Chapter 5. Computational Biology

5.1. Statistical Analysis of Cancer*

Central Dogma of Molecular Biology

It is important to understand the processes behind the genetic variables of interest. The most important concept is known as the central dogma of molecular biology. This describes that process of by which genetic information is interpreted and eventually proteins are created that are required for all life. We begin with double stranded DNA. DNA is composed of a long chain of two different base pairs (4 bases). The pairs are Adenine and Thymine, and Guanine and Cytosine. Thus at any point on the DNA strand you can have a choice of 4 different bases. Some sections of the DNA are genes. Each gene in the DNA is unzipped into one strand and then transcribed onto which is known as messenger RNA. The RNA can then be read and translated into protein. I.E. We progress as from DNA to mRNA to Protein.

Copy Number and Gene Expression

Two genetic variables of interest that are involved in the central dogma are estimated copy number and gene expression. The estimated copy number of a gene gives the estimated number of copies of a specific gene exist with in a samples genome. Normally, we would expect to see a copy number of 2 but often we see copy numbers as low as 0 and as high as 5 or 6. Most of the time this does not harm the particular patient and we expect to see a certain amount of copy number variation within any person, but it can sometimes be correlated with the incidence of cancer.

Gene expression is a variable that indicates the estimates amount of mRNA that one observes in a sample. It is often difficult to directly measure the amount of a given protein within a sample, but we often witness a correlation between the amount of mRNA in a sample and the amount of protein that is created.

Goals

Our data is comprised of 89 samples which have estimated copy number data. 41 of these samples also have gene expression data. The goals of this project is to examine the copy number and gene expression data of these sample and see if there are certain sections in the genome that have are consistently over or under expressed and possibly have an abnormal copy number. The next steps would be to examine the known processes of these genes to see if they are incorporated in any cancer-related functions such as cell growth.

Copy Number Analysis

First, the raw copy number data was imported into an application called Illumina Genome Studio. Physically, the data is obtained by measuring how bright a certain chemical fluoresces when mixed with the DNA. Genome Studio then estimates the copy numbers based on this data. Genome Studio was then used to generate a frequency plot which displays how frequently each gene was amplified or deleted,

In addition, to confirm our findings an R script was written that also generates a similar plot,

These findings were also confirmed using a program called GISTIC which is genetic analysis software which is part of the Broad Institutes Gene Pattern Server.

The next step in our analysis of the copy number data involved examining any genes that had extreme amplifications or extreme deletions. I.E. any genes in which the estimated copy number was less than .5 or greater than 5. A script was written in the R statistical package in order to detect the frequency with which the genes had major amplifications or deletions.

Future Work

The results of this research will be published in a future publication. It would be necessary to get more samples, especially ones with both copy number and gene expression data to create a full gene expression pattern. ???

Acknowledgements

Dr. Rudy Guerra, Matthew Burnstein, Dr. Chris Man, Dr. Ching Lau, Alexander Yu, Powell-Brown Lab, Rice university, Texas Children's Hospital

This Connexions module describes work conducted as part of Rice University's VIGRE program, supported by National Science Foundation grant DMS–0739420.

References

1. Rebecca Dent, Kathleen I. Pritchard, Wedad M. Hanna, Harriet K. Kahn, Carol A. Sawka, Lavina A. Lickley, Ellen Rawlinson, Ping Sun and Narod, Steven A. (2007). Triple-Negative Breast cancer: clinical Features and Patterns of Recurrence. *Clinical cancer Research*, *13*, 4429-4434.

5.2. Analysis of miRNA and mRNA associated with Epithelial Mesenchymal Transition*

Bioinformatics PFUG

Lauren Kirton

July 2010

Special thanks to:

This Connexions module describes work conducted as part of Rice University's VIGRE program, supported by National Science Foundation grant DMS–0739420.

Dr. Rudy Guerra

Dr. Sendurai Mani

Joe Taube

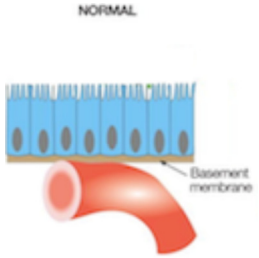
Introduction

One of the most prominent areas of research today is that of cancer study. There exists thousands of studies relating to the causes and cures of cancer. Molecular biologists at the M.D. Anderson Cancer Center research lab believe they have found a breakthrough in the cause of metastatic cancer. This breakthrough potentially answers the most important question of cancer research-how can we prevent cancer? The research at the M.D. Anderson Cancer Center involves a focus on the genetic level, with biological functions discovered only in the past few years. More information can be found in Ref. ???.

Introduction to Cancer

Over 80% of reported cancer cases are carcinomas. A carcinoma is an invasive malignant tumor composed of transformed, mutated epithelial cells.

Figure 5.1.

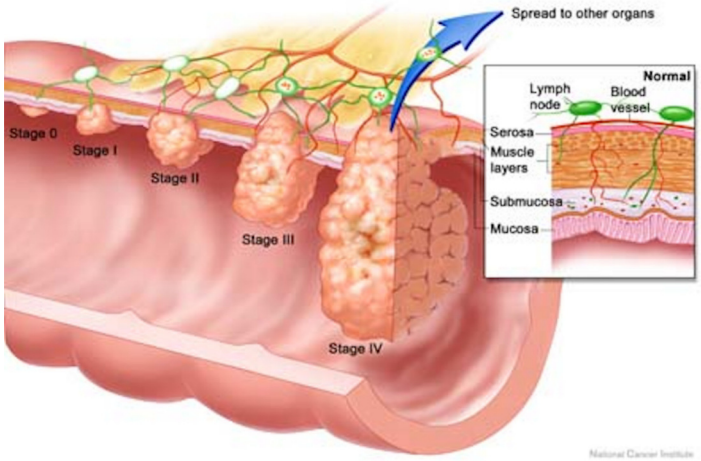


Example of Normal Cells

Normal epithelial cells line the cavities and surfaces of the structures that make up the body. These cells are strict and rigid in structure, held tight together by intercellular junctions. Epithelial cells are specialized, meaning they are made for a specific purpose or structure (i.e. skin tissue, organ tissue, etc.).

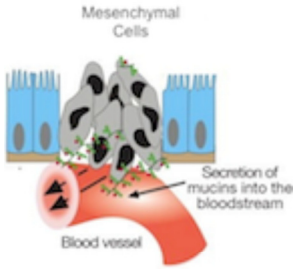
A carcinoma forms when mutated epithelial cells begin to form a tumor in a localized region. There are five stages of carcinoma, describing the extent or severity of an individual's cancer. Common elements of staging systems include location of the primary tumor, tumor size and number of tumors, lymph node involvement, cell type and tumor grade, and presence of metastasis.

Figure 5.2.



In higher stages of carcinoma, a higher chance of metastasis that is likely to occur. Carcinoma is very treatable or curable. Carcinoma tumors can be removed with surgery, or killed with radiation, chemotherapy and hormonal therapy.

Figure 5.3.



Example of Mesenchymal Cells

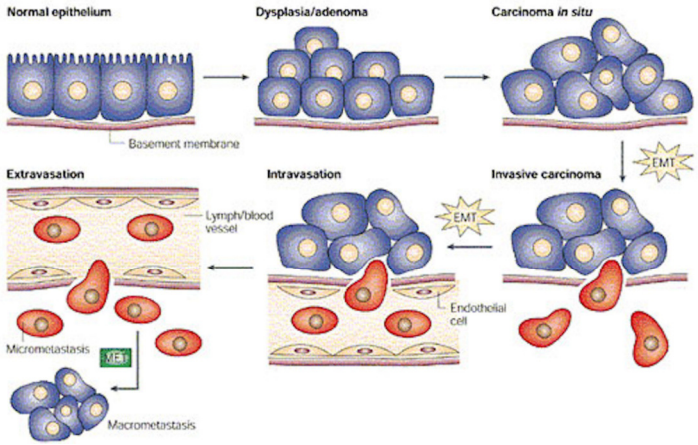
Fatality associated with cancer stems from metastasis of the cancer. Metastatic cancer refers to cancer cells that spread from a primary epithelial based tumor to another location in the body where these cells form a metastatic based tumor. These cells, known as mesenchymal cells, possess stem cell like properties and can differentiate into a variety of cell types. These cells are spindly-like and very spread apart, allowing for easy transfer throughout the body. Mesenchymal cells have long proliferation and are unspecialized, making the spread of cancer elsewhere in the body fairly easy.

Epithelial Mesenchymal Transition (EMT)

Epithelial Mesenchymal Transition (EMT) is a process of biological cells in which epithelial cells express losses of cell adhesion, repression of E-cadherin expression and allow the cells increased mobility. There are several transcription factors that induced EMT. Transcriptional factors Snail and Slug are repressors of E-cadherin, and

the expression of these two transcription factors induces EMT. The transcription factors S, T, G and B are known to induce EMT. When these four transcription factors undergo EMT, expression of FOXC2 is observed, an important factor known to induce EMT and regulate metastasis. The process of EMT begins when epithelial cells in a high stage carcinoma undergo some process that mutate these cells into mesenchymal cells. These mesenchymal cells then enter the blood stream through capillaries that cover the tumor and travel throughout the body. These cells can then exit the bloodstream and return to the carcinogenic epithelial cells and begin to form a second tumor elsewhere in the body.

Figure 5.4.



The Process of EMT

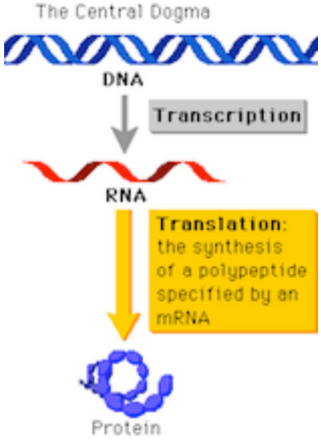
It is known that there are several main causes of cancer and metastasis. Environment and lifestyle habits are among the leading causes of cancer in patients. Genetics also plays an important role in the cause of cancer in metastasis. Looking deeper into the genes known to cause metastasis, molecular biologists at M.D. Anderson Cancer Center hope to zero-in on what induces metastasis.

The Central Dogma

The Central Dogma deals with the transfer of sequences that code genetic information and forms the backbone of molecular biology. The Central Dogma is characterized by four main steps: replication, transcription, translation and splicing.

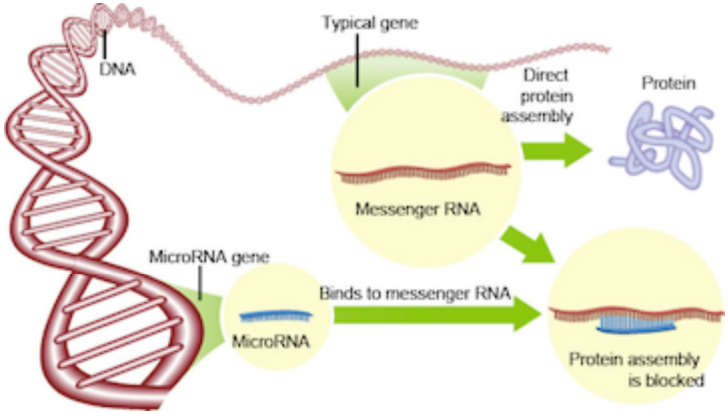
The central dogma of molecular biology begins with transcription, the process in which information in a section of DNA is transferred to a messenger RNA (mRNA). DNA sequence is copied by RNA polymerase to produce a complementary RNA strand. This strand replaces the thymine of DNA with uracil, a main difference between RNA and DNA. Transcription is usually the first step that leads to the expression of genes. Following transcription is translation, where the mRNA is translated, following the unique genetic code, into a functional protein. This protein is what regulates genes expression, overall controlling EMT.

Figure 5.5.



It is believed that microRNA (miRNA) can control EMT ultimately by regulating the proteins that induce or repress the EMT process. These miRNA are short post-transcriptional regulators that bind to target messenger RNA, most commonly resulting in gene-silencing. These miRNA are on average 20-24 nucleotides in length, but play an important role in genetic processes. In short, these bind to complementary strands of messengerRNA, preventing the functional protein from being formed, controlling the regulation of EMT.

Figure 5.6.



Protein Inhibition with miRNA

The Project

M.D. Anderson used a database to compile the interactions between different miRNA and mRNA for four particular genes known to induce EMT. The compiled data gave the specific target mRNA, miRNA, the location of the mRNA-miRNA interaction on the gene, and the interaction energy between the mRNA and miRNA. The interaction energy measures how well an miRNA binds to the mRNA site. For this study, an interaction $\geq$ abs(x) was considered useful information. The constant, x, being a certain cutoff limit. The compiled data can contain multiple mRNA- miRNA interactions. Each gene has a dataset for miRNAs that are up-regulated, implying that the greater amount of miRNA will have more protein inhibition, repressing the gene, and a dataset for down-

regulated miRNAs, in which the miRNAs are repressed, implying greater gene expression since the protein inhibition did not occur. We want to determine which miRNA's and mRNA's are significant for further research.

Using the program, R, I sorted each dataset to have each unique miRNA then number of mRNA that interact with each miRNA. From this new data, i then sorted the data into unique miRNA and the unique corresponding mRNA. Applying a filter to exclude interaction energies $\leq$ abs(x), the number of unique miRNA- unique mRNA interactions drastically reduced. Our hypothesis that certain miRNA-mRNA interactions appear significant is hard to validate-the datasets themselves lack similar data between them. We also have only one set of data as a whole to compare, and thus cannot have inference to which we cannot prove that some data is statistically significant. Comparing these results to other data from different samples will allow such hypothesis to be confirmed or rejected.