# 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.

## 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 SVM with radial kernel from *e1071 package* [cite]. We use *C-Classification* for training classification model and *eps-regression* for training regression models. 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 regression problems in order to tune c and gamma associated to SVM algorithm. We do not apply tuning algorithm for classification problems because in some cases the number of samples in subclasses is not enough to divide the data into 3 groups as “training-set”, ’tuning-set” and “test-set”. Before SVM we apply fSVA [cite] to normalize batch effects and PCA [cite] to obtain the principal components of the data. [figure 2, supplementary figure 1]

### We can distinguish growth phase, carbon sources and Mg concentrations but not for Na concentrations from mRNA data

Results indicate data has a strong signal associated with growth phase with a correct prediction rate of 94% for exponential, 92% for stationary and 69% for late stationary phase. For carbon sources data let us to do moderate predictions. With a 94% correct prediction rate glucose is the best predicted carbon source. Glycerol and lactate follows it with 77% and 66% correct prediction rates. 64% of the runs gluconate is predicted as glucose and only for 24% of the cases gluconate is correctly predicted. There is also a strong signal associated with Mg levels we can predict low magnesium 81%, base magnesium and high magnesium 71% of the time correctly. On the other hand, the signal associated with sodium levels is weak and our pipeline is not able to distinguish high Na data from base Na. Pipeline predicts base Na correctly %97 percent of the time and predicts high Na correctly only 56% of the time. [figure 3]

### Overall performance of using protein data in prediction is less than mRNA data with an exception of carbon sources

Due to the results protein concentration is a slightly weaker predictor compared to mRNA concentration with the exception of carbon sources. For different growth phases test results provide a correct prediction rate of 82% for exponential, 77% for stationary and 49% for late stationary phase. Most clean signal is related with carbon sources in protein data. Glucose, glycerol and lactate are correctly predicted 93% 92% and 100% percent of tests. The worst prediction results are associated with gluconate. Gluconate is predicted correctly 66% of the cases. When it is not predicted correctly it is mislabeled as either glucose or glycerol with 20% and 14% of the cases. Mg levels are one another variable that lost predictability for protein data compared to mRNA data. Algorithm mixes high and low Mg levels 30% percent of the cases. On the other hand, base Mg, which is predicted correctly 93% of the cases, is distinct from other two. As for mRNA concentrations protein concentrations cannot be used to distinguish base sodium samples from high sodium samples. Base sodium is predicted correctly 98% of the cases and high sodium is predicted correctly for only 48 % of the cases. [figure 4]

**Signals are getting weaker as samples translate from exponential to stationary state.**

We also analyze the data by dividing the data into two time frames exponential and stationary; the hope in here is to investigate the distinguishability of conditions through time. Overall trends indicate that the cells become more similar as time passes for both mRNA and proteins. Also in general protein concentrations look more similar to each other compared to mRNA concentrations. [figure 5]. The individual trends mRNA lactate samples become more similar to glucose as growth phase changes from exponential to stationary, similarly low Mg protein samples begin to look more like base Mg. As an exception predictability of Na concentrations based on mRNA levels increases from exponential to stationary phase. Supplementary figures [1-3]

**Predicting multiple conditions from a single model is also possible.**

We can also make predictions for all four variables for the same data which generate a prediction space composed of 16 distinct conditions. The pipeline is similar for individual parameters and includes weight factor normalization for different sample sizes in training data also cost and gamma values are assigned in the same way with individual predictions in previous section. The results indicate with 152 mRNA samples we have a prediction score of 0.58 and with 105 protein samples we have a prediction score of 0.51 [figures 6, 7]

### Combining models by combining datasets increases the prediction ability slightly.

Finding the complementary information between mRNA and protein concentrations is an important aspect of the project. We directly combine mRNA and protein data after calculating size factors and batch effects independently for both mRNA and proteins and provide combined data to SVM. To compare the gains compared to individual data we calculate prediction tables for intersection set mRNA and protein data [supplementary figure 4 - 5]. The results indicate for 102 intersecting samples proteins are better for predicting combined conditions compared to mRNA with prediction score of 0.43 and 0.49 respectively. The prediction power increases slightly compared to mRNA when we add two datasets together with a prediction score 0.52 [figure 8]. But the drops because of the decrease in data size is more dominant than the gain of combining mRNA and protein data together, which indicated the number of samples is the main limiting factor for making predictions [supplementary figure 6].

**We can make continuous predictions by using regression**

## 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 mRNA is a better predictor than proteins (which might be because of more samples associated with mRNA) 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 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 prediction results need normalization in order to be comparable between different conditions. The metric uses trace to find correct prediction percentages for each distinct category and compare the result with random and perfect predictions by scaling random as “0” and perfect prediction as “1”.

### 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.