Computational Identification and Annotation of Key-reactions in Metabolic Pathways of E. coli that Discriminate Different Growth Conditions.

Viswanadham Sridhara[1], Austin G. Meyer[1,2,5], Jeffrey E. Barrick[1,2], Pradeep Ravikumar[3], Daniel Segre[4], Claus O. Wilke[1,5,*]

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

2 Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, TX, USA

3 Department of Computer Science, The University of Texas at Austin, Austin, TX, USA

4 Department of Biology, Boston University, Boston, MA, USA

5 Section of Integrative Biology, The University of Texas at Austin, Austin, TX, USA

$*$ **E-mail: wilke@austin.utexas.edu**

Abstract

Currently, predicting bacterial growth conditions without prior knowledge of the medium is an unresolved problem. By contrast, computing bacterial metabolic output, given a set of starting conditions has become a comparatively routine task via flux balance analysis (FBA). To compute metabolic output, one specifies a set of starting conditions within the context of a complete bacterial metabolic model. Here, we selected 7 carbon and 7 nitrogen sources along with 4 more commonly used experimental media. We generated metabolic flux data using FBA in E. coli MG1655 for the 18 specified growth conditions. We then used a model selection algorithm to identify the key reactions that discriminate among the tested growth conditions. These models on average require assaying the fluxes through 9 reactions to accurately predict the correct carbon and nitrogen source used during growth. For each source metabolite, we mapped its predictive reactions onto the E. coli central carbon metabolism to highlight important metabolic regions. Our analysis provides several important physiological and statistical insights. First, by analyzing metabolic end products, we can consistently predict growth conditions. Second, despite its heterogeneity, the experimental media appears to be similarly predictable to the homogeneous media. Third, predictive reactions seem to frequently lie near the initial entry point into central metabolism for the metabolite being predicted. Finally, we found that separately predicting the carbon and nitrogen sources is better than making joint predictions. In addition, the fact that separate prediction performs better than a more sophisticated joint prediction scheme, generates several potentially interesting hypotheses regarding bacterial physiology.

Author Summary

Introduction

Flux balance analysis (FBA) is a computational technique that is routinely used for computational guidance in metabolic engineering [1]. Traditionally, FBA involves training a whole-cell metabolic model with a specified starting media and optimizing the network to produce a specific output. The goal of such analyses is to identify bottlenecks to producing various metabolic products.

Results

Reversing flux balance in the E. coli system

Improving growth rate or identifying the key enzymes for improving specific metabolite concentrations is a routine task in metabolic engineering. In general, FBA is a computational to guide experimentalists in hypothesis generation. Specifically, it can provide detailed information about potentially product-limiting reactions within a metabolic network. The insight gained from FBA can be used to remove such road blocks to metabolic output. Rather than making forward predictions about network throughput from known starting materials, we wanted to know if it is possible to reverse FBA to generate predictions about starting conditions from known output fluxes. To that end, we used machine learning techniques in addition to traditional FBA to predict growth sources from simulated flux data.

We used the iAF1260 metabolic model of K-12 MG1655 strain [25]. The complete reaction network includes transport, biochemical, and biomass exchanges containing 2382 separate reactions. Each reaction has a minimal and specific set of reactants and a similarly minimal set of products. There is no inherent promiscuity in the biochemical

pathways. Since the model is assumes perfect fidelity, each transport protein is responsible for one and only one substrate. Thus, predicting the starting material based on the unique transport protein would be a trivial result. As our intent was to predict start conditions from the simulated flux data, we removed all the transport reactions. Other details about this model are provided in methods section.

Effect of noise and prediction structure on model accuracy

We tested 7 carbon and 7 nitrogen sources in all pairwise combinations. Each flux balance simulation requires that we specify both a carbon and a nitrogen source. In addition, we generated 100 replicates of pairwise combinations to make 4900 total observations. In addition to simulating individual carbon and nitrogen sources, we generated flux data and calculated misclassification rates for other commonly used media described in (Table 2). To simulate the effect of measurement error, we added randomly selected carbon and/or nitrogen sources [6]. Keeping the metabolic uptake rate at 1% (0.2 mmol/gDWhr) of the primary sources, we changed the number of nutritional sources that we used as noise. For reference, 2% noise means that we randomly picked 2 carbon and 2 nitrogen sources from a set of 174 carbon and 78 nitrogen sources used previously [6]. Next, we generated a training set by randomly drawing from our simulated fluxes. We trained the models with a generalized linear model to predict starting conditions based on final fluxes. The remaining simulation data was used to calculate the misclassification rates. We used cross-validation in a generalized linear model to pick regression coefficients.

We find that even at relatively high noise levels of 10%, we are able to predict the growth media in greater than 80% of cases. Figure 2 illustrates that as random background carbon and nitrogen sources are added, the misclassification rate increases relatively little. Moreover, assuming a training set of sufficient size, it appears additional data does little to affect the accuracy of model predictions. Despite our simulations being run as pairwise combinations of the available carbon and nitrogen sources, there appears to be no benefit

to joint prediction. In fact, a model involving the separate prediction of the carbon and the nitrogen sources performed better in all cases tested here (Figure 2 and Figure 3). At 1% noise, the difference between joint and separate predictions is lowest; by contrast, at 10% noise the difference climbs to greater than 10% difference (Figure 3). At 10% noise level with joint prediction, the misclassification rate is 25%. On the other hand, given 7 carbon and 7 nitrogen sources, just by chance the misclassification rate is 93%. Moreover, the relative value of more data for separate prediction is substantially lower. It seems at this point, any estimates made via reverse flux balance analysis ought to involve separate rather than joint carbon source prediction.

In order to generalize our simulations to more experimentally relevant test condition, we performed similar analysis for several media that are more commonly used in experimental microbiology. Specifically, we tested autoinducer bioassay (AB) minimal media, proprietary media from the company ATCC, Davis Mingioli (DM) media and Bochneder defined minimal media. Due to the relatively small number of available starting conditions in the training set, we were able to classify our experimentally relevant media choices very accurately. Our misclassification rate was less than 1% for noise levels up to 20%. (What does this sentence mean???) There was only one feature predicted for each of 3 reactions, except for AB minimal media, for which only β_0 is non-zero and the rest of regression coefficients are 0.

Application to experimental design

To gain some mechanistic insight into the isolated reactions, we mapped these reactions onto the E. coli central metabolism model. The predictive reactions were overlayed on E. coli MG1655 central metabolism. Manual validation of these reactions indicate that most reactions seem specific to some unique growth source. For example, using acetate as a growth unsurprisingly isolates TCA cycle entry as the relevant predictive reactions. Similarly, sorbitol is the singly reduced alcohol of D-glucose and fructose enters glycolysis

just three steps away from un-phosphorylated D-glucose; thus, each possesses a predictive reaction in the relative vicinity of the glycolytic pathway (Figure 5). Mapping nitrogen sources to central metabolism reveals a similar trend. For example, L-alanine as a growth source has predictive reactions near its site of entry into the three and four carbon metabolism of the TCA cycle (Figure 6). Each of the metabolic maps (Figure 5 and Figure 6) is meant to highlight generally important areas in the E. coli central metabolism.

Despite the accuracy of our scheme, a model with a large number of predictive reactions would quickly become intractable in experimental applications; each predictive reaction means an additional unique biochemical assay. Our model required an average of 9 reactions for accurate prediction. Considering our network included a total of 1443 reactions, 9 is an impressively minimal requirement for media prediction. To understand the effect of reducing the model parameters below the optimal number, we employed a different strategy from that used above. There were 126 reactions that were assigned a non-zero coefficient in our trained model. For each of those 126 reactions, we eliminated it, trained a new model, and calculated the test accuracy. Except for 3 reactions, we found that the misclassification rate remained unchanged for each of the other 123 reactions. Thus, the model average of 9 predictive reactions does not represent a unique solution; for experimental purposes, the number of assayed reactions could very likely be further reduced without sacrificing much predictive power.

Discussion

We have developed a method for making predictions regarding growth media from known metabolic flux data. We generated flux by simulating the complete E. coli metabolic network for 7 carbon and 7 nitrogen sources. Then, we divided the data and employed machine learning with a generalized linear framework to train a model to predict growth media. We found that even at high noise levels, we could make reliable predictions regarding growth media for all tested sources. Also, extending our prediction algorithm

to more experimentally relevant growth media, showed comparably accuracy to the extremely minimal simulated media types. Moreover, in our scheme, independent prediction of carbon and nitrogen sources always performed better than joint prediction. Although this is likely influenced by the volume of training data, it very likely says something important about rate-limiting reactions in the E. coli metabolic network. In addition, our results indicate that for most input metabolites at least one predictive reaction commonly occurs near its entry point to central metabolism. Finally, we found that the number of reaction fluxes required to make accurate predictions is relatively small, and can probably be reduced further with few trade-offs. Thus, predicting growth media from metabolic flux data should be an experimentally tractable problem.

We have shown that given simulated metabolic flux data, growth conditions can be accurately predicted via machine learning. Although the fractional noise can have a dramatic affect on model accuracy, the misclassification rate remains acceptably low even with 10% or 20% noise. The addition of noise revealed one interesting and unexpected physiological hypothesis about E. coli metabolism. Namely, as noise increases from 1% to 20%, our model increasingly predicts acetate as the default carbon source and ammonia as the default nitrogen source. Generally with any input growth source, the reactions in the pathway that lead to TCA should have some reasonable amount of flux calculated by FBA, if the key is to produce biomass. In order words, these sources may not be so surprising when one considers acetate's central role in the TCA cycle; what is essentially the center of bacterial metabolism. Further ammonia is one of the few, if not the only, source of nitrogen without any associated carbon atoms. To be sure this was not an artifact of nutrient limitation (carbon versus nitrogen), we changed the uptake rates of carbon and nitrogen sources so that there is no limiting factor and re-analyzed by training a new model. There seems to be no effect of these uptake rates.

It was surprising to us that given the same number of observations in the training set, separate prediction for starting materials always performed better than joint prediction.

There are two likely explanations for this result. First, making joint predictions requires discriminating between 49 different pairwise combinations. By contrast making individual prediction only requires discriminating 7 different conditions in two different sets. Thus, one possible explanation for the lack of predictive power is that we simply did not have the appropriate level of training data. Indeed adjusting the amount of training data appears to have a dramatic effect on joint prediction in particular. On the other hand, such an issue represents an important experimental concern. Namely, often the size of the training set, being experimentally determined, is more limiting than the size of the testing set. As a result, our analysis indicates employing a separate prediction strategy will be more useful for experimental application. Second, although it is not yet clear to us, separate prediction may gain additional power due to the physiology of the organism. For example, if the initial steps of metabolic entry were often predictive (as they appear to be), then the two predictions could easily be performed separately. By contrast, if one were using a single metabolite as a combined carbon and nitrogen source, we may expect different results.

Finally, we have shown that there is no obvious experimental restriction for applying FBA and machine learning to predict initial growth media from final metabolic flux data. As nine reaction fluxes on average provided the optimal solution to our regression model, it is evidently not a unique solution. There are very likely other possible alternative solution that may garner similar predictive power. By individually eliminating reactions and retraining the model, it appears the minimum number of critically important reactions is three. With such a small number of necessary reactions, it should be possible to immediately apply this technique to experimental data.

Perhaps a paragraph on caveats... not really sure.

Materials and Methods

We used MATLAB and R for this study. For flux balance analysis, we used COBRA toolbox [29] with MATLAB and for multinomial regression, we used GLMNET package [24] with R. The methods are described in detail below.

Flux Balance Analysis:

In FBA, the steady-state solution for reaction fluxes is calculated. S(m,n) is the stoichiometric matrix for "m" metabolites and "n" reactions and is represented as S hereafter. The other variables used are v and x, where v is a vector of reaction fluxes for all the reactions involved and x is the concentration of the metabolites. In steady-state, the rate of change of metabolite concentrations is 0. So, we can formulate the above problem with the set of equations, as described below:

$$dx/dt = Sv_i, \tag{1}$$

$$Sv_i = 0. \tag{2}$$

The constraints that are typically used are:

$$\alpha_i < v_i < \beta_i \tag{3}$$

where α and β are lower and upper bounds of the reaction fluxes.

Hence, we solve this set of linear equations with interested constraints by linear programming.

$$max.c^T v, s.tSv_i = 0. \tag{4}$$

where $c^T v$ represents the biomass composition reaction.

Generally in these metabolic networks, the number of reactions are more than number of metabolites resulting in multiple solutions. However in metabolic engineering applications, we are interested in optimizaing a certain function, for example here, we are

interested in maximizing biomass composition which would result in a particular solution.

Our motivation to introduce noise is 2-fold. First, we would like to find alternate optimal solutions using FBA methods so that the final prediction results still hold for varying noise levels. The second reason is to simulate replicates for a particular growth condition to train the mathematical models. Flux variability analysis can also be used for finding alternate optimal solutions, but apart from generating alternate optimal solutions and generating replicates, we can also introduce noise and find its effect in key-reaction prediction using the method described above.

E. coli model:

From the BiGG database, we downloaded the iAF1260 model, as it has shown to be used rigorously in various studies involving metabolic engineering and seems to agree well with the experimental data. Generally these models are stored in SBML format, which is becoming a common format for systems biology related models i.e., signaling pathways, gene regulatory networks, metabolic pathways etc [30]. In the current iAF1260 model, there are 2382 reactions, 1668 metabolites. The biomass composition reaction is also included in the model. Except for the input growth sources (Carbon and nitrogen sources used in this study) used, we did not change any defaults that are used in this model. The upper bounds on 2377 reactions is set to 1000 mmol/gDWhr, i.e., there is no limit on the production of metabolites involved in these reactions. But for 5 reactions, i.e., ATPM, CAT, FHL, SPODM, SPODMpp, the upper bound was set to 50 mmol/gDWhr. On the other hand the lower bound for almost 1800 reactions is set to 0 mmol/gDWhr, which means these reactions cannot uptake any metabolites from the media. For ATPM reaction, the lower bound is set to be the same as upper bound at 8.39 mmol/gDWhr. For the rest, the lower bounds were set to -1000 mmol/gDWhr, except for glucose and oxygen. We did not change the oxygen uptake rate (-18.5mmol/gDWhr), but we set the

lower bounds of glucose and ammonia to zero. If we used glucose/ammonia as growth sources, we then set the lower bounds of these sources accordingly.

Growth conditions:

We picked the growth conditions manually from [6] that seemed more common in experiments. In our study, we used pairwise combinations of 7 carbon sources, and 7 nitrogen sources. 7 carbon sources when used alone did not result in any growth. On the other hand, the nitrogen sources except ammonia contributed to E. coli biomass composition that is non-zero. These carbon and nitrogen sources are listed in Table I. Depending on the input growth, we set the lower bound of that particular exchange reaction to -20 mmol/gDWhr. This lower bound of -20 mmol/gDWhr is previously used as reasonable uptake amount in many studies [6]. For 49 pair-wise combinations of the sources, we generated 100 replicates of data. Apart from these growth conditions, we also used 4 growth media, generally used in E. coli K-12 MG1655 experiments as cited in EcoCyc database [Cite URL of EcoCyc].

We changed the uptake rate of nitrogen source to -1000 mmol/gDW/hr, keeping the carbon source at -20mmol/gDW/hr to make sure carbon sources don't lack nitrogen source. We also repeated the analysis keeping the carbon source uptake at -1000 mmol/gDW/hr and nitrogen source uptake at -20mmol/g/DW/hr.

Background noise levels:

To generate replicates and may be alternate optimal solutions using these conditions, we incorporated different background noise levels. For this, we used a subset of the 174 carbon and 78 nitrogen sources, previously used in Feist et. al [6]. We used different background noise levels, ranging from 1% to 20%. For example, if we want to set 5% noise level, we randomly picked 5 Carbon and 5 nitrogen sources and set their lower bounds to -0.2 mmol/gDW. Please note that we generated the flux data for a pairwise combination of 1

carbon and 1 nitrogen source along with the background noise as described here.

For each noise level, we used half of the dataset as test set. We used subsets of the remaining half as training (i.e, 240,480,2400 observations). On the training sets we did 3-fold cross validation. We used cross-validation in GLMNET package for model selection. Model selection means picking the regression coefficients at the lambda value that had the lowest misclassification rate with 3-fold cross-validation. We then used this model to calculate the misclassification rate on the test set. We repeated this step to calculate the misclassification rates at different noise levels (1%,5%,10%) and different training data sizes (240,480,2400 observations).

Regression based on regularization:

We used GLMNET package with R. We used 3 fold cross-validation. From the flux balance analysis, the observations we generated for different noise levels were divided into 2- halves, one for training and the other for test set. We did training and 3-fold cross validation to pick the lambda that has the lowest misclassification rate using the GLMNET. We did this for both joint prediction as well as the separate prediction and then making a joint call.

1. Separately:
   i) Take data set
   ii) Train model to predict C sources
   iii) Train model to predict N sources
   iv) Predict C and N separately and calculate joint prediction accuracy.

2. Jointly
   i) Take data set
   ii) Train model to predict C and N jointly
   iii) Predict C and N jointly and calculate prediction accuracy.

Since the cross-validation accuracy cannot be used to compare the GLMNET results for separately predicting to that of joint prediction, we used the test set to determine the prediction accuracy.

Below are the equations used in multinomial regression setting. If Y is a categorical response variable with "m" levels (m¿2), and x is a predictor, this will result in

$$logP_r(Y = l/x)/P_r(Y = m/x) = \beta_0 l + x^T \beta_l, l = 1, 2, ..., m - 1. \quad (5)$$

From this, we can model

$$P_r(Y = l/x) = e^(\beta_0 l + x^T \beta_l) / \sum_{k=1}^{m} e^(\beta_0 l + x^T \beta_l). \quad (6)$$

The above model can be fit by maximizing the penalized log-likelihood, as explained in detail elsewhere [27]

$$max 1/N \sum_{i=1}^{N} logp_g i(x_i) - \lambda \sum_{l=1}^{m} P_\alpha(\beta_l) \quad (7)$$

Here β are the regression coefficients and N is the total number of observations. However, in LASSO, there is a tuning constant λ that puts the strength on the penalty introduced to achieve sparsity. We direct the reader to Friedman et. al., [27] for further details.

We used GLMNET package with R, instead of MATLAB as the supercomputing cluster at TACC has R installed on it and we were able to easily add the GLMNET package to it.

Mechanistic Insights:

Mechanistic insights of the results obtained from multinomial regression can be obtained by understanding the role of features that are predicted for each growth condition. For this, we used E. coli map downloaded from BiGG database. For overlaying reactions (features) onto E. coli central metabolism map, we used modules in COBRA toolbox.

We deleted 4 reactions in the map that seemed inconsistent with the E. coli model used. The predictors from GLMNET are highlighted with a different color, along with the metabolites involved in these reactions.

Acknowledgments

We would like to thank Segre lab members at Boston University for useful discussions on flux balance analysis. We would also like to thank BCG and TACC at UT for resources. VS would like to thank Piyush Rai for useful discussions on multinomial classification.

Author Contributions

Conceived and designed the experiments: V.S. and C.O.W. Performed the experiments: V.S. Analyzed the data: V.S and C.O.W. Wrote the paper: V.S, J.E.B, P.R, D.S. and C.O.W.

References

1. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245-8.

2. Edwards JS, Palsson BO (2000) The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97: 5528-33.

3. Karp PD, Riley M, Paley SM, Pelligrini-Toole A (1996) Ecocyc: an encyclopedia of escherichia coli genes and metabolism. Nucleic Acids Res 24: 32-9.

4. Ouzounis CA, Karp PD (2000) Global properties of the metabolic map of escherichia coli. Genome Res 10: 568-76.

5. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). Genome Biol 4: R54.

6. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. Mol Syst Biol 3: 121.

7. Orth JD, Palsson B (2012) Gap-filling analysis of the ijo1366 escherichia coli metabolic network reconstruction for discovery of metabolic functions. BMC Syst Biol 6: 30.

8. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, et al. (2013) Ecocyc: fusing model organism databases with systems biology. Nucleic Acids Res 41: D605-12.

9. Brandes A, Lun DS, Ip K, Zucker J, Colijn C, et al. (2012) Inferring carbon sources from gene expression profiles using metabolic flux models. PLoS One 7: e36947.

10. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2: 886-97.

11. Burgard AP, Vaidyaraman S, Maranas CD (2001) Minimal reaction sets for escherichia coli metabolism under different growth requirements and uptake environments. Biotechnol Prog 17: 791-7.

12. Eker S, Krummenacker M, Shearer AG, Tiwari A, Keseler IM, et al. (2013) Computing minimal nutrient sets from metabolic networks via linear constraint solving. BMC Bioinformatics 14: 114.

13. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186-9.

14. Oliveira AP, Ludwig C, Picotti P, Kogadeeva M, Aebersold R, et al. (2012) Regulation of yeast central metabolism by enzyme phosphorylation. Mol Syst Biol 8: 623.

15. Costenoble R, Picotti P, Reiter L, Stallmach R, Heinemann M, et al. (2011) Comprehensive quantitative analysis of central carbon and amino-acid metabolism in saccharomyces cerevisiae under multiple conditions by targeted proteomics. Mol Syst Biol 7: 464.

16. Hyduke DR, Lewis NE, Palsson BO (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9: 167-74.

17. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320.

18. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium escherichia coli. Nature 427: 839-43.

19. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B-Methodological 58: 267-288.

20. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714-21.

21. Casanova R, Whitlow CT, Wagner B, Williamson J, Shumaker SA, et al. (2011) High dimensional classification of structural mri alzheimer's disease data based on large scale regularization. Front Neuroinform 5: 22.

22. Casanova R, Hsu FC, Espeland MA, Initi ADN (2012) Classification of structural mri images in alzheimer's disease from the perspective of ill-posed problems. Plos One 7.

23. Wang H, Nie FP, Huang H, Yan JW, Kim S, et al. (2012) From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. Bioinformatics 28: I619-I625.

24. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33: 1-22.

25. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. Bmc Bioinformatics 11.

26. Menendez P, Kourmpetis YA, ter Braak CJ, van Eeuwijk FA (2010) Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. PLoS One 5: e14147.

27. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432-441.

28. Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional ising model selection using l(1)-regularized logistic regression. Annals of Statistics 38: 1287-1319.

29. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. Nat Protoc 6: 1290-307.

30. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524-31.

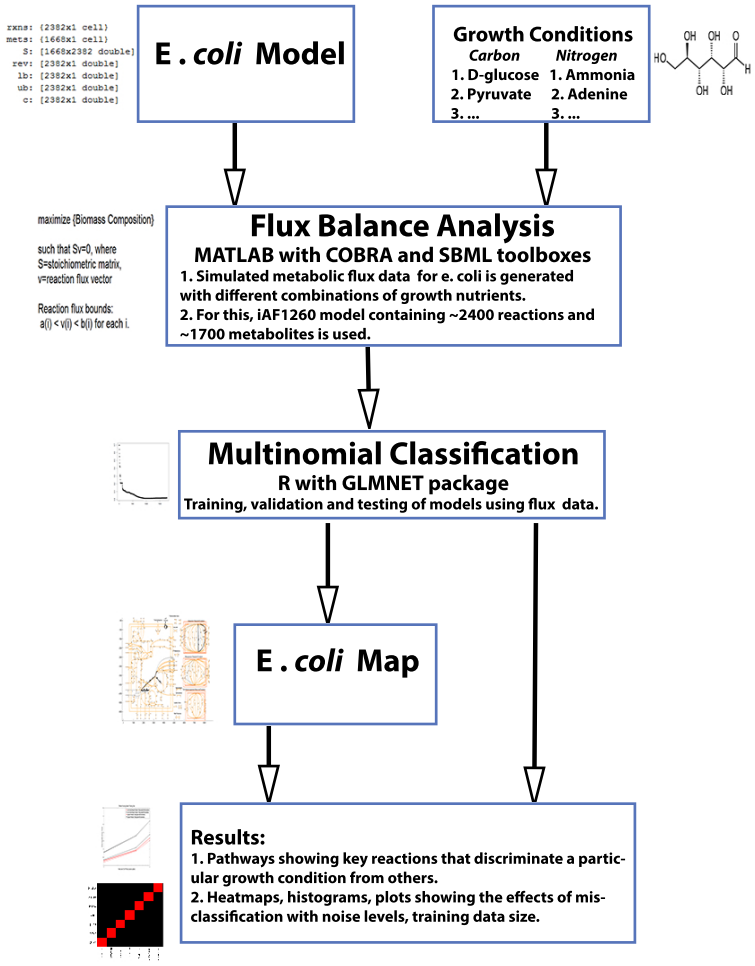

Figure 1: We obtained E. coli model and map from BiGG database. The key steps involved are Flux Balance Analysis and multinomial classification routines.

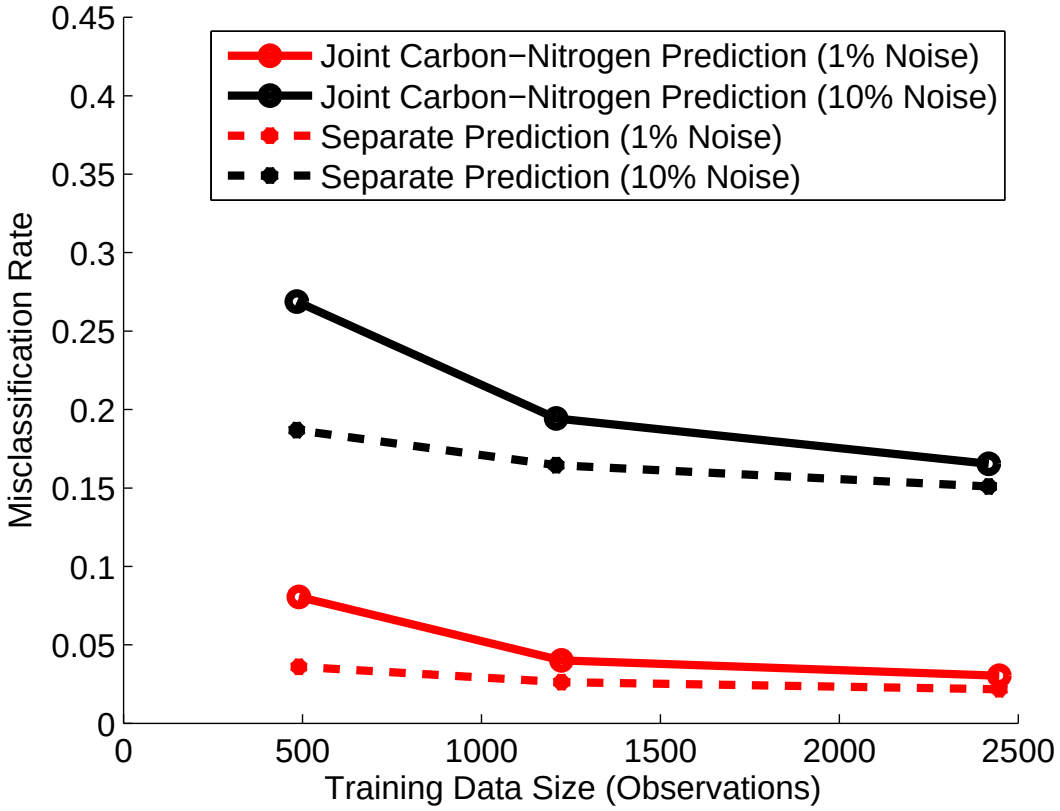

Figure 2: This plot shows that as training data size increases, the misclassification rate increases. This is tested for 2 different noise levels (1% and 10%).

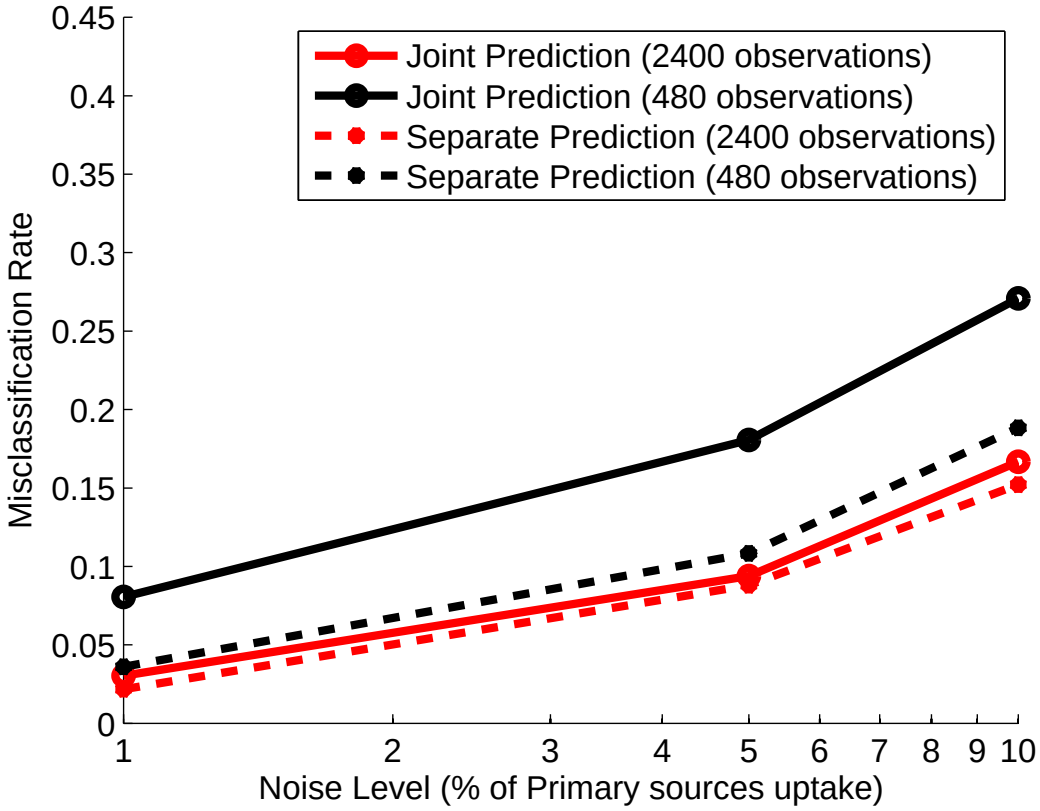

Figure 3: This plot shows that misclassification rate increases as noise increases in FBA models. This is tested for 2 different replicate sizes (480 and 2400 observations).

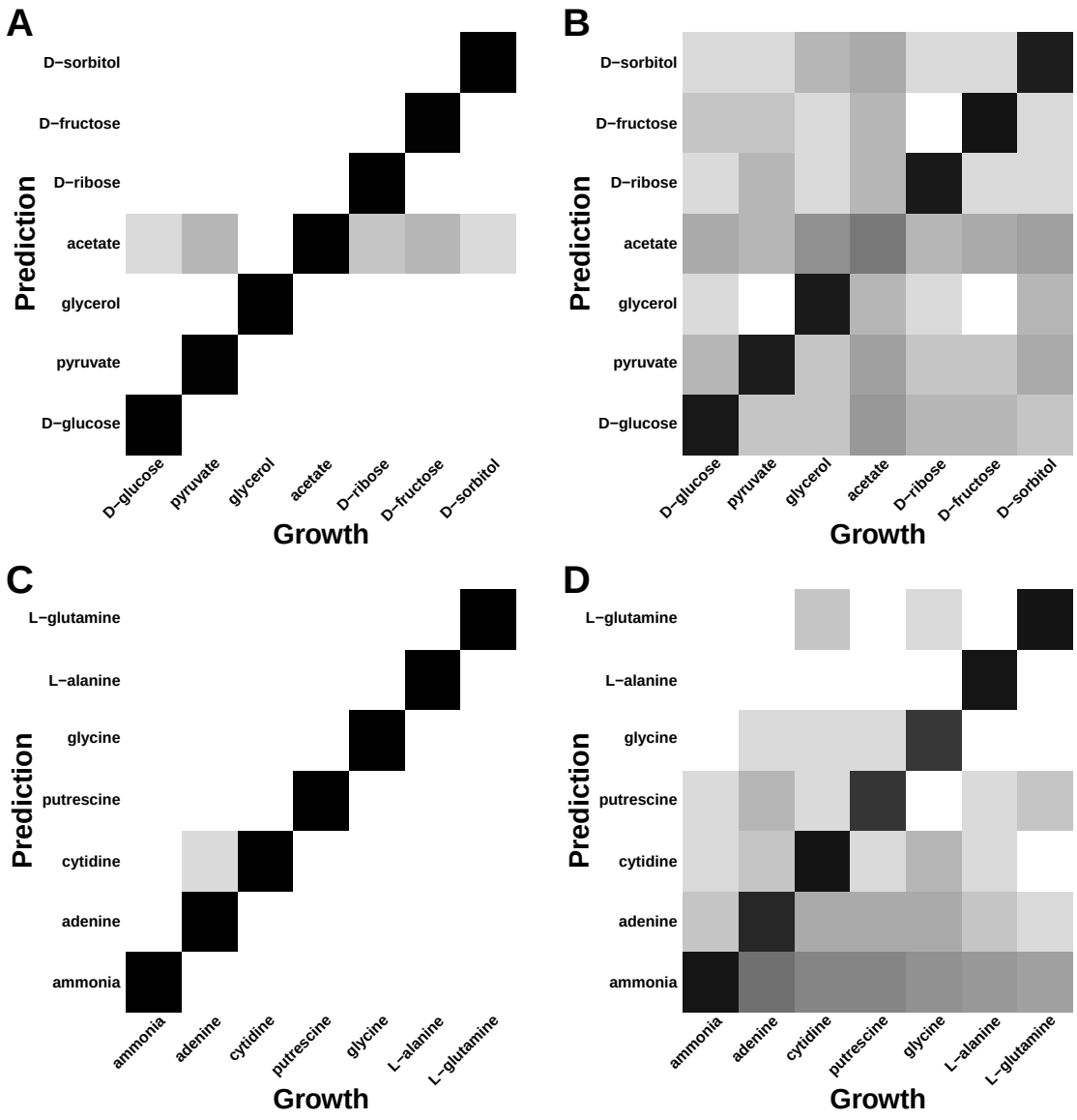

Figure 4: At 20% noise, most C sources are predicted to be acetate and most N sources are predicted to be ammonia.

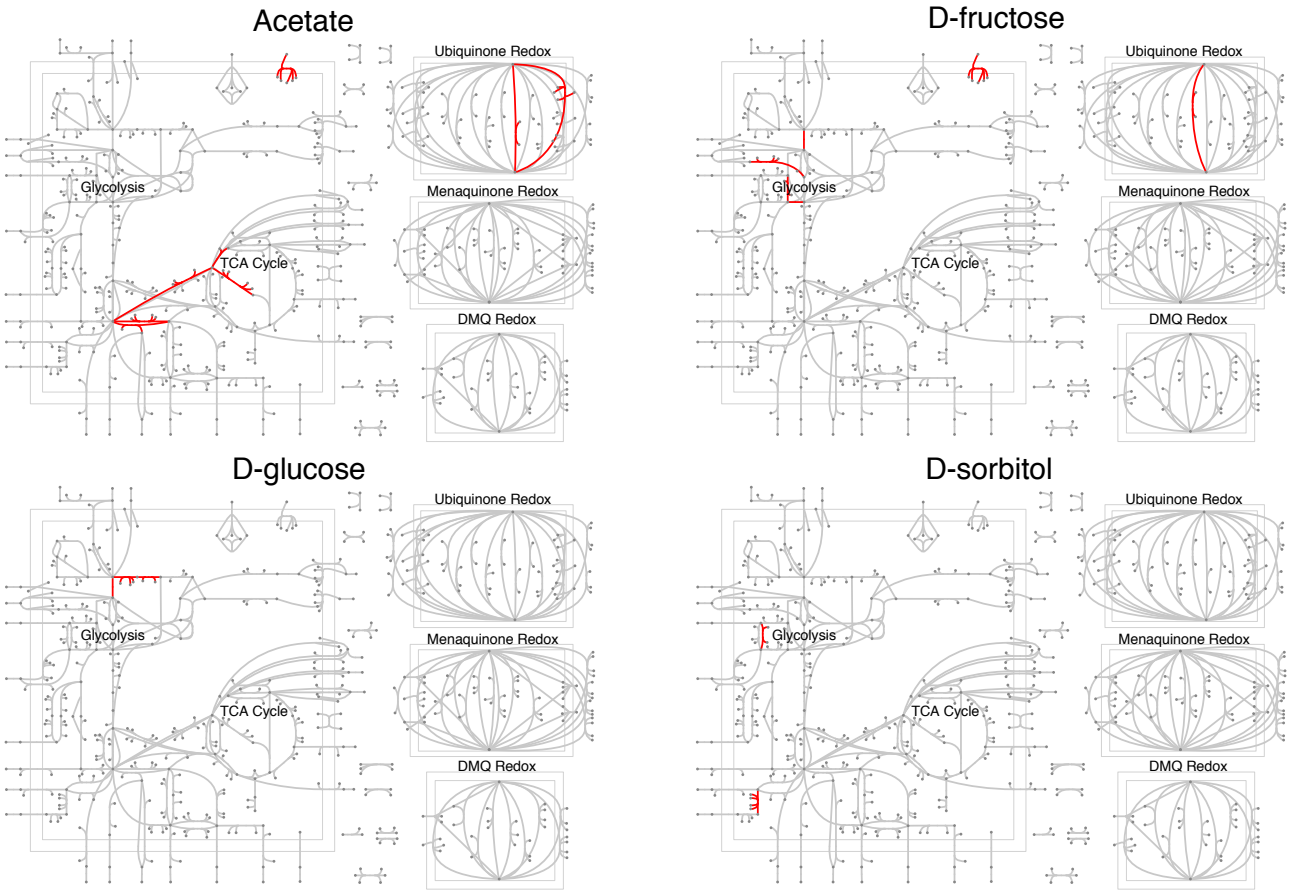

Figure 5: The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Out of 7 carbon sources, here we show 4 carbon sources and the key-reactions.

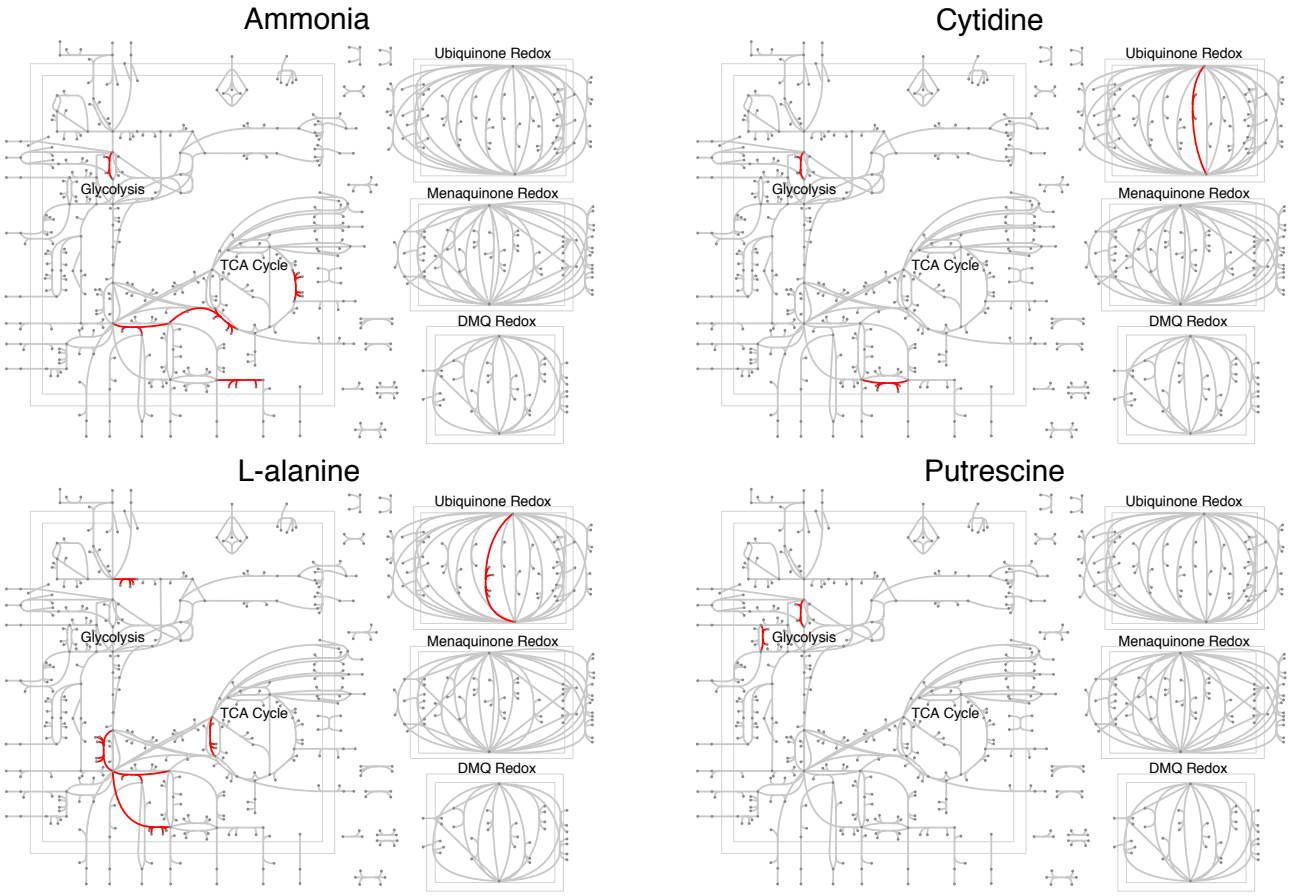

Figure 6: The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Here, the growth medium used are generally used for K-12 MG1655 strain.