Using Statistical Divergence to Cluster Genes by Unseen Variant Probabilities

Thesis Overview

Ethan Ashby

9/10/2020

Project Overview

Identifying the primary site (or tissue of origin) for cancers is an important problem in precision oncology. Cancer is a malady characterized by runaway mutation, and so the problem of primary site classification is addressed by analyzing a tumor's profile of non-heritable genomic mutations, or **somatic variants**. Somatic mutation analysis traditionally focuses on a small number of frequently occurring or "hotspot" mutations known to associate with particular tissues. However, this approach ignores potential signal encoded by the diverse collection of rare somatic alterations. For context, greater than 90% of mutations in The Cancer Genome Atlas (a clinical panel of over 10,000 whole exome sequenced tumors) are singletons, meaning they appear once across all sequenced tumors. Thus, cancer is a malady characterized by rare, unique mutation, and traditional somatic mutation analysis neglects this vast preponderance of rare mutational data and the tissue-specific signals that they encode.

Previous research (Chakraborty et al. (2019)) applied a Bayesian, nonparametric method (*Smoothed Good-Turing Frequency Estimation*) developed in computational linguistics to rare somatic variant mutations in major cancer genes. This method produced a **probability of encountering a previously uncatalogued or unseen mutation** in a future sequenced tumor for each gene. In several genes, these unseen variant probabilities showed tissue specific patterns, meaning that these previously uncatalogued mutation probabilities contain information that could improve cancer primary site classification.

My thesis will try to extend this analysis from 400 select cancer genes to all (~20,000) protein coding genes in the cancer genome. The majority of genes were excluded from Chakraborty et al's analysis due to sparse mutation data. We can address this problem by aggregating the mutation data within gene groups, or **metagenes**.

The chief technical challenge of this project is what criteria do we use to build these gene groups. We would like to create gene groups where the tissue specificty of its unseen variant probabilities is maximized. To do this, I will represent each gene as a [2xk] discrete probability distribution, where $[1, i]$ corresponds to the probability of encountering a previously unseen mutation in the gene under tissue type i and $[2, i]$ as its complement. I will use **statistical divergence** to calculate pairwise "distances" between the probability distributions of different genes. The resulting distance matrix will form the basis for a clustering scheme to generate the gene groups with enriched tissue-specific signal.

For my thesis, i will explore different divergence metrics developed in statistics and information theory (such as the **Jensen-Rényi divergence**) for measuring pairwise "distances" between probability distributions. After a literature seach, I will conduct a simulation study to test the suitability and performance of different divergence metrics in grouping these genes. I hope to apply incorporate these divergence metrics into a clustering scheme to generate gene groups of high clinical value.

References

Chakraborty, Saptarshi, Arshi Arora, Colin B. Begg, and Ronglai Shen. 2019. "Using somatic variant richness to mine signals from rare variants in the cancer genome." *Nature Communications* 10: 1–9. https://doi.org/10.1038/s41467-019-13402-z.