CS230

Final Report CS230-Spring 2020: Modeling Cellular Nutrient Conditions

Jensina Froland
Department of Bioengineering
Stanford University
jfroland@stanford.edu

Abstract

In a world where transcripomic data is increasingly easy to obtain, a model predicting the growth conditions of an organism from their transcriptome would be desirable. Such a model could predict the media conditions for a modified organism to have a certain gene expression profile. Alternatively, this kind of model could indicate the media condition a given organism is sensing, allowing for targeted fermentation condition improvements. Here a three layer neural network is implemented with linear layers ending in a softmax layer to classify a transcriptome by one of four media conditions from 12 classes. Accuracy was highly variable due to a very small train and test set.

1 Introduction

The goal of this project is to predict the nutrient environment and growth rate sensed by yeast based on their gene expression profile. This would be particularly interesting in application to predicting media conditions for a cell to have a certain gene expression profile.

To this end I locally obtained the transcriptomic, metabolic, and proteomic data sets from [1]. All the data for the 47 samples taken is in a form easily loaded into excel and converted to a csv. Only transcriptomic data was loaded as this first model is closely mimicking the supervised model of DeepMetabolism [2]. The labels were constructed to have the first four elements representing if the media represented one of four conditions using a 0 or 1. The fifth element represents growth rate or dilution rate of the culture, a value that was either, 0.07, 0.1, or 0.2. The input layer was the log2 fold change of activity of 9335 Affymetrix probes, representing a transcriptome profile. The metabolic and proteomic data are available for use in the future should they be useful for auto encoding different layers.

The end to end model was locally implemented in tensor flow [3] and has highly variable results in comparison to DeepMetabolism. This is not surprising given only the 48 available samples were used. Ways to expand the data set will be discussed in next steps.

2 Related work

There are multiple approaches to classifying cell states from transcriptomics, but few works focus on using deep learning to predict media conditions from transcriptomic data. More common methods include principal component analysis (PCA) and flux balance analysis(FBA)[4]. Multiple papers have used constraint based modeling including FBA in combination with neural networks [5,6], but not towards predicting media conditions. PCA across transcriptomic data allows for clusters of cell

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

types to be elucidated as well as different cell states[1,5], however overlapping clusters may make this a less than ideal way to classify an unknown transcriptome. Models of how cells interact with their media condition often incorporate FBA, which use genome-scale metabolic network reconstructions to calculate the flow of chemicals through the metabolic network. This has the benefit of indicating which reagents may be limiting growth and a way to predict yield of desirable metabolites or products, but is not organism specific in the measurements taken, making it's application somewhat limited in directing feedback. Notably [7], was able to predict the growth conditions from internal metabolic fluxes using FBA and multinomial classification, in a purely synthetic model of E.coli. This however has limited application in serving as a tool to direct media design as it requires measurements of internal fluxes of chemicals. DeepMetabolism notably has been able to predict cell phenotype with high accuracy from transcriptomic data, serving as a good jumping off point for predicting nutrient conditions.

Here I will be focusing mainly on the approach used by the DeepMetabolism model. While there are other deep learning metabolic models that use omic data, none to my knowledge have been used to predict phenotype or media conditions from transcriptomic data. There are instead more models being used machine learning methods including support vector machines to predict metabolic fluxes through the cell [1]. The DeepMetabolism model was constructed using an unsupervised autoencoder and a supervised 3 layer network that had pruned connections between layers based on biological knowledge of essential genes and the phenotypes they regulate. The final supervised model takes in transcriptomic expression levels reported in normalized log2 fold change values and outputs a given phenotype, specifically tested on ethonal and succinate yield as well as growth rate. This simple model is different from DeepMetabolism in that it is without an autoencoder or biological prior knowledge used to alter connections between network layers.

3 Dataset and Features

Describe your dataset: how many training/validation/test examples do you have?

The data set sources from [1] consists of 4 media classes, each representing samples taken from a culture of yeast grown in carbon, nitrogen, phosphorus, or sulfur limiting conditions. As these cultures were grown in a chemostat, the growth rate was controlled such that they grew at a specific growth rate of 0.07/h, 0.1/h, or 0.2/h. This gives us 12 total classes of samples with 9335 features each, representing the probes used to detect gene expression levels. Within this data set there are 47 samples. Each class has 4 samples, except for the class of sulfer limited media grown at a growth rate of 0.2, which had 3 samples. The mean values and standard deviations the (Figures 1-4). Furthermore the [1] reports how each media class is separable with principal component analysis (PCA) (Figure 5). Figure 5a shows that the carbon limited class is well separated and further does not illustrate great separability between the clusters of nitrogen, phosphorus, or sulfur. However the line in figure 5a illustrates a consistent pattern in all 4 media classes to have growth rate increase along PC1. This figure indicates that there may be difficulty accurately distinguishing nitrogen, sulfur, and phosphorus limited media types if growth rate is not taken into account.

As the standard deviation across a given condition is on the order of magnitude of the mean for many values, this implies that there may be features within the full 9335 set of probes that are more important than others in their expression towards adapting to a given media type. This would imply that not all features might be relevant to predicting media conditions and therefore could be excluded.

Thus the features were filtered by essential genes as was the case with DeepMetabolism. There are 1110 genes essential to the survival and replication of Saccharomyces cerevisiae. Maps to these genes were sourced from the database of essential genes (DEG)[8] and from one set of 870 essential genes from the Saccharomyces Genome Database (SGD)[9] by [10]. Within the DEG essential gene set, all genes essential to eukaryotes are included, giving us 1961 essential genes. Essential gene labels were mapped to probe labels using either SGD IDs and Gene Primary DBID, aquired from the SGD.

Before being input into the model, each feature was normalized by taking the log2 fold change against the reference values provided.

Synthetic data was generated to expand the data set by using a spline function to interpolate between raw transcriptome feature values along specific growth rates between 0.07/h and 0.02/h. However,

given time constraints, the models discussed here were only trained on the 47 real samples described above.

4 Methods

The transcriptomic reference ("REF") sample values within the data set was used as the baseline to calculate the fold change. The formula for this transormation is log2(reference value - sample value). This processed transcriptomic data was then used in the input layer. The network architecture tested were LINEAR -> RELU->LINEAR->RELU->LINEAR->SOFTPLUS and LINEAR –>LINEAR->LINEAR->SOFTPLUS and the same architectures with a softmax layer instead of a softplus layer. The formula for the softmax layer is below and a soft plus layer has the formula f(x) = ln(1+exp x).

$$y_c = \varsigma(\mathbf{z})_c = \frac{e^{z_c}}{\sum_{d=1}^{C} e^{z_d}} \quad \text{for } c = 1 \cdots C$$

DeepMetabolism used the second architecture with a softplus layer, however with layer sizes specific to their organism E coli. The data was intentionally split into test and train sets making sure that a sample from each of the 12 condition types was randomly selected and placed in the test set. The layer sizes tested here were changed to 9335 and then 1961 for that many input features, 1110 for that many essential genes in yeast, and 5 for 5 input features. The end to end model was then implemented using tensor flow [3].

In improving this model I realized that all the weights were nan, due to the labels containing zeros and the cost function sourced from Deep Metabolism dividing by the labels. They fixed this by adding 0.00001 to each label value. Once this was implemented the performance decreased to 0.0 accuracy. Next I changed layer sizes to 9335, 1110, 5 from 1110, 500, 5 which did nothing to improve the scores from 0.0 accuracy. Then I changed the cost function to tensor flow's sofmax cross entropy with logits, which returned higher accuracy ranging from about 0.5 to 0.7. Finally I tuned the learning rate, which increased the accuracy to include 1.0 at times.

A random forest model was created as a machine learning baseline for such a small data set. It was implemented with scikit learn with leave out one cross validation on the 47 samples using raw transcriptomic data as well as the normalized log2 fold change transcriptomic profiles. Three types of random forests with different classes were created. First I labeled the data according to media condition and growth rate combinations, resulting in 12 classes. Then I labeled the data according to just media, 4 classes, and just growth rate, 3 classes. I was able to then get the average accuracy from each approach on this data set.

5 Experimental Results and Discussion

The results of the random forest suggest that training separate networks to predict either media nutrient conditions or growth rate may allow for better results as the accuracy increased for the random forests when just media or growth rate was used to define classes:

LOOCV Accuracy Scores:

12 classes by nutrient and growth rate: 0.19148936170212766

4 classes by media: 0.8723404255319149

3 classes by growth rate: 0.6382978723404256

The DEG essential gene set was chosen as the two sets overlapped by 820 features, making the DEG set more inclusive of important features. Accuracy was used to guide hyper parameter tuning. The final model consisted of three layers with an primary linear layer with a node for each of the features used (1961), a second linear layer with 1110 nodes representing the number of yeast essential genes, and finally a third linear layer with 5 nodes representing the output. No activation function was used in between layers as this decreased accuracy, and the final layer was fed through a softmax layer rather than a soft plus layer as with Deep Metabolism. A mini batch size of 18 was chosen

as there were only 35 training examples. Learning rate was tuned to 0.000001, which yielded a consistent decrease in cost, while learning rates of 0.00001 had exploding costs, and 0.0000001 did not minimize cost as much. Test and Train Accuracy varied depending on the split of the data into test and train sets, while test and train sets had an even representation of all 12 classes. This indicates the need for more samples. Some example outputs include:

Train Accuracy: 0.85714287, Test Accuracy: 1.0, PCC: 0.99999999

Train Accuracy: 0.9142857, Test Accuracy: 0.75, PCC: 0.79056941

The metrics used by DeepMetabolism include Pearson's correlation coefficient (PCC) between the original observations and corresponding model predictions, and the percentage of matched predictions within 20% error of original observations. I chose to compare results based on the PCC as I made a classifier rather than a model that would serve as a quantitative phenotype predictor. The PCC of this model ranges from 0.9999 to 0.5963 so far depending on how the test and train data is split within this smaller data set. This indicates the need for training on a larger data set to attain comparable results to DeepMetabolism, which was able to get a PCC of 1.0 for each of it's three phenotypes.

To understand how this model may be running into difficulty distinguishing between classes, I created a confusion matrix just for one of the test train splits, ignoring growth rate.

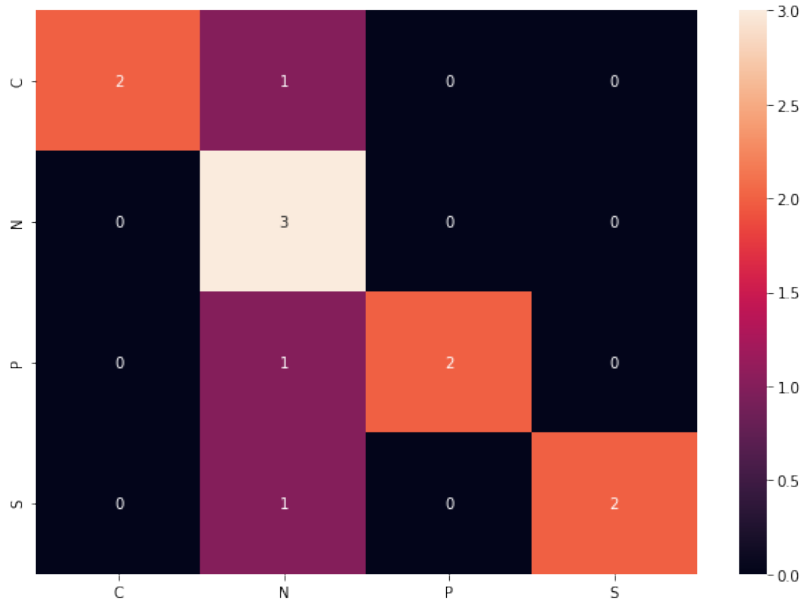


Figure 1: Confusion Matrix: C: Carbon limited, N: Nitrogen limited, P: Phosphorus limited, S: Sulfur limited

A larger test set would be better to determine if the network could accurately distinguish classes, despite the issues of separability noted by the PCA analysis. However this confusion matrix would indicate that the model was not particularly more accurate with carbon limited class over the nitrogen, phosphorus, or sulfur limited classes. Given that the there are 47 samples over all, over fitting is likely. This would be indicated by a lower test accuracy, which seems to depend on the test train split given the variation in test and train accuracy shown above.

6 Conclusion and Future Work

To summarize, performance of the final model was variable, but at times indicated high accuracy and comparable results to DeepMetabolism (PCC 1.00). This indicates a need for a larger training and test set.

The two main goals for improving the model are 1) to train on synthetic samples and source more experimental samples from the same distribution as the data from [1] and 2) to improve the architecture

of the model to be adaptable to predict growth rate, which a softmax layer cannot capture. The first objective I have is to find more transcriptomic samples with similar labels to be able to test for over fitting in future experiments

First I would like to see the impact of changing the labels of the data being fed into the model, potentially predicting media and growth rate separately and adjusting the cost functions accordingly.

Towards generating more training data, I generated more data by using an interpolation method. However, given the relationship between expression at different growth rates is not being accounted for in this interpolation method, this may not be a great way to represent each feature's distribution. This may also lead to over fitting or erroneous results, which could be observed in changes in the test and train accuracy or by using data from another paper that uses nutrient limited media and a similar growth rates.

The second approach I would pursue is to use the distribution of each feature to randomly select error to add to transcriptome values, assuming a Gaussian distribution for all genes. Should these methods prove to have issues over-fitting or poor results, noisy data can be generated by randomly adding error to the 47 transcriptome profiles from the same distribution of the array used to take the transcriptomic measurements, while keeping the labels constant.

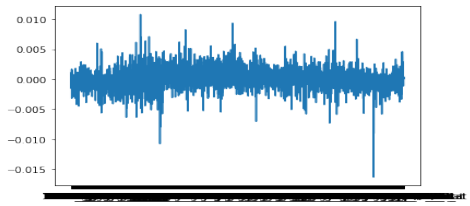
To aid in training a larger data set, another measure to Implementing an auto-encoder to use unsupervised learning to generate the starting weights for the supervised mode's layers. I would need to source and process transcriptomic data obtained from GEO for this specific yeast strain.

Other architectures that more closely model the relationship between transcriptomes and environmental conditions could be implemented. Should these not result in a PCC close to 1 and a high percentage of matched predictions, a biological prior knowledge could be used to limit connectivity between layers, allowing each layer to represent an intermediate protein that regulates the response to a given media condition, which has been characterized for this organism.

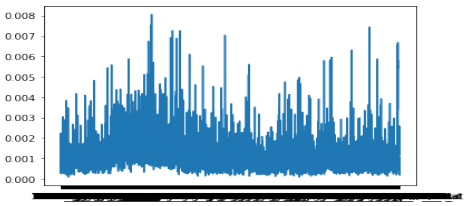
References

[1]Castrillo, J.I., Zeef, L.A., Hoyle, D.C. et al. Growth control of the eukaryote cell: a systems biology study in yeast. J Biol 6, 4 (2007). https://doi.org/10.1186/jbiol54 [2] Guo, W., Xu, Y., |& Feng, X. (2017). DeepMetabolism: a deep learning system to predict phenotype from genome sequencing. arXiv preprint arXiv:1705.03094. [3]Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. [4]Orth, J., Thiele, I. Palsson, B. What is flux balance analysis?. Nat Biotechnol 28, 245–248 (2010). https://doi.org/10.1038/nbt.1614 [5]Zampieri, G., Vijayakumar, S., Yaneske, E., Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. PLoS computational biology, 15(7). [6][Wu, S. G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R., Zhang, X., ... |& Bao, F. S. (2016). Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS computational biology, 12(4). [7] Sridhara, V., Meyer, A. G., Rai, P., Barrick, J. E., Ravikumar, P., Segrè, D., Wilke, C. O. (2014). Predicting growth conditions from internal metabolic fluxes in an in-silico model of E. coli. PloS one, 9(12). [8]Ren Zhang, Hong-Yu Ou and Chun-Ting Zhang, (2004) DEG, a Database of Essential Genes. Nucleic Acids Research 32, D271-D272. [9] S.A. Chervitz, E.T. Hester, C.A. Ball, et al. Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. 27(1):74–78, 1999. [10] N. LeMeur and Z. Jiang, (2020) Synthetic Genetic Interaction in Yeast genes. https://rdrr.io/bioc/SLGI/man/essglist.html

7 Figures

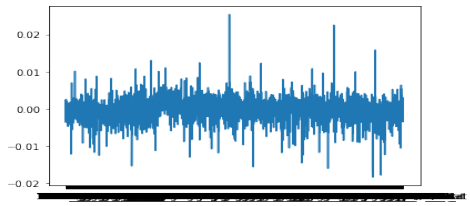


(a) Mean Carbon Limited

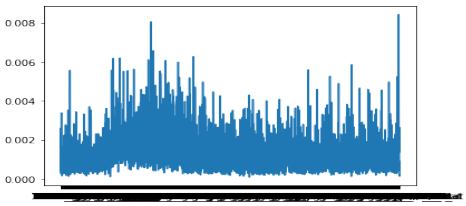


(b) Standard Deviation Carbon Limited

Figure 2: Carbon Limited Transcriptome

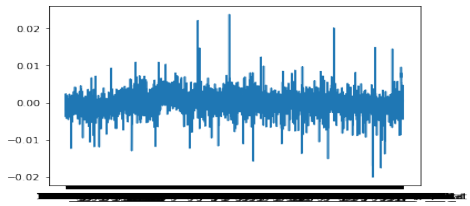


(a) Mean Nitrogen Limited

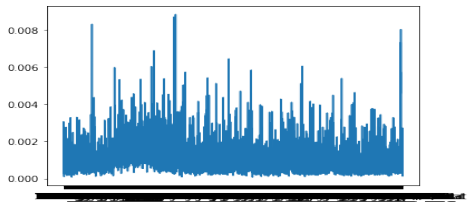


(b) Standard Deviation Nitrogen Limited

Figure 3: Nitrogen Limited Transcriptome

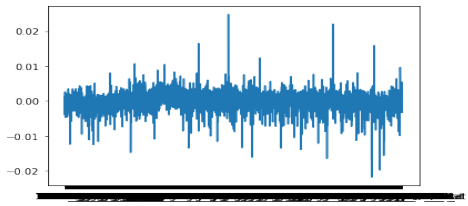


(a) Mean Phosphorus Limited

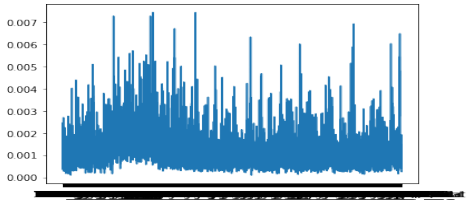


(b) Standard Deviation Phosphorus Limited

Figure 4: Phosphorus Limited Transcriptome



(a) Mean Sulfur Limited



(b) Standard Deviation Sulfur Limited

Figure 5: Sulfur Limited Transcriptome