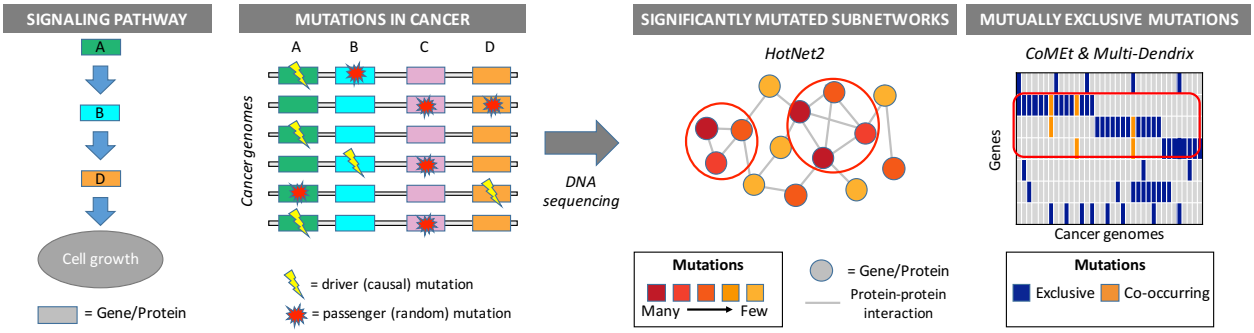

Figure 1: (left) The *driver* mutations responsible for cancer target key signaling pathways, such that different combinations of mutations cause cancer in different patients. (right) My current research is focused on identifying driver mutations from DNA sequencing data by searching for *combinations* of mutations occurring in patterns expected for pathways. In that vein, I developed algorithms to search for mutations that cluster in a biological interaction network or are mutually exclusive across cancer genomes (outlined in red).

in this research area: (1) new algorithms for identifying combinations of driver mutations; and, (2) a platform for integrating and annotating mutation data.

1.1 Algorithms for Identifying Combinations of Driver Mutations

I developed three algorithms to find the members of genetic pathways that are significantly mutated in a cohort of tumors. These methods have several advantages over other approaches. First, in contrast to methods that test databases of known pathways and gene sets, these algorithms can identify novel pathways, as well as crosstalk between existing pathways. Second, the algorithms identify multiple pathways simultaneously, corresponding to the widely accepted theory that cancer is caused by mutations to multiple pathways.

The algorithms utilize different knowledge of how driver mutations target pathways in order to perform the search. The HotNet2 algorithm [Leiserson et al., 2015c] searches for significantly mutated subnetworks of a protein-protein interaction network, which are expected to correspond with pathways targeted by driver mutations. The key feature of HotNet2 is that it encodes the topology of the network and the mutations in the genes *simultaneously* using a heat diffusion process, allowing it to overcome the biases of the network and the mutation data. The CoMEt algorithm [Leiserson et al., 2015a,b] and its prequel the Multi-Dendrix algorithm [Leiserson et al., 2013] are combinatorial optimization algorithms to search for sets of genes with mutually exclusive mutations, a pattern expected for driver mutations in genes in the same pathway. One key advantage of Multi-Dendrix and CoMEt is that they require no prior knowledge, and thus do not need to take into account the biases of any prior information and work even when prior information is not available.

In addition to developing these algorithms, I showed that they perform well in practice. I benchmarked each of the algorithms on simulated and real data, showing they outperform related methods. I also applied them to real datasets and in collaboration with biologists. I led the application of HotNet2 to The Cancer Genome Atlas (TCGA) Pan-Cancer dataset, where HotNet2 identified subnetworks overlapping pathways and complexes both with well-known and less well-characterized roles in cancer [Leiserson et al., 2015c], as well as cancer studies from TCGA [Hoadley et al., 2014; TCGA Research Network, 2014, 2015]. I demonstrated

Multi-Dendrix and CoMEt by identifying members of multiple cancer pathways in glioblastoma tumors, and applied CoMEt in a pan-cancer study [Lu et al., 2015]. I also applied earlier versions of these algorithms as part of collaborations with cancer biologists [Kanchi et al., 2014; Kandoth et al., 2013; Ley et al., 2013].

1.2 Annotating, Integrating, and Visualizing Mutation Data

The computational challenges for leveraging large public datasets (e.g. from TCGA) for follow-up studies is a major hindrance to cancer genomics research. While researchers can inform the study of private cohorts by taking advantage of the large number of samples in public datasets, this remains computationally challenging. Furthermore, follow-up studies revealing new insights into these public datasets remain scattered in the literature and disconnected from the datasets.

To address these challenges, I developed MAGI (`http://magi.brown.edu`), a web application for annotating, integrating, and visualizing public and private mutation data [Leiserson et al., 2015d]. MAGI allows users to create custom visualizations of the combined public and private datasets by uploading their own mutation data with a simple web form and integrating it with large public mutation datasets. Thus, researchers can quickly compare and contrast the mutations in their small tumor cohorts with the mutations in the large public datasets. Moreover, MAGI allows users to view and add literature annotations of mutations, and is initialized with a database of nearly 40,000 annotations.

2 Future Research

I plan to continue developing algorithms to tackle computational problems in cancer genomics. Many important and difficult computational problems remain unsolved, and new datasets will provide new computational challenges that must be overcome to answer key questions in cancer research. In that vein, my future research will continue to focus on the problem of identifying combinations of driver mutations, as well as investigate research questions for predicting cancer drug targets. I also aim to take advantage of new cancer genomics datasets such as from the International Cancer Genome Consortium (ICGC).

2.1 Identifying Combinations of Driver Mutations

One problem I am interested in is identifying combinations of driver mutations to regulatory elements in the intergenic regions that make up 99% of the human genome. This question remains open despite the great strides that have been made in developing methods to identify driver mutations. Most research to date has focused on genetic mutations in cancer, but mutations can also alter cellular function by targeting elements that regulate gene function in the intergenic regions of the genome. However, the data for performing a comprehensive analysis of mutations to regulatory elements has only recently become available through whole cancer genome sequencing efforts from consortia such as the ICGC.

As such, I aim to investigate whether incorporating regulatory mutations can inform the search for combinations of driver mutations by taking advantage of these new datasets. This is a particularly pressing research question as single-gene analyses of regulatory mutations have identified relatively few regulatory driver mutations. However, it may be possible

to leverage the database of regulatory interactions from the recently concluded ENCODE project to evaluate whether or not a regulatory mutation is a driver. One major challenge with this project is to search for the genetic and regulatory driver mutations by analyzing both types of mutations and both physical and regulatory interactions simultaneously.

2.2 Predicting New Cancer Drug Targets

I am also interested in expanding the scope of my research to new problems in developing personalized cancer treatment. One such area I am interested in is identifying new targets for cancer drugs. New technology and datasets may make it possible for computational approaches to identify a new class of drug targets. The key insight is that if a cancer cell has a mutation disrupting the function of gene X, then it may be vulnerable to drugs that target a gene Y that is compensating for the change in X's function. This is a critical insight because there are no known drugs to target many important cancer genes. However, several key computational challenges remain.

One major challenge is to identify and prioritize potential drug targets computationally. Recent studies have published high-throughput experiments that measure how vulnerable different human cancer cells are to drugs targeting thousands of different genes, but these datasets are extremely noisy. However, it may be possible to utilize genetic networks or pathways to identify and prioritize the drug targets, because the targets are expected to be functionally related to a mutated gene. Furthermore, the recent advent of CRISPR/Cas9 for precise gene editing may lead to larger and higher quality versions of these datasets, making methods for analyzing them a priority.

References

K. A. Hoadley et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

K. L. Kanchi et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature Communications*, 5:3156, 2014.

C. Kandoth et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471): 333–339, 2013.

M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Computational Biology*, 9(5):e1003054, 2013.

M. D. M. Leiserson, H-T. Wu, F. Vandin, and B.J. Raphael. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology*, 16(1):160, 2015a.

M. D. M. Leiserson, H-T. Wu, F. Vandin, and B.J. Raphael. CoMEt: A Statistical Approach to Identify Combinations of Mutually Exclusive Alterations in Cancer. In *RECOMB*, pages 202–204. Springer, 2015b.

M. D. M. Leiserson et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2015c.

M. D. M. Leiserson et al. MAGI: visualization and collaborative annotation of genomic aberrations. *Nature Methods*, 12(6):483–484, 2015d.

T. J. Ley et al. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

C. Lu et al. Patterns and Functional Implications of Rare Germline Variants across 12 Cancer. 2015. To appear: *Nature Communications*.

TCGA Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, 2014.

TCGA Research Network. Comprehensive molecular characterization of papillary renal cell carcinoma. 2015. To appear: *New England Journal of Medicine*.