Bibliography First Draft

Ethan Ashby

September 25 2020

1 Oncology and Math references

Cancer is a malady caused and characterized by mutation. But mutational heterogeneity is a major problem that needs adjustment for in cancer genetics. Analysis of 27 cancer types showed that median mutation frequency varied upwards of 1000-fold between cancer types [9]. It's also important to note that this heterogeneity has important evolutionary justification, arising from the stochastic nature of Darwinian evolution [1].

The data I will be working with is somatic variant mutation data in 6676 non-hypermutated human tumors acquired through the NCI The Cancer Genome Atlas (TCGA) project [7].

The Bayesian, nonparametric methods that I will use to calculate tissue specific mutation probabilities are introduced in [6] and more in-depth derivations and justification are provided in the supplementary materials to that paper [5].

Good-Turing methods allow for estimation of probabilities of unseen events, but require smoothing of regions of vastly different accuracy. A simple logarithmic smooth was found to provide the best smoothing of the raw frequencies [8]

Once these probabilities are generated, we need a rigorous way to measure "distance" between the probabilities associated with each gene. Statistical divergence is a way to measure distance between probability distributions. I will use [2] as a good jumping off point for learning about divergence, as this textbook contains a number of useful definitions and examples of using divergence for inference problems. [4] will provide a useful overview of the interpretations and connections between many common distance/divergence metrics. I will the annotated bibliography [3] as a resource to identify some papers using divergence in the context of applied problems. [11] may be of particular interest, since this study uses f-divergences to select diagnostically-relevant (class label specific) genes from microarray datasets. A interesting family of divergences that warrants further exploration are the Rényi divergences, which are introduced here [12]. The information theoretic interpretation of some f-divergence is provided here [10].

2 References

References

[1] Pan-cancer Analysis et al. "Pan-cancer analysis of whole genomes". In: *Nature* 578.February (2020). DOI: 10.1038/s41586-020-1969-6.

[2] Basu Ayanendranath. *Statistical Inference: The Minimum Distance Approach*. ISBN 978-1420099652. Boca Raton, Florida: Chapman and Hall/CRC Press, 2011.

[3] Michèle Basseville. "Divergence measures for statistical data processing — An annotated bibliography". In: *Signal Processing* 93 (2013), pp. 621–633. DOI: 10.1016/j.sigpro.2012.09.003.

[4] Sung-hyuk Cha. "Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions". In: *International Journal of Mathematical Models and Methdos in Applied Sciences* 1.4 (2007).

[5] Saptarshi Chakraborty et al. "Supplementary Information for " Using Somatic Variant Richness to Mine Signals from Rare Variants in the Cancer Genome "". In: *Nature Communications* 10 (2019), pp. 1–20.

[6] Saptarshi Chakraborty et al. "Using somatic variant richness to mine signals from rare variants in the cancer genome". In: *Nature Communications* 10 (2019), pp. 1–9. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13402-z. URL: http://dx.doi.org/10.1038/s41467-019-13402-z.

[7] Kyle Ellrott et al. "Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Article Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines". In: *Cell Systems* 6 (2018), pp. 271–281. DOI: 10.1016/j.cels.2018.03.002.

[8] William A Gale and Murray Hill. "Good-Turing Smoothing Without Tears". In: *Journal of Quantitative Linguistics* 2 (1995), pp. 1–24.

[9] Michael S Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499 (2013), pp. 214–218. DOI: 10.1038/nature12213.

[10] Friedrich Liese and Igor Vajda. "On Divergences and Informations in Statistics and Information Theory". In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

[11] Pradipta Maji. "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data". In: *IEEE Transactions on Biomedical Engineering* 56.4 (2009), pp. 1063–1069.

[12] Alfréd Rényi. "On measures of entropy and information". In: *The 4th Berkeley Symposium on Mathematics, Statistics and Probability*. Berkeley, CA: University of California Press, 1960, pp. 547–561.