Host- Microbiome Omics Integration for Cancer Analysis And Diagnostics

# Background

Currently trying to combine data from TCMA and this data set.

TCMA:

<https://www.sciencedirect.com/science/article/pii/S1931312820306636>

<https://tcma.pratt.duke.edu/>

Gene 5000 features for all cancers:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5728678/

Now trying to use SNF data

# To Do

class in balance is horrible because a lot of samples fallout. Try to either manually get data from TCGA or use other data set to solve this.

Firstly if the same problem exists for all the data sets on TCMA. Then try the other file from the same source of genes and see if it’s the same (they are the same patient wise).

(The problem does exist for all of them. And the other file for genes is the same. STAD has the most normal samples at 39. Also, SNF only has colon as an overlapping set, and it has less gene samples than the one from Akash.)

Try to find preprocessed TCGA gene data with more samples otherwise preprocess it yourself

Possibly promising:

<https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Stomach%20Cancer%20(STAD)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443>

<https://ucsc-xena.gitbook.io/project/faq/advanced-data-and-datasets>

Otherwise

<https://portal.gdc.cancer.gov/repository>

https://portal.gdc.cancer.gov/repository?facetTab=cases&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22content%22%3A%7B%22field%22%3A%22cases.project.project\_id%22%2C%22value%22%3A%5B%22TCGA-STAD%22%5D%7D%2C%22op%22%3A%22in%22%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.submitter\_id%22%2C%22value%22%3A%5B%22%2A4188%2A%22%2C%22%2A4280%2A%22%5D%7D%7D%5D%7D&searchTableTab=cases

(a lot of STAD normal samples are missing even here? nevermind)

Firehose is also no good. Xeno also doesn’t have the samples I need for stomach. The studies in my documents also don’t really have the data.

Looks like I have to do it manually, but the datafiles are really big. Try to download specific samples and genes also, looks like I need authorization?

So, it seems like Xenos has some samples that the other one doesn’t equal amount of controls in total, so I could try to append them and then get the preprocessing to be the same: It doesn’t have any control samples which the author doesn’t have, only tumor samples. Where the bacterial samples correct and the other samples? YES

Look closer into SNF set. Doesn’t have any control samples and only has aa ones

Could also try with another datatype first. With Zeno’s, methylation and CNV have zero normal overlap for STAD.

Also, see about other data sets from Zeno’s: READ has zero from the four and COAD has 3

the one from Akash also barely has normal samples, only two in the methylation data set, and zero overlap. I asked them for additional data.

I explored the normal versus tumor of the TCMA data sets, I should also do that for the gene expression data set.

I should also investigate the normal samples missing to see if they really are in TCGA. I only check for one sample. It seems to be there, but in another file. Also, the xenos data doesn’t have all the cases of STAD for example and also not as many normal samples.

I can now try to find other DNA data sets, but also other microbial data sets.

Also, try to see if data set of the big TCM a TCGA study is available

if this doesn’t work, then either I generate data,

Visualize the data with PCA or t-SNE:

hi visualize the overlapping samples GE and MB. After this, try to visualize all of the microbial samples.

I’ve visualized all of the microbial samples TCMA for phylum and genus. A lot of overlap between tumor and no tumor.

I’ve visualized also the overlap with Aakash GE data and phylum. Next do it with genus.

Visualize Genus x Aakash GE PCA: done. Not much difference with phylum overlap

what about just GE? Is there less separation than when combining with microbes? Need to get sample metadata to differentiate project. I did this and there is a good amount of separation.

I need to create a file for a data descriptor. This is done and apparently there really are a good amount of normal sapamples in the GE data set. It might be fruitful to find another microbial data set.

I need to create a more flexible pipeline for visualization and everything and handling the data.

Also reorganize files unloading and processing of data and delete files: reorganization all files and processing of data and cleanup is done. I introduced a loader class.

Next step is trying out heat maps with overlapping samples with GE and also saving the plots for all of the overlapping samples. Try with TS NE and umap and see difference between combined and separate

Try subtype detection? Or survival prediction or tumor stage prediction

Think about feature selection.

Fix bug with joining with GE overlap stage. The discrepancy in tumor versus normal is caused by duplicates. Fixed it.

Work package: hypotheses, task, result, discussion

Hypotheses: what were the results of running a prediction model using overlapped microbial and gene expression data on tumor versus normal classification?

Task: run an SVM model to predict tumor versus normal and compare it with separate predictions using only microbial or only gene expression. For all cancers. I ran it with genus and GE and calculated the average support per fold per class and average specificity, F1 score, accuracy and precision per class.

Result: only one result with a decently high support is better than GE only for the overlapped. Everything else is better with the GE. It’s also worse than only genus

Analysis: this might be because of the horrible class imbalance and lack of data for the overlapped

Hypotheses: there might be more separation between classes when overlapping than separately when using t-SNE

Task: run t-SNE on all data sets

Result: pretty similar results to PCA, except this time to overlap with phylum is not the same as the overlap with genus (probably caused by overrepresentation of GE). Also, ESCA seems to have even better separation when overlapping with microbial samples with t-SNE.

Hypotheses: does it matter whether you integrate with phylum or genus?

Hypotheses: could there be benefits to having less features in terms of class separation or prediction tasks

Task: if run feature selection (simple univariate or multivariate) chi² test separately on gene expression and abundance data and overlapped data for multiple amount of features and see which works best for class separation visually assessed with PCA or t-SNE or with prediction tasks such as tumor stage prediction or tumor versus normal prediction

Result: there does seem to be slightly more separation for example for COAD for 10 features when comparing GE and overlapped with genus and definitely more than with genus alone. Although these are separate data points. Especially for COAD 10 features is better separation than 5 for gene expression and provides even more separation then when using all features.

For ESCA and STAD, the integration of genus with GE performs better than just GE. For STAD this is only with feature selection.

Analysis: doesn’t seem to be because the values are wrong in my new pipeline. Not sure why this happens, perhaps it’s just predicting everything is the same for everything and overfitting. The imbalance is also very strong. Have to see what happens if I try with stage data.

Hypotheses: are there benefits to integrating microbial and omics data when predicting tumor stage?

Task: summarize tumor stages into 123 or 4 and use a multinomial logistic regression model to predict them for all cancer types when using genus and when using only gene expression data and when using overlapped data. Do this using all features, five features and 10 features. Don’t have tumor stage data for genus.

Result: all the metrics seem to be higher for the overlapped set, for all of them except HNSC. READ again doesn’t have enough samples to be included in the overlapped set.

Analysis: maybe there is something screwed up with the prediction model, such as always predicting certain stages or something

Work package: visualize results better

Task: Visualize all PCA and t-SNE for stage and tumor, across selected feature amounts and for all cancers. And then do the same when the same samples are selected.

Result: helps with visualization

Hypotheses: results could be more fair if you use the same samples to compare GE genus and overlapped

Task: Add a function to loader which selects the same samples for all three data sets

Hypothesis: does selecting a higher quality of features give better results?

Task: try a grid search on different amounts of features

Hypotheses: visualize prediction performance with graphs

Hypothesis: using a different metric than precision or recall might be more appropriate especially during imbalanced classes.

Task: use PR AUC instead, make avg fun for multiclass stage. Possibly reverse classes because PR AUC is when there are few positive classes. Try to focus on F1 score possibly.

Result: F1 score is clearly higher across a lot of cancers for tumor prediction one five or ten features are selected. It seems to be higher across the board for stage prediction.

Hypotheses: improvement when integrating could be more apparent when visualizing stage separation for PCA and t-SNE of genus, GE, overlap.

Task: do this for these three data sets. I can only do it on overlapped samples, because I don’t have separate stage data for the microbial data.

Result: I don’t really see a pattern.

Hypotheses: proper feature selection without information leakage could give more durable results

Task: create a proper feature selection pipeline where I split the data in testing and training and then spit training again into testing and training so I can select the proper amount of features with a part of the data set possibly using grid search and then do the rest of the downstream analysis

Hypotheses: visualize abundance data for genus

Task: take the mean value for all taxa of abundance values for all rows. Also take out the number of zero values and then plot it.

Result: Of 221 taxa, 102 of them were zero for the genus data set which overlaps with the gene expression data set. Some taxa are clearly also overrepresented.

Hypotheses: reinterpret stage data as a continuous task and normal samples as stage zero

Task: set normal samples as stage zero. Use linear regression to perform prediction. Clamp prediction values and regression round to entire number. Normal samples had tumor stage????

Result: possibly need to compare with SVM

Hypotheses: selecting features with linear regression could be better than chi2

Task: use feature selection version of in your regression and SVM. For regression make sure features are scaled the same way. Make sure there is a good split no data leakage.

Result: Linear regression gives prediction values which were way outside of the range. This results in dubious supports. Due to magnitude of the predictions, the round function to convert from continuous to categorical also gave an error preventing the generation of a classification report. Had to clip and round numbers.

For tumor prediction, regression selection flips GE and integrated performance of STAD&ESCA and revives COAD and HNSC flat lines. When using chi2, STAD has much higher performance with integrated. For stage, some reversal, but not much. COAD has higher integrated performance when feature amount increases. ESCA has higher performance when integrating for both selection types.

Analysis: results were obtained using the entire set for selection, so there was information leakage

Hypotheses: model might be stupid, leading to few variation in predictions

Task: create a pipeline to save prediction outputs and generate confusion matrices to inspect generate the predictions for tumor and stage across all cancers, features, sampling and selection types

Result: the flat lines are because of prediction model collapse, especially for tumors. For stage prediction, there is often some collapse with only predicting stage II, but not always.

Hypotheses: making feature selection part of the prediction pipeline (using a training set) could afford additional power

Task: when having a split, train the feature rankings on the training set, and then use that amount of features on the testing set

Result: the results make more sense due to additional power.

Hypotheses: performance from integrated performance might be due to feature dominance from one modality

Task: keep track of the features selected across different cancers, feature amounts, feature selection types, targets and data sets. Then visualize these using a graph

Result: especially for higher feature amounts, the models are mostly using gene expression features

Analysis: this might simply be because there are way more gene expression features

Hypothesis: increasing the amount of features, using an 80/20 split instead of a 70/30 one, removing cross-fertilization,using more random sampling iterationscould give more useful insights

Task: do these things. Make sure the random sampling split is stratified.

Result: the graphs are showing more consistent and logical results

# Notes

try to better understand the source of your data, especially the microbial data, like what type of sequencing is it and how it dealt with batch effects

see if the differences are statistically significant between prediction results metrics

I need to follow his writing guide and reorganize my paper and then also try to reorganize my results section with the new results

how much effect does the different normalization of microbial, gene expression, and in the future integrated data have on the model?

When integrating I have realized that there are only 353 samples for the stage endpoint, yet the overlapped file still has 401 samples for stage endpoint

I just realize, why am I selecting 200 features in my pipeline for microbial data even though there is only 100 or so nonzero data features?

**Towards January 23 meeting above this**

the expression data set is already feature selected if you think about it, try to do the same with the microbial data set.

Attempt to use survival data

checkout other microbial data sets

**From 31 October 2022, after greenlight meeting**

Possibly optimize prediction set up (NEED TO):

For some reason,the results are a bit different for future values lower than the maximum which indicates that there were some feature selection differences on different runs even though it

Possibly cap the maximum amount of features or display them.

Running tcma\_gen tumor COAD 0 random\_sampling linreg

Running for tcma\_gen COAD

to

Running tcma\_gen\_aak\_ge tumor HNSC 200 random\_sampling chi2

Running for tcma\_gen\_aak\_ge HNSC 200

Maybe replace “0” or “all” feature count with exact amount of features

Maybe do more random selection iterations. Also investigate features being selected. Maybe also more selected features. Experiment with normalization of confusion matrix.

My actual vs predicted is flipped?? I screwed up twice, thereby fixing my screwup. I flipped y\_predicted and y\_test in the list comprehension and thus when saving it, but accidentally swapped the parameters when generating the confusion matrix.

Possibly add stratification to chi2 selection pool to prevent selecting all normal samples.

(I’m missing a normal sample for COAD microbial data)

Add things that cannot change per graph as a title for the whole plot instead of per graph

**From September 28, 2022 pre- greenlight meeting**

maybe regenerate the summary data

Survival?

doesn’t selecting features with all data give me an unfair advantage and have information leakage?

Select features using each cancer separately or all cancers?

Create pipeline to run all tests at once, so run predictions and visualization generations. (you also increase the amount of maximum features selected to 50)

Create more stage descriptor files? Try simple train test split? Get stage for microbe data?

What happens if you do relative abundance for the gene expression as well

try to select the same samples for gene expression and microbiome

Find out how many features are needed. Possibly run experiments with feature selection of different sizes for each of the three modalities and find the numbers which work the best. Then do other experiments.

Are the metrics I am using the best ones? Maybe ROC and MCC

my overlap integration method might have missed some samples because there are samples for TCMA which there are both a and B. Why would it matter? Might just lead to mismatch. But I didn’t know anyways which metadata sample specifically belongs to which gene sample. Oh yeah, it’s because there are samples with only AB and not a at the valve type. When selecting overlapped samples be attention to whether a or B variant was taken in the overlap integration

(write down the steps below as hypotheses, also the things that you are doing like different sampling)

# To Do

-(23 January presentation)

If there is time:

* Change the feature robustness vertical axis to be absolute
* Make it x-axis more spaced
* Make visualization tables of only all selected genus in overlap set
* Fix the fact that the bottom of the pictures in my CNF are not visible
* Make my tables more professional and pretty
* Clean up questions and future plans section

Presentation:

* Don't make the font size of the figures tooself small, and have legible text [x]

Thesis:

* Provide better caption for images (the meaning of axes, what the colors mean, the different modalities) [x]
* Rearrange the introduction: first you give the context, then angles for problem statement, then state-of-the-art (which it seems is my last one right now of previous research), then research question and then solution.
* Make the data title in the result section more clear (e.g. characterization of data) [x]

Further experiments:

* Look into feature robustness
* Look into what (genus) features are being selected. (Look into the literature, does it makes sense)
* Try to project features into latent space (e.g. nmf, encoder)
  + have to make sure my output is between zero and one and the order is the same of the samples to keep that factor constant, contrast one model with everything against separate models per cancer
* Rerun pipeline with nonzero genus features removed

Code:

- Cleanup code: remove all unnecessary data files, cleanup pipeline [x]

Procedural:

- Fill in two documents from BrightSpace

- Apply to RPF

- Request and submit DCO

- Figure out committee for defense

<https://www.tudelft.nl/en/student/eemcs-student-portal/education/graduation-msc/composition-thesis-committee>

Meetings: 26 January, 22 February, 22 March

Thesis committee form: 2 meetings before defense ( 3 months beforehand ). So approved before 22 February. Has to be handed in 5 days prior to the meeting, so before 17 February.

Final examination form: 6 weeks before defense date, last meeting before defense. Approved before 22 March, handing in before 23rd of February.

-(19 December presentation)

Presentation:

- Can take stuff from previous meeting to do

- Make feature selection absolute numbers for fraction of modality [x]

- RBF kernel does not have C parameter [it does!]

Model:

- Display RMSE is instead of F1 score for stage

Thesis:

- Make research proposals separate section and more clear [x]

- Visuals: for exploration remove REA if D. Have a title for each group of observations. Don’t have double information, for example no PCA component labels everywhere. Can just have it for one column and one row. Don’t have the color legend within the graph, either outside or describe it. In caption, describe everything needed to know to understand picture from only caption. Don’t use green and red because of color blindness. [x]

- Be more clear about what you want to show. Don’t show stad, you can remove all the other cancers and say you chose one representative example. Be consistent with it throughout the paper. Include the other ones in the appendix and mention you did so. Explain similarities and differences with the other cancers. It’s

- Method: describe what you have done in the results section, so move what you have there now into the results. Has to be very specific, for example with data. What parameters do you have with your patient samples, what portal did you get it from. Enough information to reproduce.

- Result: describe why she did something, how, what observations you made and what can be concluded. Shouldn’t have sub sub sections, just have it as a label. Group stuff differently. Main conclusions are there is no difference quantitatively or qualitatively. The first group can stay, but the 2nd want you can group in more logical groupings.

-(14 November presentation)

\* presentation:

- for the stage class in balance slides try to expand the table with cancers as columns [x]

- Use bigger text [x]

- Put the main message of what you found instead of the method used in prediction [x]

- Use fewer figures and focus more on what you want to show( for example, for just one representative cancer) [x]

- more descriptive text in figure legends [x]

- how you can improve even more instead of just results are inconclusive

\* Model:

- What happens if you also look at how much each feature there even is softened initial sets and account for it [x]

- Vary feature selection stage

- Try different model (elastic net) for prediction

- Try different model (elastic net) for selection

- different sampling method: sampling with replacement from underrepresented classes

- check feature selection robustness

- rerun linreg after fixing bug with feature selection ordering (and elif for chi2 selection?) [x]

- deal with zero features in genus abundance

\* Thesis:

- Introduction: includes background, knowledge gap and proposal to fill knowledge gap and not separate background section [x]

- Methods and materials: includes detail on how to reproduce paper. Data for example with table of samples and sample types. [x]

- Results: each subsection (4 or 5) will be the main messages with subparagraphs of why you think it is. For example, in the exploration phase it seems that genus and gene expression are overlapped. [x]

- Make clear what experiments you intend to do further to investigate this.

-(17 October presentation)

\* Presentation:

- make clear what microbial features are and how many samples [x]

- what is in the overlap set? [x]

- give updated data on tumor stage distribution

\* prediction set up:

- remove cross-fertilization, reorder amount and ordering of features, use more features and bootstrapping iterations [x]

- try an 8020 or 70 30 split [x]

- for linear regression and ch2, just do splits and use feature selection on training part and testing on testing part [x]

\* evaluation:

- give R bars [x]

- explore features selected and robustness of feature selection ( how do you deal with random seed in different iterations? Is it always to same features? How do you compare when the amount of features of each modality is different? What percentage of features are from each modality? Should I ever show all features?)

Its not fair to compare all features if each modality has different amount []

- don’t have to connecting lines (categorical?) [z]

\* Model

- try hyper parameter tuning, use elastic net instead of linear regression

- look at SVM gamma and c

\* Thesis:

- have an outline and empty tables on what you want filled in

- make hypothesesfrom packages and write in thesis draft based on commits and to do’s and notes.

-(28 September presentation)

\* Show performance with normal samples as stage zero

-- Possibly contrast with performance when that’s not the case (were normal samples being misinterpreted as stage 1?)

\* Show performance when using new prediction set up with random sampling and additional fewith atures

\* Show performance when using new feature ranking with linear regression (discuss data leakage)

\* Show performance when using new prediction with linear regression for stages

\* Show analysis of prediction results across experiments with confusion matrices

\* If time: investigate dominance of each feature in integrated modality

\* If time: update tables

-(6 September presentation)

\* display predictive performance for each cancer separately with increasing feature amounts on the X axis instead of different cancers, and then each omics type separately in each graph [x]

\* try MSE error

\* fix the y-axis to have the same units [x]

\* if time: try a different set up with random sampling to estimate performance

\* start initial subheading and grouping of current thesis document [x]

\* for presentation

- Think about what I really want to present instead of a big bundle of images.

- Fix cancer sucks thing and find source for how many people die from it. [x]

look more into HMP?

Investigate performance more. Are all my features from one modality?

What kind of prediction set up? Random sampling based? And then select a different amount of features and perform sampling to estimate performance

What if I use other omics types and compare?

# Presentation

17 October

Try different stages of integration

- Feature selection

- Prediction task on stage and tumor

two fold cross validation chi squared feature selection\

tumor: svm, rbf kernel, c=1

stage: mlgr,

- PCA and t-SNE visualizations for all cancers and data sets with tumor status data

- also do this for stage data, but only for overlapped samples

- running algorithm on overlapped data sets

- making graphs for visualization of prediction performance

-

- Handle information leakage properly?

- Properly selected feature amounts

- Hyperparameter selection

Thomas lab meeting

# Questions

* Feature selection, different visualizations, t-SNE, overlapped samples, stage prediction, setting baseline, visualizing prediction performance, dealing with information leakage, more information on experimental setup

is the next step good?

Should I still look for more data?

How should I go about doing this?

Sep 28, 2022 Meeting

Information and results do not seem conclusive. Also some contradicting things, such as STAD being better one integrated with chi2, but worse than GE with linreg selection

Dealing with information leakage of linear regression

is this good enough, what is needed for greenlight

Oct 17, 2022 Meeting

Nov 14, 2022 Meeting

Is the separation between materials and methods and results good? I.e. the data split amount and ratio and iterations of random sampling. Subheadings good?

What about design decisions such as using GE data, in results

Can I maybe confided in a more logical way that doesn’t duplicate as many headings and subsections?

Where is the thesis rubric?

Is it okay to cherry pick different cancer visualizations within exploration and prediction performance for example

computing power?