# Predicting bacterial growth conditions from mRNA and protein abundances.

Mehmet U. Caglar1, 2, 3, Claus O. Wilke3, 4, 5\*

1Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA

2Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas, USA

3Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas, USA

4Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas, USA

5Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas, USA

\*Corresponding author: [wilke@austin.utexas.edu](mailto:wilke@austin.utexas.edu) (COW)

## Abstract

Predicting bacterial phenotype from external perturbations is a problem investigated many times; on the other hand, the question of predicting external conditions by using phenotype is not investigated much. Here we use a *E.coli* transcriptomic and proteomic dataset to predict the growth conditions. Our findings indicate that we can make far from random predictions related with carbon source, salt levels and growth phase of the bacteria from both transcriptome and proteome data. We can also predict multiple conditions for a sample up to 75% accuracy. There is a small synergistic effect related with combining transcriptome and proteome data. It is also possible to make predictions for continuous parameters like the time that the sample was taken and salt concentrations. The analysis indicates 2 clear signals exponential data have a higher predictive value and combining mRNA and protein data results in a small but consistent increase in prediction power independent of the machine learning model used.

## Introduction

## Results

### Data structure and pipeline design

We used data from previous study [cite] to generate predictive models based on mRNA and protein abundances to generate predictive models that try to figure out growth conditions that the sample is collected. Although the methods and procedures is general and can be applied to different growth conditions here we focus on prediction of four different parameters that are systematically varies in the data; growth phase, carbon source, Mg and Na concentrations. Data set composes from 155 samples, mRNA abundances were measured for 152 of them and protein abundances were measured for 105 of them. For 102 of the samples we have both mRNA and protein concentrations. [figure 1]

For the analyze of data we generated a pipeline using four different machine learning models including SVM with radial kernel, SVM with sigmoidal kernel, SVM with linear kernel and random forest algorithm with the help of *e1071* [cite] *and random forest* [cite] *packages*. We use *C-Classification* for training classification model and *eps-regression* for training regression models which we use for predicting growth rates. We adjusted weights of samples in a way that each class ends up with equal weight in order to prevent the prediction bias in favor of more populated classes.

We also generate a tuning loop for the analyze in which we were optimizing cost value for models SVM with linear, radial and sigmoidal kernels, and gamma for SVM with radial and sigmoidal kernels. For random forest algorithm, we optimize mtry, ntrees, and nodesize parameters. We use the multi conditional f1 score [cite], in order to weight all conditions equally and assign same importance to false positives and false negatives.

Before SVM we apply DeSeq2[cite] for size factor normalization, fSVA [cite] to normalize batch effects and PCA [cite] to obtain the principal components of the data. Despite the fact that our dataset has strong batch effects [cite], we believe we get rid of the batch effects as much as possible before training our data with the help of fSVA algorithm. We write our tuning algorithm that enables us to divide the dataset into subsets semi-randomly in a way that the ratios of samples tried to be constant as much as possible between training tuning and test sets. We also calculate the conditional class weights for each training data in tuning process [figure 2].

So overall our pipeline is designed for tuning four different models, SVM with linear kernel, SVM with radial kernel, SVM with sigmoidal kernel, and random forest; independent of weight and batch effects and by using principal components in order to prevent overfitting and assigning same importance to false positives, false negatives through all individual conditions.

### We can make predictions on both datasets with all four models

Predicting 4 different components carbon source, Mg levels, Na levels and growth phase which makes 16 distinct conditions at the same time on the test dataset is a challenging task. We apply our pipeline and the results indicate we can make reasonable predictions by using all our algorithms there are clear winners in the tuning stage. The table1 shows the winning models for mRNA and protein data in the tuning stage, as can be seen SVM with radial kernel is the winner in mRNA data, and SVM with sigmoidal kernel is the winner in protein data. Although the test set scores are less significant and more similar to each compared to tuning set results, the same trends can be observed in test set records (Figure2). The dramatic drop between tuning set results and test set results is due to xxx. Although the results on test set is much less significant compared to tuning set the results for test set are still far from random, enable us to predict the 4 components of the sample conditions correctly 62%of the time for mRNA and 54% of the time for protein data if the number of test set examples are equally distributed and independent from the actual sample number distribution that the raining set use corresponding to *multi condition F1 score* of 0.63 and 0.56 respectively. (Figure 3)

### Combining mRNA and protein data causes a significant increase in predictability

We can use the pipeline to gather some information about internal workings of biological system. Figuring out the amount of distinct information between mRNA and protein data is an important task and can show how much of the information in mRNA level is lost in protein level and how much of new information generated by post transcriptional regulations. To see this, we run our pipeline on subsets of data that matches in between proteins and mRNAs. The results indicate protein data includes more information compared to mRNA data in three of four models used, and combined mRNA protein data includes more information compared to both mRNA and proteins with an exception of for radial model f1 score distribution associated with combined data is not different from protein data (Table 2, Figure 5). By using this information, we can say that after normalization for number of samples protein data contains more information about external conditions the bacteria lived in than mRNA data, in addition to that combined mRNA protein data contains more information that individual samples which indicates some information about external conditions in mRNA data was lost during translation process.

### Information about external conditions lost as time passes

To see how the information content changes in time we used the sub-data sets that only have either exponential or stationary phase samples for both protein and mRNA data. Results indicate the prediction power of data for the external conditions decreases from exponential to stationary phase for both protein and mRNA data. This trend is independent from the model used. We can say that the bacteria begin to have similar composition in terms of protein or mRNA abundances as time passes regardless of external conditions (Figure 6).

### Most predictable external condition is growth phase for mRNA data and carbon source for protein data

The results indicate that the performances of models for specific data types and experiments are similar to each other and chosen model is not the main driver of successful prediction on the other hand there is a big variation between conditions. The most predictable external condition is the growth phase from mRNA data and the least predictable one is carbon source. On the other hand, the most predictable external condition is carbon source for protein data and least predictable one is the Mg+2 concentration (Figure 7 and supplementary figure xx). Comparing the results of mRNA and protein data does not make sense because of different sample sizes. These results are consistent with the “Clustering of mRNA and protein abundances by different growth conditions” table generated for the same dataset (cite)

## Discussion

We try to find methods to classify growth conditions of bacteria by analyzing its mRNA and protein concentrations, this is the inverse problem of finding concentrations of specific mRNA’s and protein under given growth conditions. Overall results indicate that we can predict the growth conditions up to 90% percent of the time. We used the dataset [cite], which includes corresponding mRNA and protein reads for analyze.

Overall there are two trends after get rid of the bias because of different number of samples in proteins and mRNA we can say proteins have more prediction power than mRNA’s and predictability decreases as phase changes from exponential to stationary

The biggest handicap of the work is the sample size, although the sample size is big compared to similar studies [cite], the comparison between multi variable and multi variable intersection analyses for both mRNA and proteins the prediction power decreases with larger sample set. This indicates we are not in stationary regime in terms of number of samples.

The second problem seems to be associated with sample number bias, although we made a correction with weight factors and it seems still there is a correlation between sample size and prediction performance.

## Materials and Methods

### Data

We use the same data sets that were used by the paper [cite]. For all single variable test, and also for multi-variable prediction tests except for combining mRNA and proteins we use all available data. For combination tests we use intersection of mRNA and protein samples (102 sample).

### Prediction Methodology and Parameters

The initial preparation of the data is similar to the paper [cite]. After finding suitable subsets of the data for the tests and summing up technical replicate results for proteins, we calculate size factor normalized data with DeSeq2 [cite] and apply variance stabilizing transformation (vst) on it.

We then divide the data into two subsets; training and test. The division is semi random i.e. algorithm preserves the ratios of different conditions in training and test subsets. We preserve the condition labels for training data but we delete the labels of the samples for test set. We then apply frozen Surrogate Variable Analysis (fSVA) [cite] to get rid of the batch effects in the sample. The algorithm can correct the batch effects on both training and test data without knowing the labels of the test data. We defined individual conditions by using labels of different Mg, Na concentrations, different growth phases and different carbon sources to define individual conditions for fSVA algorithm. After fSVA we use principal component analysis (PCA) to define principle axis of training set and rotate the test set with respect to principal axis of training set. We then pick the most significant top “*d*” axis, where “*d*” is square-root of number of samples in training set. Then we calculate weights for each different condition in training set. Weights are inverse of number of samples for each specific condition in training set. Finally, we apply support vector machine (SVM) algorithm from e1071 package [cite] with c-classification and radial kernel, with chosen parameters of cost “c” as 1 and gamma “γ” as 1/d to predict the labels of test sets. [figure 2]

For combined data we calculate size factors and batch effects individually for mRNA and protein data then combine 2 datasets and apply PCA on it

We repeat the pipeline for thousand times with different semi randomly chosen training and test sets. We calculate the percentages of predictions for each distinct condition and report them as tables.

### Calculation of the score metric

The metric we use is multi condition F1 score [cite] that normalized over individual condition; i.e. each condition have equal weight instead of each sample.

### Statistical analysis and data availability

## References

## Acknowledgments

This study was funded by Army Research Office (ARO,<http://www.arl.army.mil/>) grant W911NF-12-1-0390 to COW. COW also acknowledges support from the NIH (R01 GM088344, R01 AI120560) and the NSF (Cooperative agreement no. DBI-0939454, BEACON Center). The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided high-performance computing resources.

## Contributions

M.U.C, C.O.W. conceived the study and designed the pipeline and analyze the data.

contributed computer code used for data analysis.

M.U.C, C.O.V. prepared the figures.

M.U.C., C.O.W. wrote the initial paper draft. All authors reviewed and edited the final manuscript

## Competing interests

The authors declare no competing financial interests.