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

Viswanadham Sridhara[1], Austin G. Meyer, 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

Biochemical information about the enzymes catalyzing reactions in metabolic pathways of microbes is obtained from databases, such as BiGG, BioCyc. This information along with flux balance analysis (FBA) is routinely used in metabolic engineering applications, such as predicting phenotypes given known substrate uptake or gene deletions. In this study, we developed a computational framework in identifying the growth conditions, given the phenotype (simulated flux data). For this, we generated metabolic flux data using FBA on E. coli iAF1250 model for different growth conditions. Using this simulated flux data, we then used model selection algorithm, LASSO, for identifying the key reactions that discriminate different growth conditions. In this study, we used 7 each of carbon/nitrogen sources along with more commonly used experimental media to grow K-12 MG1655 strain. We showed that separately predicting the carbon and nitrogen sources in carbon-nitrogen mixture is better than jointly predicting. On average, 9 features (reactions) were predicted per growth condition. The misclassification rate seems to increase with increasing background noise levels and decreasing training data size. We then mapped these features (reactions) on the E. coli central carbon metabolism to visually show the key genes/metabolites specific to a particular growth condition. For the more commonly used growth media, the models seems to predict at high accuracy too.

Author Summary

In the last decade, there is gaining popularity for predicting the reaction fluxes in metabolic network using criteria such as maximizing the biomass composition or maximizing the co-factors yield using genome based metabolic models. Moreover, metabolic pathways provide us a framework to integrate diverse kinds of high-throughput data i.e., transcriptomics/proteomics/metabolomics data. Here, we generated metabolic fluxes for different growth conditions and then used machine learning techniques to predict the key-reactions that discriminate different growth conditions. Such automatic identification of key-reactions would in turn also help experimentalists to quantitatively measure the respective metabolites involved (mass-spec based metabolomics), proteins that catalyze the reactions (mass-spec based proteomics) or the genes that encode these enzymes (sequencing based RNA-seq experiments).

Introduction

Microbial systems biology is developing at a rapid pace, with advances in sequencing technologies. Genome sequence along with the available annotation can be used to build a biochemical network of metabolic pathways. These metabolic pathways, when represented by mathematical models provide relationship of phenotype to its genotype. Such phenotype prediction given the growth nutrients or the mutant type (gene deletion etc) is generally carried out using Flux Balance Analysis (FBA) [1] on genome scale metabolic models (GEM) [2–4]. So, given the growth condition, reaction flux vector (phenotype) could be predicted. But can we predict the growth condition, given the simulated flux data? This is the question we are interested in answering as it has many applications. For example, microbes can be engineered in lab for useful purposes as well as deliberate attacks as shown in anthrax mailings. To identify the source of pathogen, mathematical models that can predict the growth condition given the cellular composition of the pathogen would be useful. Hence, in this study, the goal is to identify the reaction fluxes that discriminate the input growth conditions.

Here, we used E. coli iAF1260 GEM model [5] that is comprehensively and qualitatively well annotated by 2 groups for almost 2 decades [2,5–7]. For the current analysis, we used 7 each of carbon and nitrogen sources, to generate simulated flux data. These sources were previously shown to result in growth with this iAF1260 model [5]. Once we have the flux data, we then used machine learning techniques to predict the growth sources. To our knowledge, there were no studies that identified growth sources from flux data.

Machine learning algorithms use with computational biology has also been growing, given the high-dimensional nature of the problems in computational biology. In this biological data, generally the number of samples is far less than features. So, linear models cannot be used without reducing the number of covariates. For example, differential expression of tens of thousands of genes in microarray studies or identifying the SNPs in GWAS studies is a routine task in biomedical research now-a-days. Even though the sequencing technologies are becoming cheaper day-by-day, the number of samples sequenced is still low, compared to the number of features. In such cases, model selection algorithm LASSO [8] seems to perform well. LASSO methods are used in the past for genomics studies [9]. LASSO seems to perform well with other kinds of data, for example, classifying structural images of brain using MRI data [10–12].

So, in this study, we used LASSO [13] along with FBA to answer the following questions: 1. Can we predict the nutrient source on which the microbe is grown, given the simulated flux data? 2. In a mixture of different growth sources, can we predict each growth source sperately or only joint prediction a possibility? 3. Can we make any mechanistic insights into the predicted features (here, reactions in metabolic pathways)? These features (reactions) can later be linked to quanitiatve measurements in targeted OMICS studies.

Results

Improving the growth rate or identifying the enzymes key for improving specific metabolite concentrations is a routine task in metabolic engineering. FBA can be regarded as a technique to computationally guide experimentalists, for example, identifying a specific mutant or growth condition that increases a particular engineering target of interest. Here, we used FBA with machine learning techniques to predict growth sources from simulated flux data.

1. For FBA, we used COBRA toolbox [14] with MATLAB.

2. For classification, we used GLMNET [13] package with R.

3. We used E. coli model iAF1260, downloaded from BiGG database [15].

Below we describe the results obtained in this study.

Simulating metabolic flux datasets:

The iAF1260 metabolic model of K-12 MG1655 strain was obtained from BiGG database [15]. This model has 2382 reactions, that included the transport, biochemical and biomass composition reaction. For regression analysis, we removed all the transport reactions in iAF1260 model as these are collinear to the input growth sources, which we are trying to predict from flux data. Other details about this model are provided in methods section.

For different input growth conditions that are listed in Table 1, we generated fluxes. To see the effect of noise, we added randomly picked carbon and nitrogen sources from previously known list [5], in addition to the primary sources we want to predict. Use of noise also allows us to generate many replicates for the same combination of sources, although generating replicates by other ways exists, for example, by finding alternate optima [16].

We generated 100 replicates of pairwise combinations of 7 carbon and 7 nitrogen sources, i.e., 4900 observations. We then divided the dataset into training and test sets. We trained the models with training data using GLMNET package and used the test set for calculating the misclassification rates. We generated flux data and calculated misclassification rates in the same manner for other commonly used media described in Table 2. Below, we summarize the results.

Noise and training data size effects

Contamination in the growth media would generate a different reaction flux vector than without contamination. This would then have an effect on the classification rate of the nutreint sources from simulated flux datasets. To see the effect of varying noise (contamination) levels, we generated simulated flux datasets at different noise levels. We accomplished this by keeping the uptake rate of randomly picked carbon/nitrogen sources at 1% (0.2 mmol/gDWhr) of the primary sources. For example, 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 [5]. We then used the cross-validation GLMNET (logistic regression) to pick the model (regression coefficients) that classifies the sources. This cvGLMNET module outputs the misclassification rate on the test set, which we used to present our results.

Figure 2 shows that as background carbon and nitrogen sources interference increases, the misclassification rate increases. This is the case for both joint prediction and separate prediction of carbon and nitrogen sources. At 1% noise, the difference between joint and separate predictions is not a lot, but clearly at 10% noise, misclassification rate is lower for separate prediction. From this analysis, it looks like separate prediction performs better than joint prediction. This result has 2-fold advantages. 1. The classification models can be built by traning individual sources, making the classification relatively easy. 2. Since the number of experimental observations are generally small, joint prediction has less data to train compared to individual prediction. Even at 10% noise level with joint prediction, the misclassification rate is 25% using these models. On the other hand, given 7 carbon and 7 nitrogen sources, just by chance the misclassification rate is expected at 93%.

To make sure the above result has no biases in the uptake amounts, we looked at these uptake distributions for carbon and nitrogen sources. The distributions can be found in supplementary info as S1 figure. Most of the carbon sources uptake seems to be limited by nitrogen and few nitrogen sources seems to have limited carbon for their growth. So, 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. These results also suggest that separate prediction works better than joint prediction.

Noise or contamination is not the only cause for misclassification with these models. Traning data size has a dominant effect on classification too, as a very small training set would not generate a model that can classify at higher accuracy. To see this effect, we took 3 subsets of original dataset that correspond to 240, 480 and 2400 observations. Using this data, we trained new models with GLMNET package. Our results show that as training data size increases, misclassification decreases. For example, in figure 3, clearly, with a large training data set size of 2400 observations, the classification is better compared to that of 480 observations. The low number of observations is more common with realworld experimental datasets and clearly separate prediction outperforms joint prediction with a training data size of 480 observations.

To understand what the models mean, we investigated more into the features that classify both the carbon and nitrogen sources. On average, the number of features for each source seems to be 9. This is relatively small given the total number of reactions (1443). In one study, it has been shown that for a specific growth condition, the total number of reactions that seem to have high-flux are very low [17], although the reactions with low-flux can be indicative of the growth source too. To understand this result in a different way, we used a simpler strategy. For each reaction that has a non-zero weight in the model, we knocked-off the reaction flux, trained a new model (separate prediction) and calculated the test accuracy. Except for 3 reactions (predictors), the misclassification rate with the newer models seems to be the same for the rest 126 key-reactions (features predicted for 7 carbon and 7 nitrogen sources). This indicates that alternate models exist, depending on the input set of reaction fluxes.

The possible growth nutrients could be many and some of these sources are a different variants of others. For example, maltose is 2 units of glucose. To see how a model classifies a growth source that is not used in training it, we generated flux data using maltose and nitrogen sources that were used above in the study. Since maltose is 2 units of glucose, we expect the model to classify this dataset as glucose (separate prediction), or glucose + nitrogen source (joint prediction). However, the result is quite interesting. Individual prediction seems to classify maltose as glucose more than 85% of the times and the nitrogen sources were predicted correctly more than 95% of the times. But when we tried to predict using the joint model, the nitrogen sources were mispredicted most of the times and the carbon source was predicted as D-sorbitol most of the times.

Mechanistic Insights

To have mechanistic insights into the misclassification caused by noise levels, we drew the heatmap (Figure 4) of the actual source against the predicted sources at 2 noise levels (1% and 20%). As seen in figure 4, the results indicate that as noise level increases from 1% to 20%, carbon sources get predicted as actate, while N sources get predicted as ammonia most of the times. So, we looked into the reactions that are key in classifying acetate and ammonia sources. Looking at these reactions that are predicted by GLMNET, it seems that they either are part of the TCA cycle or lead to the TCA cycle. This might be due to the objective function used, which is maximizing biomass composition. 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.

We mapped the predicted reactions (from separate prediction at 1% noise level and using 2400 observations for training data size) for each growth source onto the E. coli central metabolism model. We overlayed these reactions on E. coli central metabolism along with the metabolites involved in the reaction. Manual validation of these reactions indicate that indeed some reactions seem specific to growth source, for example glycerol dehydrogenase, glycerol kinase and glycerol-3-phosphate dehydrogenase reactions are specific to ”glycerol” source [**?**]. E. coli metabolic map showing predicted reactions for 4 carbon and nitrogen sources is shown in figures 5 and 6. All reactions that are predicted by GLMNET is provided in supplementary information as Table S1.

Large-scale prediction of Carbon/Nitrogen sources

Since individual prediction is shown to perform better than joint prediction, we did large-scale prediction of the comprehensive list of 174 carbon and 78 nitrogen sources used previously [5]. We used the similar methodology as we used earlier i.e., we generated simulated fluxes for all pair-wise combinations of carbon/nitrogen sources (174*78) for 2 replicates. We used 1 replicate for training the model and then we used the 2nd replicate for testing. The misclassification rate for carbon sources is X%, while the misclassification rate for nitrogen sources is Y%. It should be noted that joint prediction would result in 174*78 classes with 1 replicate for training, while the separate prediction would have 78 replicates for training carbon sources and 174 replicates for training nitrogen sources (pair-wise designs in FBA).

Further validation with other media

We did similar analysis for the media that are more commonly used in E. coli experiments. The media we used were AB minimal media, ATCC media, Davis Mingioli media and Bochneder defined minimal media. Since the number of different media used in our analysis was 4, we were able to classify at a higher rate even at higher noise levels and lower number of replicates. The misclassification rate with the test set is found to be less than 1% for noise levels upto 20%. 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.

Discussion

The biochemical network of metabolic pathways, along with FBA, are used for many diverse kinds of applications. For example, a study identified minimal set of reactions that are necessary for growth for a particular growth condition [18]. There is another study published recently that identified minimal set of nutrients for growth conditions [19]. Another study by Ibarra et. al., [20] looked at the growth of E. coli K-12 MG1655 strain on glycerol for 40 days (700 generations) and they saw an increase in the growth rate with generations. Other than the above studies, this set of pathways when integrated with other diverse types of data, have many applications as discussed in these 2 reviews [21,22]. However, in this study we use the simulated flux data (obtained from different growth conditions) to train mathematical models that can predict the input growth condition.

In the current study, we used E. coli metabolic model for the analysis. We used iAF1260 genome-based metabolic model of K-12 MG1655 strain with COBRA toolbox to generate the flux data. We then used GLMNET package to predict different growth conditions, including the media typically used to grow this K-12 strain. Although there is a latest realease of this model [7], our results should still hold with the new model. One of the reasons we used the previous model is because iAF1260 model is used in many FBA studies and is shown to correlate well with experimental data. We used the constraints of maximizing growth to find a particular reaction flux vector for a given set of growth conditions. Although, the experimental data can be used to refine the constraints [23], we used the default lower and upper bounds in the model. We generated replicates for the primary growth sources of interest, by solving for reaction flux vector at different varying background levels of growth sources (contamination). We then used the simulated flux data to train mathematical models to predict the input growth sources.

Our results indicate that separate prediction of the mixture of growth sources performs well than joint prediction at a noise level, for a given size of training data. This result is interesting, as we intially thought there would be lot of cross-talk between the carbon and nitrogen sources. This result is advantageous, as the model trained on known growth sources would predict those sources, when present, at a high confidence level when tested on species grown on unknown nutrients. Although the simulated flux data is a good first step to train the models, ideally we would like to integrate these models to use the experimental data. For example, combining flux analysis data with phosphoproteomics data to deduce functionality of enzymes is described in this study [24]. An interesting study to understand the adaptation of yeast metabolism to the growth conditions is studied, using enzymes in metabolic model pathways to validate the quantitative proteomics data [25].

To our knowledge, LASSO methods are not used in the analysis of metabolic pathways. However there are studies that used LASSO on other biochemical networks such as gene regulatory networks [26]. In machine learning algorithms, there are other regression methods that use regularization techniques other than LASSO. Graphical LASSO [27] and Ising Markov Random Field models [28] are also used to complement with LASSO. We would like to explore these and other models to see if we can improve the classification rate.

Materials and Methods

We used MATLAB and R for this study. For flux balance analysis, we used COBRA toolbox [14] with MATLAB and for multinomial regression, we used GLMNET package [13] with R. The methods are described in detail below.

Flux Balance Analysis:

Biochemical network represented mathematically (metabolites and reactions) is used in 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.

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

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, a certain objective function of interest is optimized, for example here, we used a constraint to maximize biomass composition. This resulted in uniqe solution.

Equation (4) can be solved for different input growth conditions i.e., by changing lower bounds of transport reactions that uptake a specific nutrient. To create replicates, for each design, we introduced varying noise levels. 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 [29]. 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 [5] 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 [5]. 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 [5]. 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

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

From this, we can model

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

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

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

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, Marx lab members at Harvard 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, A.G.M, 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. 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.

6. 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.

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

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

10. 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.

11. 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.

12. 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.

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

14. 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.

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

16. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5: 264-76.

17. 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.

18. 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.

19. 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.

20. 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.

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

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

23. 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.

24. 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.

25. 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.

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. 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.

Tables

Table 1. Carbon and nitrogen sources used in the study.

Carbon Sources	Nitrogen Sources
D-glucose	Ammonia
Pyruvate	Adenine
Glycerol	Cytidine
Acetate	Putrescine
D-ribose	L-glycine
D-fructose	L-alanine
D-sorbitol	L-glutamine

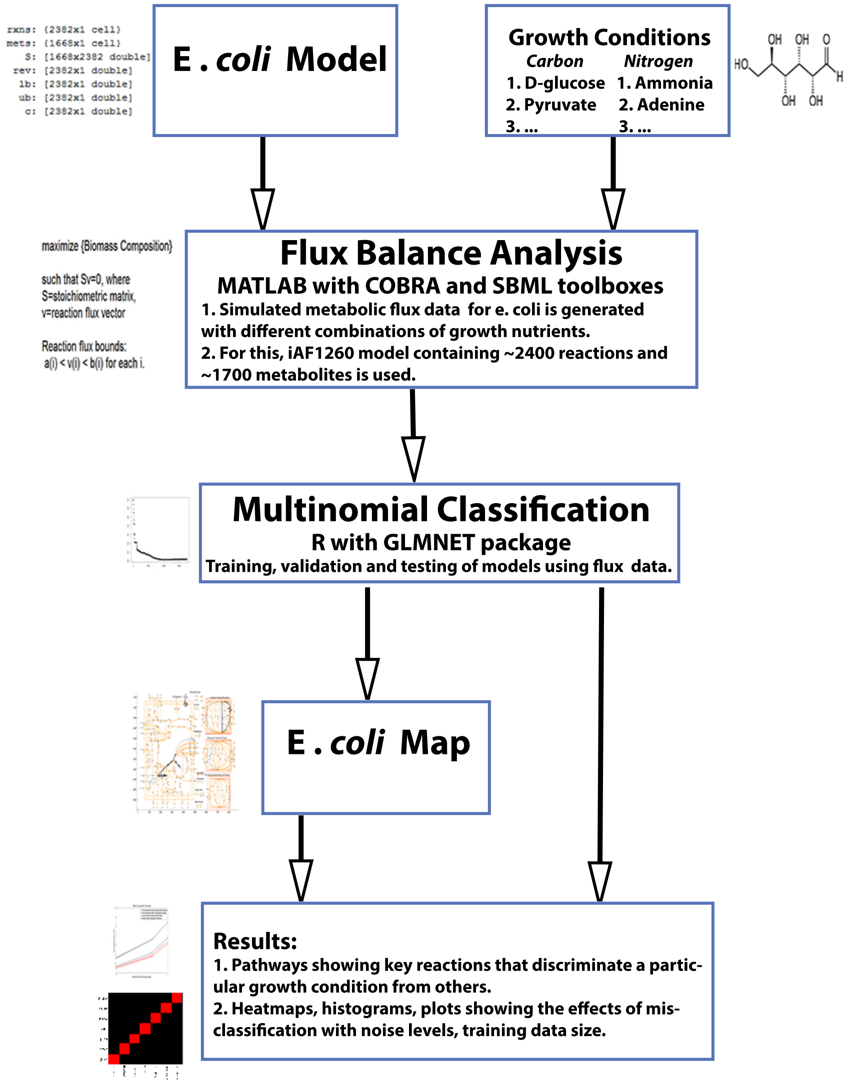
7 carbon and 7 nitrogen sources are used in this study.

Table 2. Growth media for K-12 MG1655.

Other media	Composition
AB medium	D-glucose/Ammonia
ATCC medium 57	Glycerol/L-lysine/Ammonia
Bochner medium	Pyruvate/Ammonia
Davis Mingioli medium	D-glucose/Citrate/Ammonia

4 commonly used growth media for E. coli MG1655 strain that were picked from EcoCyc website were also used in this study.

Figure Legends



```
rxns: {2382x1 cell}
mets: {1668x1 cell}
   S: [1668x2382 double]
 rev: [2382x1 double]
  lb: [2382x1 double]
  ub: [2382x1 double]
   c: [2382x1 double]
```

1. D-glucose
2. Pyruvate
3. ...

1. Ammonia
2. Adenine
3. ...

maximize {Biomass Composition}

such that Sv=0, where
S=stoichiometric matrix,
v=reaction flux vector

Reaction flux bounds:
a(i) < v(i) < b(i) for each i.

1. Simulated metabolic flux data for e. coli is generated with different combinations of growth nutrients.
2. For this, iAF1260 model containing ~2400 reactions and ~1700 metabolites is used.

Training, validation and testing of models using flux data.

1. Pathways showing key reactions that discriminate a particular growth condition from others.
2. Heatmaps, histograms, plots showing the effects of misclassification with noise levels, training data size.

Figure 1. Flowchart describing methodology used in this study. We obtained E. coli model and map from BiGG database. The key steps involved are Flux Balance Analysis and multinomial classification routines.

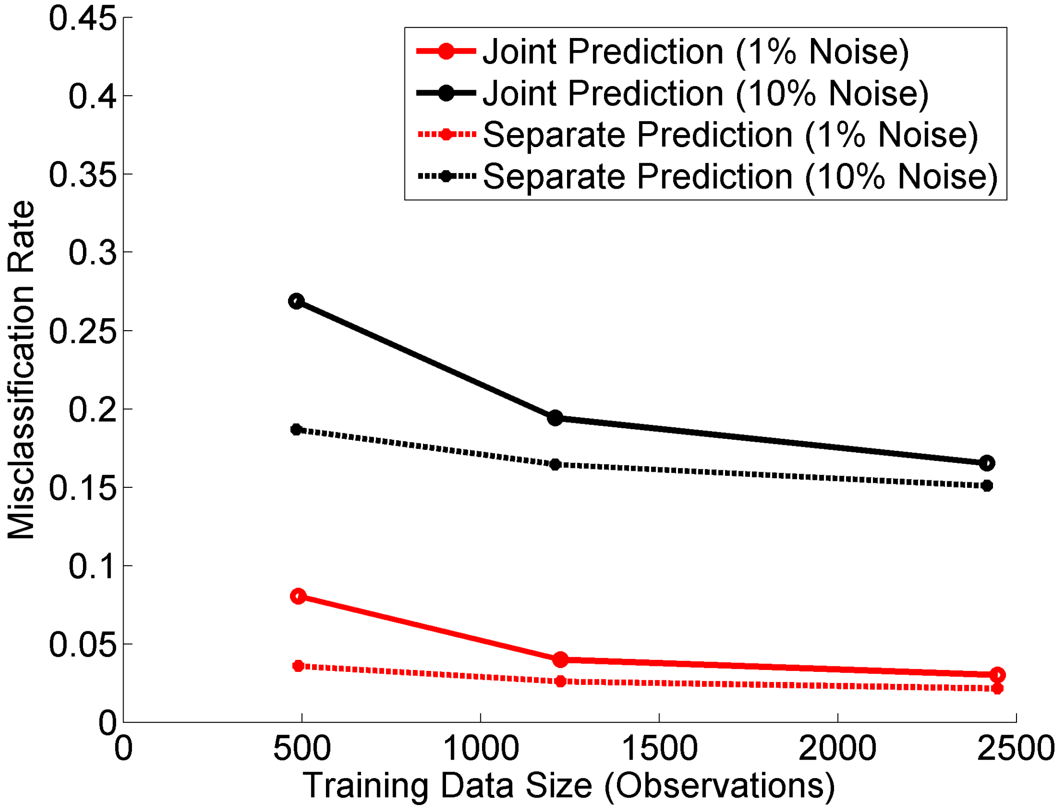


Figure 2. Misclassification rate for different replicate sizes. This plot shows that as training data size increases, the misclassification rate increases. This is tested for 2 different noise levels (1% and 10%).

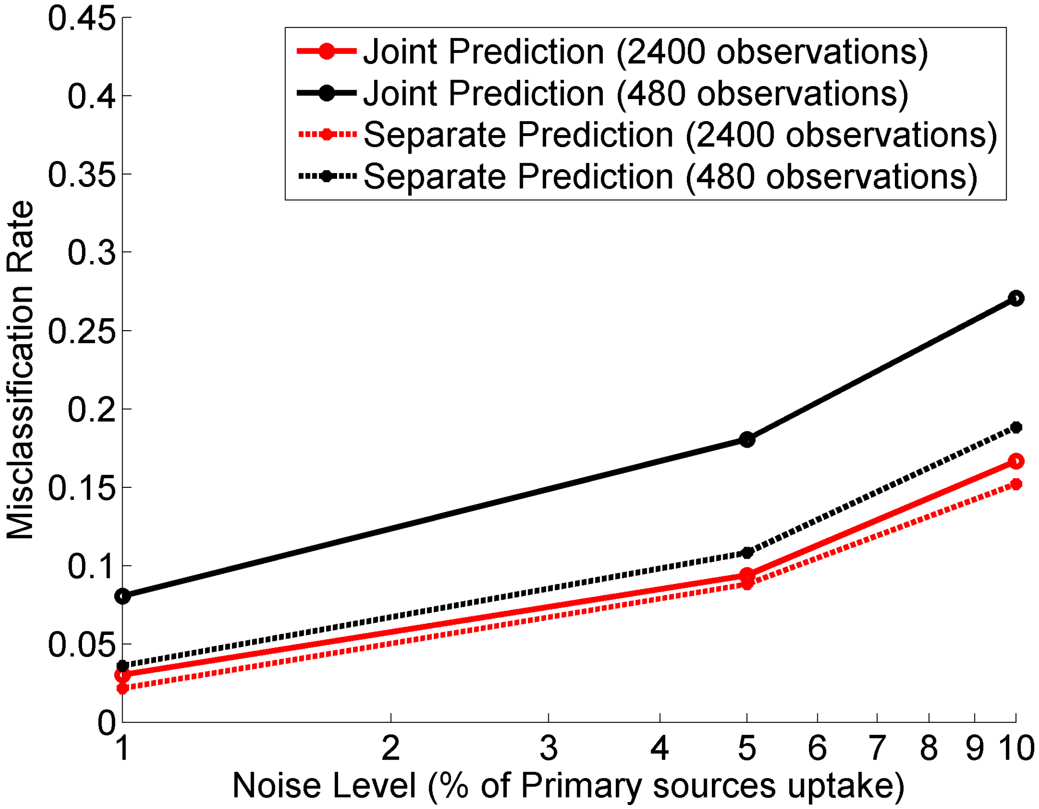


Figure 3. Misclassification rate at different noise levels. This plot shows that misclassification rate increases as noise increases in FBA models. This is tested for 2 different replicate sizes (480 and 2400 observations).

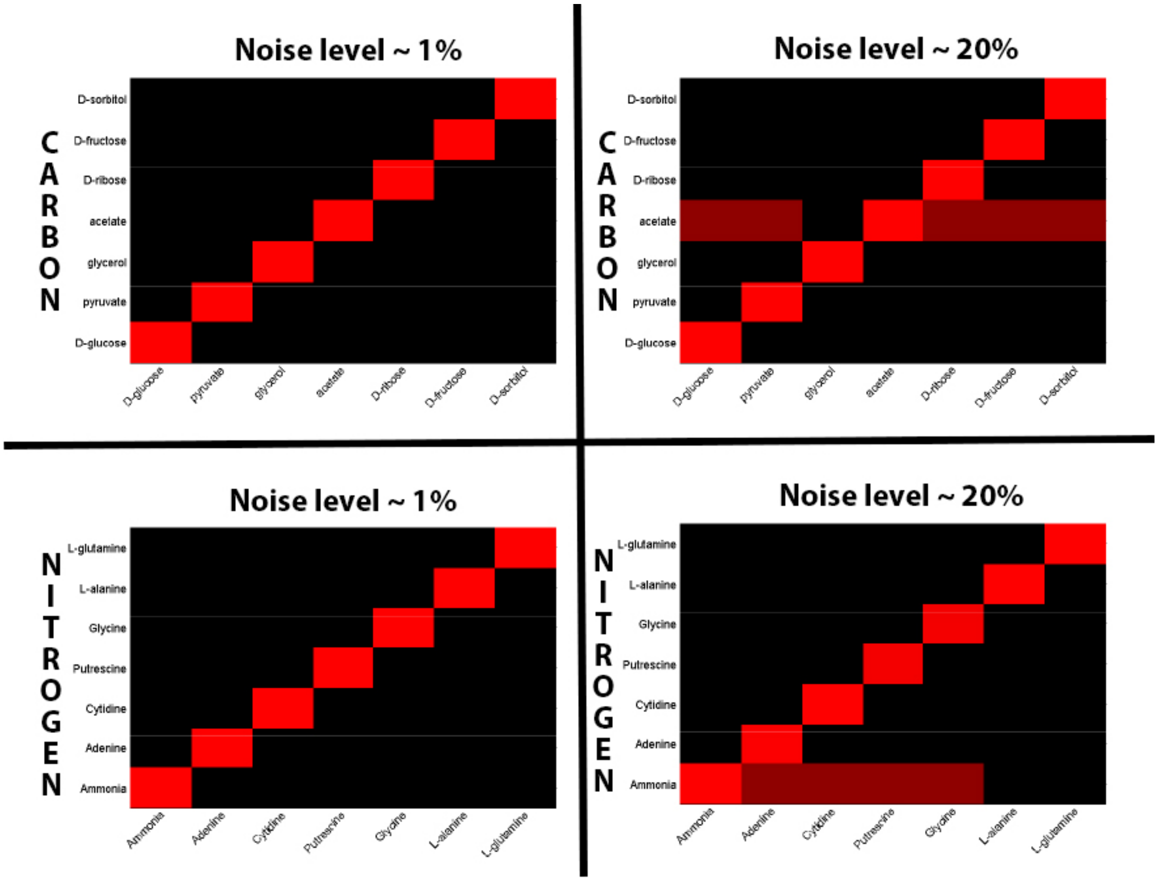


Figure 4. Heat maps with actual sources as columns and predicted ones in rows. At 20% noise, most C sources are predicted to be acetate and most N sources are predicted to be ammonia.

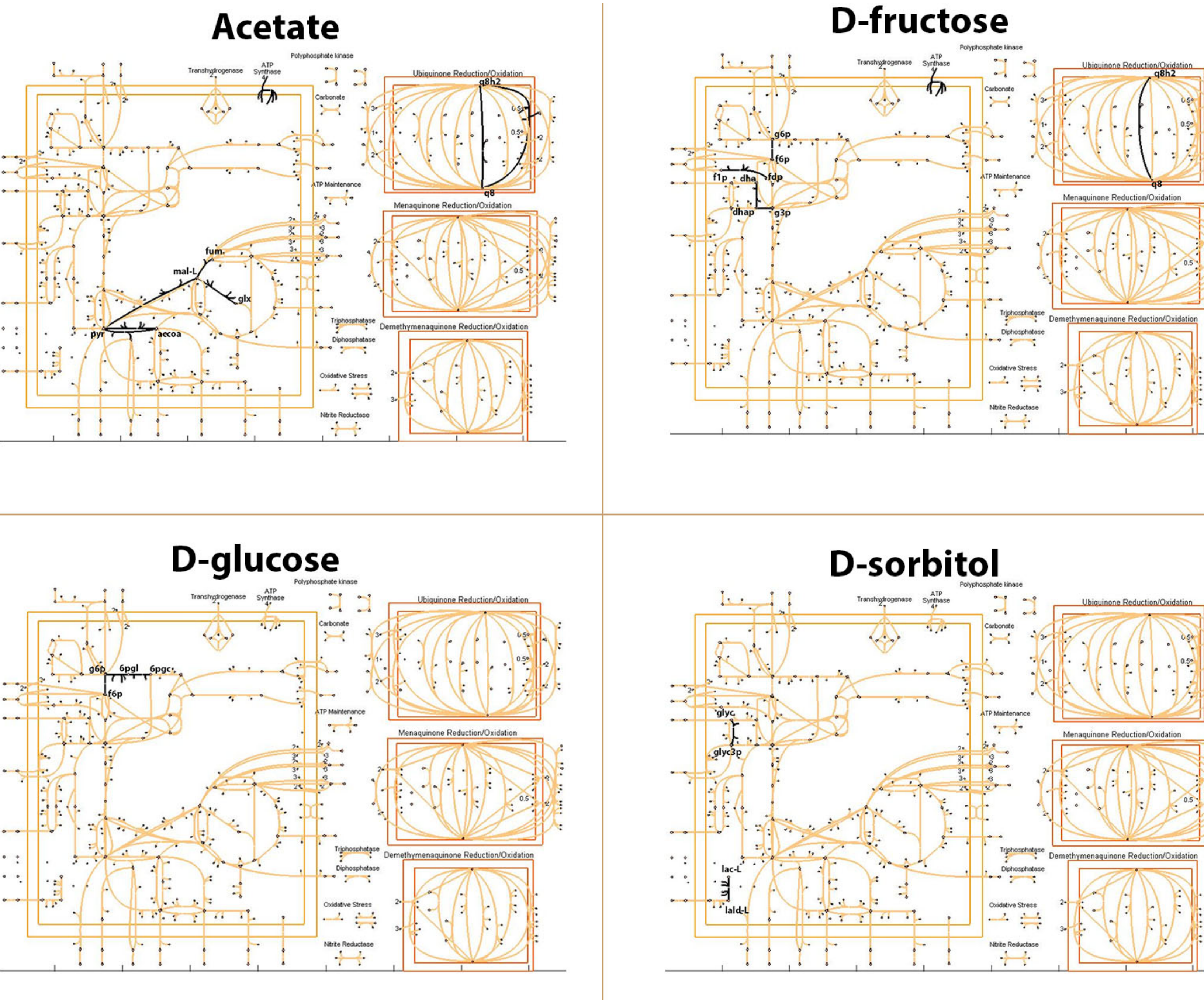


Figure 5. Key-reactions that discriminate Carbon sources. The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Out of 7 carbon sources, here we show 4 carbon sources and the key-reactions.

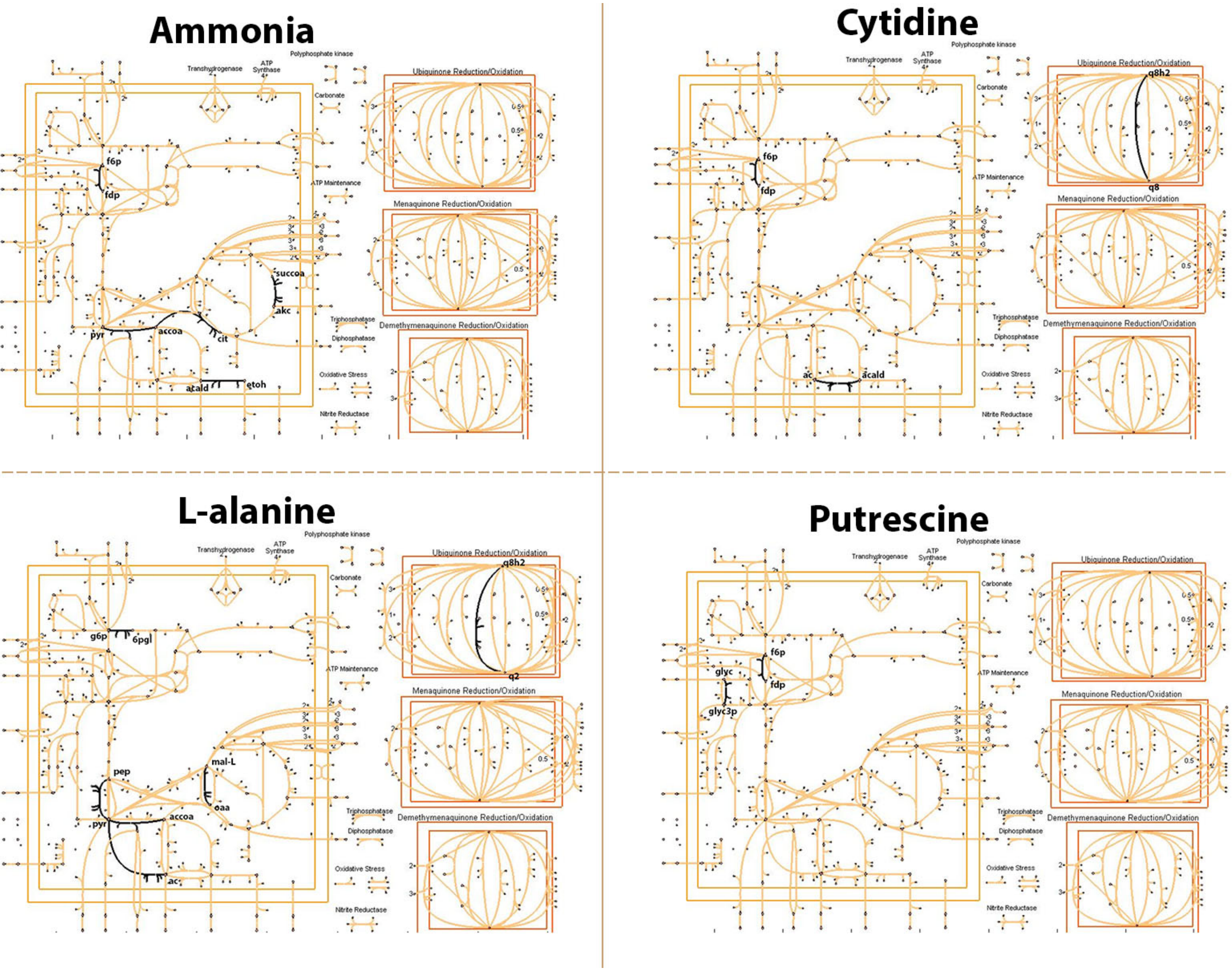


Figure 6. Key-reactions that discriminate Nitrogen sources. The key-reactions identified by GLMNET package were mapped onto E. coli central metabolism to visually show the differences between different growth conditions. Out of 7 nitrogen sources, here we show 4 nitrogen sources and the key-reactions.