Causal Discovery for Cancer

Tasks Introduction

Cancer is a genetic disease. It is caused by changes in genes that control the way cells grow and multiply. Somatic genomic alterations (SGAs) in DNA are largely responsible for initiating cancer. In particular, SGAs cause cancer through differential expression of genes (DEGs). Those SGAs that cause DEGs are good candidates as drivers of cancer. The task is to identify SGAs that cause the DEGs in breast cancer.

Cancer Dataset and Demos

The dataset provided for this Hackathon is the Mini-TCGA Cancer Dataset from Center for Causal Discovery.

The Mini-TCGA Cancer Dataset contains data of Somatic Genomic Alteration (SGA) and Differentially Expressed Gene (DEG).

SGA indicates both nonsynonymous somatic mutations and copy number alterations; a gene sequence variable is coded as 1 if altered (SGA) and 0 if not (not SGA). The selected 6 SGAs in the dataset are well known breast cancer drivers.

DEG indicates which genes are differentially expressed, relative to a baseline (e.g., normal tissue); DEG = 1 means a gene is differentially expressed, e.g. expression more than 2 SD from mean of normal-cells distribution. Otherwise DEG = 0. The selected 60 DEGs are most frequently regulated by the SGAs according to the TCI algorithm.

This Jupyter notebook shows how to load toy data and the Mini-TCGA Cancer data, as well as a few demos to call an off-the-shelf algorithm (e.g. fast greedy equivalence search (GES)) for causal discovery, i.e. predict a Directed Acyclic Graph (DAG) from the data.

Significance

Causal discovery for cancer is an important and challenging biomedical problem. Computational methods and machine learning are driving the development of new algorithms that meet the needs of biomedical researchers. The ability to discover and model causal relationships accurately is a key step in more fully realizing precision cancer diagnosis, prognosis, and therapy.

Helpful tools and resources

Please feel free to explore these tools and resources. It is feasible to choose a method from these references to address the task. You may also consider further development upon the chosen method and apply it to the tasks. Alternatively, you may develop your own methods for the task.
https://causal-learn.readthedocs.io/en/latest/index.html
https://www.cmu.edu/dietrich/causality/causal-learn/
https://bd2kccd.github.io/docs/

https://www.ccd.pitt.edu/biomedical-science/

More details about greedy equivalence search (GES)

- Searches over equivalence classes of Bayesian networks (patterns).
- Uses a Bayesian score.
- It is a greedy algorithm that has a forward stepping phase and a backward stepping phase.
- If a Bayesian network (BN) is generating the data (and it is a perfect map), GES is guaranteed to find the data generating BN in the large sample limit.
- FGES is an optimized version of GES for a single processor and can be parallelized to run on multiple processors

More data sources

You may find more data in the Cancer Genome Atlas (TCGA) dataset for breast and head and neck cancer. Include: measurements of somatic mutations, copy number alterations, DNA methylation, gene expression, microRNA expression, reverse phase protein array (RPPA) data, a.k.a functional genomic data.

The Cancer Genome Atlas Program (TCGA): https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
UCSC Xena: https://xenabrowser.net/datapages/
Broad GDAC Firehose: https://gdac.broadinstitute.org/

More challenging tasks

The goal is to develop computational methods to model causal relationships between different biological factors and cancer. This will help discover the genomic drivers of tumors and the cell signaling pathways that are being abnormally affected, driving the development of cancer. In particular, causal machine learning can be exploited to analyze The Cancer Genome Atlas (TCGA) data to discover the causes for cancers, such as breast and head and neck cancer.

- One task is to determine which somatic genetic alterations are causal drivers of a given tumor.
- Another task is to identify which somatic genetic alterations are causal drivers shared across different tumors.
- Further, it is also interesting to discover causal relationships among signaling proteins in cancer pathways.

Cancer Genomics Overview

This section provides some background that might be interesting to know. Cancer is a group of diseases caused by changes in DNA that alter cell behavior, causing uncontrollable growth and malignancy. These abnormalities can take many forms, including DNA mutations, rearrangements, deletions, amplifications, and the addition or removal of chemical marks. These changes can cause cells to produce abnormal amounts of particular proteins or make misshapen proteins that do not work as they should. Oftentimes, a combination of several genomic alterations work together to promote cancer.

Basic Background on Cancer Genetics

Cancer is a genetic disease. It is caused by changes in genes that control the way cells grow and multiply. Cells are the building blocks of your body. Each cell has a copy of your genes, which act like an instruction manual.

Genes are sections of DNA that carry instructions to make a protein or several proteins. Scientists have found hundreds of DNA and genetic changes (also called variants, mutations, or alterations) that help cancer form, grow, and spread.

Cancer-related genetic changes can occur because:

- random mistakes in our DNA happen as our cells multiply
- our DNA is altered by carcinogens in our environment, such as chemicals in tobacco smoke, UV rays from the sun, and the human papillomavirus (HPV)
- they were inherited from one of our parents

DNA changes, whether caused by a random mistake or by a carcinogen, can happen throughout our lives and even in the womb. While most genetic changes aren't harmful on their own, an accumulation of genetic changes over many years can turn healthy cells into cancerous cells. The vast majority of cancers occur by chance as a result of this process over time.

Genetic alterations can be inherited from one's parents, caused by environmental factors, or occur during natural processes such as cell division. The changes that accumulate over one's lifetime are called acquired or somatic changes and account for 90–95% of all cases of cancer.

The field of cancer genomics is a relatively new research area that takes advantage of recent technological advances to study the human genome, meaning our full set of DNA. By sequencing the DNA and RNA of cancer cells and comparing the sequences to normal tissue such as blood,

scientists identify genetic differences that may cause cancer. This approach, called structural genomics, may also measure the activity of genes encoded in our DNA in order to understand which proteins are abnormally active or silenced in cancer cells, contributing to their uncontrolled growth.

Once cancer-causing changes are identified, scientists can gain a better understanding of the molecular basis of cancer growth, metastasis, and drug resistance. This is done using clinical data that describes how patients responded to cancer treatment, laboratory experiments using cell lines and model organisms, and big data analysis techniques. Putting large genomic datasets together and sharing them with researchers worldwide is an increasingly important strategy for cancer research, as this boosts the power of the data and opens new opportunities for discovery. Scientists at the National Institutes of Health (NIH), as well as around the world, are working diligently to identify genetic changes underlying cancer, determine their roles in tumor development and metastasis, and harness these findings to fight cancer.

How do genetic changes cause cancer?

Genetic changes can lead to cancer if they alter the way your cells grow and spread. Most cancer-causing DNA changes occur in genes, which are sections of DNA that carry the instructions to make proteins or specialized RNA such as microRNA.

For example, some DNA changes raise the levels of proteins that tell cells to keep growing. Other DNA changes lower the levels of proteins that tell cells when to stop growing. And some DNA changes stop proteins that tell cells to self-destruct when they are damaged.

For a healthy cell to turn cancerous, scientists think that more than one DNA change has to occur. People who have inherited a cancer-related genetic change need fewer additional changes to develop cancer. However, they may never develop these changes or get cancer.

As cancer cells divide, they acquire more DNA changes over time. Two cancer cells in the same tumor can have different DNA changes. In addition, every person with cancer has a unique combination of DNA changes in their cancer.

What kinds of genetic changes cause cancer?

Multiple kinds of genetic changes can lead to cancer. One genetic change, called a DNA mutation or genetic variant, is a change in the DNA code, like a typo in the sequence of DNA letters. Some variants affect just one DNA letter, called a nucleotide. A nucleotide may be missing, or it may be replaced by another nucleotide. These are called point mutations. For example, around 5% of people with cancer have a point mutation in the *KRAS* gene that replaces the DNA letter G with A.

This single letter change creates an abnormal KRAS protein that constantly tells cells to grow. Cancer-causing genetic changes can also occur when segments of DNA—sometimes very large ones—are rearranged, deleted, or copied. These are called chromosomal rearrangements.

For example, most chronic myelogenous leukemias (a type of blood cancer) are caused by a chromosomal rearrangement that places part of the *BCR* gene next to the *ABL* gene. This rearrangement creates an abnormal protein, called BCR-ABL, that makes leukemia cells grow out of control.

Some cancer-causing DNA changes occur outside genes, in sections of DNA that act like "on" or "off" switches for nearby genes. For example, some brain cancer cells have multiple copies of "on" switches next to genes that drive cell growth.

Other DNA changes, known as epigenetic changes, can also cause cancer. Unlike genetic variants, epigenetic changes (sometimes called epimutations) may be reversible and they don't affect the DNA code. Instead, epigenetic changes affect how DNA is packed into the nucleus. By changing how DNA is packaged, epigenetic changes can alter how much protein a gene makes.

Some substances and chemicals in the environment that cause genetic changes can also cause epigenetic changes, such as tobacco smoke, heavy metals like cadmium, and viruses like Epstein-Barr virus.

Why Genomics Research Is Critical to Progress against Cancer

The study of cancer genomes has revealed abnormalities in genes that drive the development and growth of many types of cancer. This knowledge has improved our understanding of the biology of cancer and led to new methods of diagnosing and treating the disease.

For example, the discovery of cancer-causing genetic and epigenetic changes in tumors has enabled the development of therapies that target these changes as well as diagnostic tests that identify patients who may benefit from these therapies. One such targeted drug is vemurafenib (Zelboraf), which was approved by the Food and Drug Administration (FDA) in 2011 for the treatment of some patients with melanoma who have a specific mutation in the *BRAF* gene as detected by an FDA-approved test.

Over the past decade, large-scale research projects have begun to survey and catalog the genomic changes associated with a number of types of cancer. These efforts have revealed unexpected genetic similarities across different types of tumors. For instance, mutations in the *HER2* gene (distinct from amplifications of this gene, for which therapies have been developed for breast, esophageal, and gastric cancers) have been found in a number of cancers, including breast, bladder, pancreatic, and ovarian.

Researchers have also shown that a given type of cancer, such as breast, lung, and stomach, may have several molecular subtypes. For some types of cancer, the existence of certain subtypes had not been known until researchers began to profile the genomes of tumor cells.

The results of these projects illustrate the diverse landscape of genetic alterations in cancer and provide a foundation for understanding the molecular basis of this group of diseases.

References

[CausalLearn Github](https://github.com/py-why/causal-learn)
[Center for Causal Discovery](https://www.ccd.pitt.edu/biomedical-science/)
Bailey MH, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018.
Cooper G, Cai C, Lu X: Tumor-specific Causal Inference (TCI): A Bayesian Method for Identifying Causative Genome Alterations within Individual Tumors. bioRxiv 2017.

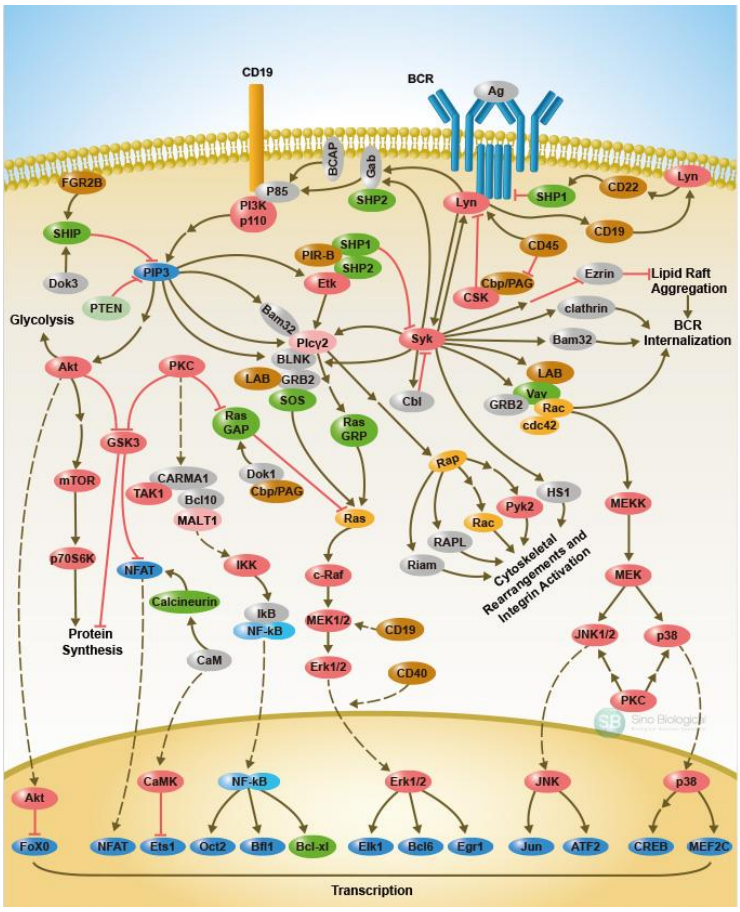
https://www.cancer.gov/about-cancer/understanding/what-is-cancer#cell-differences

https://www.cancer.gov/about-cancer/causes-prevention/genetics

https://www.cancer.gov/about-nci/organization/ccg/cancer-genomics-overview

https://www.cancer.gov/research/areas/genomics

https://www.cancer.gov/about-cancer/causes-prevention/genetics


CD19
BCR
Ag
BCAP
Gab
FGR2B
P85
PI3K p110
SHP2
Lyn
SHP1
CD22
Lyn
SHIP
CD19
CD45
PIR-B
SHP1
SHP2
Dok3
PIP3
Etk
Ezrin
Lipid Raft Aggregation
Cbp/PAG
CSK
PTEN
Bam32
clathrin
Glycolysis
BCR Internalization
Plcγ2
Syk
Bam32
Akt
PKC
BLNK
LAB
GRB2
SOS
LAB
Vav
GRB2
Rac
cdc42
Cbl
Ras GAP
Ras GRP
GSK3
Rap
mTOR
CARMA1
TAK1
Bcl10
MALT1
Dok1
Cbp/PAG
Ras
Pyk2
HS1
MEKK
Rac
RAPL
Riam
Cytoskeletal Rearrangements and Integrin Activation
p70S6K
NFAT
IKK
c-Raf
MEK
Calcineurin
IkB
NF-kB
MEK1/2
CD19
JNK1/2
p38
Protein Synthesis
CaM
Erk1/2
CD40
PKC
Akt
CaMK
NF-kB
Erk1/2
JNK
p38
FoX0
NFAT
Ets1
Oct2
Bfl1
Bcl-xl
Elk1
Bcl6
Egr1
Jun
ATF2
CREB
MEF2C
Transcription