Thanks John.

An addition to point 3: The ~2,000 genes in the dataset are known cancer genes, and genes selected based on their functional association with cancer genes.

Marieke

On 04/02/2015 06:41 AM, John Quackenbush wrote:

On 3/31/2015 3:44 PM, Ping Chen wrote:

    Dear Marieke,

    Nice meeting you last week. My biology knowledge is mainly from high school. I wonder whether you mind to answer some questions I have.

    <!--[if !supportLists]-->1. <!--[endif]-->The data we have is gene-mutation data, right? Mutated genes are not necessarily expressed, even expressed, they don't always lead to cancer, right? Are these two the main sources of uncertainty/noise? So they will put an upper bound on how much we can achieve.

    The data we have are mutational data but I would not limit this by looking at expression. In most cases, the genes are expressed--although we don't have expression data on most samples.

    A bigger constraint on what can be achieved is likely the background mutation rate in cancer--which likely differs

    <!--[if !supportLists]-->2. <!--[endif]-->Do we know that certain combinations of expressed genes will cause cancer with a high probability? Will the order of mutations or expressions matter?

    The order if mutation is thought to be important in the development of many cancers, but that this order (and the set of mutations) may be different in different cancers of in different subtypes of "the same" cancer. In the data we have, it may be same to assume that the relevant mutations have occurred since these are full-blown tumors. It is also worth noting that there are methods we could use based on the frequency of common mutations, to try to map out that order.

    <!--[if !supportLists]-->3. <!--[endif]-->For the ~2000
genes in the dataset, how much do we know about their correlation with cancer? Nothing at all or there is some knowledge embedded in medical literature?

There is quite a bit of information in the biomedical literature. However, we recognize that it is partial information. A good part of what we want to do in this project is to go beyond what has been published to identify new subtypes. If we already had defined all of the subtypes based on mutation, then we could just filter the data we have and "rediscover" the same subtypes that have been reported.

4. For the 22 cancers, do we know anything about their subtypes?

Some like breast cancer has a fair consensus as to the subtypes, but these are generally defined by gene expression, not necessarily by mutation. In other cancers we have very poor consensus as to what the subtypes are.

5. What is the main goal of the project? Using gene mutations to predict cancer?

Our hope is that NMF will allow us to discover new subtypes defined by a common core of mutations and that these subtypes will have clinically relevant differences such as in survival.

Thanks,

Ping

From: Marieke Kuijjer [mailto:mkuijjer@jimmy.harvard.edu]
Sent: Tuesday, March 31, 2015 9:39 AM
To: Wei Ding; Heather Selby
Cc: Henry Lo; Dawei Wang; Ping Chen; 强继朋; John Quackenbush
Subject: Re: Compute the Cure Meeting 3-26-2015 == tumor types for class project

Dear Wei,

Thanks for your message. I think this will be an interesting selection of tumors. Good luck with preparing the class.

Best,
Marieke

On 03/30/2015 09:03 PM, Wei Ding wrote:

Hello Marieke,

As we discussed, I plan to select a subset of the cancer data and use it in my Applied Machine Learning class. I
would like to guide students to run different experiments to see whether we can find any interesting patterns.

Based on the Cophenetic Correlations (slide 25 of the meeting minutes prepared by Henry and Dawei), the following 4 tumors seem to be interesting:

1. CESC, cervical squamous cell carcinoma (slide 25, top row, rightmost)
2. PRAD, prostate adenocarcinoma (slide 25, 4th row, leftmost)
3. READ, rectum adenocarcinoma (slide 25, 4th row, 2nd left)
4. UCEC, uterine corpus endometroid carcinoma (slide 25, 5th row, rightmost)

The changes of Cophenetic Correlations of the 4 tumor types provide 4 unique patterns.

What do you think about the 4 tumors? I plan to assign the term project tomorrow at 4 PM.

Thanks,

Wei Ding

From: Wei Ding
Sent: Monday, March 30, 2015 8:54 PM
To: Marieke Kuijjer; Heather Selby; John Quackenbush
Cc: Henry Lo; Dawei Wang; Ping Chen; 强继朋
Subject: Compute the Cure Meeting 3-26-2015

Dear All,

The meeting minutes for March 26 Thursday meeting can be accessed from:

http://www.cs.umb.edu/~ding/out/compute_the_cure/previous%20meetings/

Our next steps are
1. The whole data set
   NMF: Find the best K using Cophenetic Correlations
   1) using average of 50 runs on consensus matrices
   2) visualization of the CC plots
   3) meta mutated-genes
   4) visualization of W & H along with mutated-gene profiles
   Contrast Data Mining: What are the best contrast patterns within the clusters? (focus on patterns longer than 1 genes)

2. Inside each tumor
   Which tumor types are further separable? NMF: Find the best K using Cophenetic Correlations
   1) using average of 50 runs on consensus matrices
   2) visualization of the CC plots
   3) meta mutated-genes
   4) visualization of W & H along with mutated-gene profiles
   Do the subtypes within individual tumor types overlap with clusters of the whole set?

3. Variations of LDA
   Will investigate whether LDA (Latent Dirichlet Allocation) can find more meaningful clusters.

Best Regards,

Wei Ding