Assignment 1: Abstract

Bioinformatics research is focused on gaining a thorough understanding of the genetics that underpin cancer growth and progression. Its success can lead to important mechanistic, diagnostic, and therapeutic breakthroughs. The discovery of genes that cause tumours to appear after a mutation is a big step in this direction. Recent developments in computational biology have demonstrated the ability to classify cancer driver genes by integrating genetic overview statistics that reflect the mutational burden in genes with biological networks, such as protein–protein interaction networks. These methods superimpose summary statistics on network nodes, followed by unsupervised propagation of node scores across the network. However, in this unsupervised environment, no knowledge of well-known cancer genes is used, which may be a valuable resource in identifying novel cancer drivers.

Introduction:

Cancer is a disease characterised by uncontrolled cellular growth caused by genetic changes in cancer driver genes such as mutations, copy number variations, or gene fusions. These changes can affect the gene's activity as well as its cellular function, and they can be classified as activating (proto-oncogenes) or loss of function (tumor suppressor genes and DNA repair genes). One of the key aims of oncogenic research is to identify cancer driver genes, which helps with mechanistic, diagnostic, and therapeutic insights.

Statistical measures that determine the mutational burden of a gene may be used to identify cancer genes. However, the extensive mutational heterogeneity complicates such analyses: several genes are mutated in a small number of samples, and only a few genes exhibit substantial mutation throughout several samples. This phenomenon makes it difficult to distinguish between genes that only bear passenger mutations and cancer genes that are seldom mutated. One reason for the variation of candidate genes is that genes interact in a variety of pathways and protein complexes, and a cell's cancerous potential is a result of the pathway's destruction rather than the mutation of a single gene, but not usually a mutation in a single gene along the pathway.

This interaction-based approach to cancer biology has recently been adopted in research: a combination of biological networks and summary statistics that quantify each gene's association with cancer has aided in the discovery of novel cancer driver genes. Nodes represent genes, and edges represent relationships between adjacent genes in those networks. There are a plethora of biological networks that are derived from various sources and span various scales. PPI networks, in particular, are an intriguing representation of gene interactions since they often integrate data from various data sources, tissues, and molecular processes at various scales, Those PPI networks, however, are far from complete, and our understanding of them is skewed toward well-studied genes. Information contamination refers to the fact that well-studied (cancer) genes have more interactions in networks. The potential for a significant effect on network analysis perception must be understood and accounted for, as it can cause results to be muddled.

Methods that use networks to represent molecular relationships frequently begin by superimposing scores on the nodes. These scores assess the gene's marginal association with the disease of interest. The MutSig P-value is a popular choice for representing each gene's association with cancer: it is a meta-P-value that describes whether there is a statistically significant difference in I the mutational burden, (ii) mutation clustering, and (iii) the functional impact of mutations in a gene between healthy and cancer tissues. A plethora of methods for analysing such gene scores in conjunction with network information to identify altered subnetworks of genes within the original network have been developed. They are broadly classified as clustering methods, which aim to find modules of associated genes that cluster together in a network, and methods that detect altered subnetworks by using network diffusion or network propagation. Both methods are based on the shared paradigm that genes influencing the same phenotype interact in a network. Network propagation methods, in particular, have demonstrated success in identifying novel cancer driver genes. Network propagation methods, on the other hand, exploit the flow of information between genes along paths, and the longer the paths, the more information is diluted. This makes detecting cancer genes that do not share short paths with other cancer genes more difficult.

NetSig is another successful approach that does not rely on this assumption. It detects cancer genes solely based on a network's local neighbourhood of genes. The calculation of an empirical P-value for each gene, which defines the aggregation of genes in the direct neighbourhood with low MutSig P-values, is at its core. The scale of a gene's local neighbourhood affects the NetSig statistic due to information contamination. To avoid this, NetSig employs a variety of permutation schemes that take the node degree into account, thereby correcting for this bias.

We present a novel approach to supervised cancer gene classification based on cancer gene annotations from the Cancer Gene Census in the COSMIC database. This is accomplished by resolving the problem of identifying novel cancer driver genes as a node-classification problem in an interaction network.