Bioinformatics, 39(5), 2023, btad186
https://doi.org/10.1093/bioinformatics/btad186
Advance Access Publication Date: 11 April 2023
Original Paper

OXFORD

Systems biology

Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling

Joshua A.M. Kaste [1,2] and Yair Shachar-Hill [2,*]

[1]Department of Biochemistry and Molecular Biology, Michigan State University, 603 Wilson Rd., East Lansing, Michigan 48824, United States
[2]Department of Plant Biology, Michigan State University, 612 Wilson Rd., East Lansing, Michigan 48824, United States

*Corresponding author. Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA. E-mail: yairhill@msu.edu
Associate Editor: Inanc Birol

Received 28 September 2022; revised 18 January 2023; accepted 6 April 2023

Abstract

Motivation: The accurate prediction of complex phenotypes such as metabolic fluxes in living systems is a grand challenge for systems biology and central to efficiently identifying biotechnological interventions that can address pressing industrial needs. The application of gene expression data to improve the accuracy of metabolic flux predictions using mechanistic modeling methods such as flux balance analysis (FBA) has not been previously demonstrated in multi-tissue systems, despite their biotechnological importance. We hypothesized that a method for generating metabolic flux predictions informed by relative expression levels between tissues would improve prediction accuracy.

Results: Relative gene expression levels derived from multiple transcriptomic and proteomic datasets were integrated into FBA predictions of a multi-tissue, diel model of Arabidopsis thaliana's central metabolism. This integration dramatically improved the agreement of flux predictions with experimentally based flux maps from ^{13}C metabolic flux analysis compared with a standard parsimonious FBA approach. Disagreement between FBA predictions and MFA flux maps was measured using weighted averaged percent error values, and for parsimonious FBA this was169%–180% for high light conditions and 94%–103% for low light conditions, depending on the gene expression dataset used. This fell to 10%-13% and 9%-11% upon incorporating expression data into the modeling process, which also substantially altered the predicted carbon and energy economy of the plant.

Availability and implementation: Code and data generated as part of this study are available from https://github.com/Gibberella/ArabidopsisGeneExpression
Weights.

1 Introduction

A grand challenge for systems biology is the ability to accurately predict complex phenotypes from omic datasets based on functional principles and mechanisms. Patterns of cellular metabolism—flux maps—are one such complex phenotype (Ratcliffe and Shachar-Hill 2006), for which tools to predict phenotypes from basic assumptions have proven useful in exploring and designing metabolic capabilities (Burgard et al. 2003; Orth et al. 2010; Chen et al. 2011). Methods to quantify flux maps from labeling data now allow the testing of such predictions in both simpler and multicellular systems. However, the integration of omic data to improve the accuracy of flux predictions is still at an early stage.

Metabolic flux predictions are also important for real-world applications since modifying an organism's metabolic activity in order to achieve some practical aim, such as overproducing a specific metabolite, is central to many biotechnology projects. As in other areas of engineering, metabolic engineering can benefit from mathematical models that describe and predict the behavior of the relevant system(s). Researchers have developed two major modeling approaches to address this need: (i) ^{13}C-metabolic flux analysis (^{13}C-MFA) and (ii) flux balance analysis (FBA; Orth et al. 2010; Antoniewicz 2015). With ^{13}C-MFA, steady-state or kinetic isotopic labeling data for metabolites in a small- to medium-sized network are used to obtain estimates of the net and exchange fluxes through that network (Antoniewicz 2015). These metabolic flux maps are regarded as the most reliable measures of in vivo metabolic fluxes; however, the throughput of this technique is limited by the large amounts of isotop-ic labeling data and other measurements needed to generate each flux map. FBA, which is based on applying conservation principles to a network of reactions using one or more assumptions about the functional objective(s) driving biological organization, requires substantially less experimental input data and is therefore an attractive and commonly used metabolic modeling technique.

© The Author(s) 2023. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

FBA and related metabolic modeling methods in microbial systems, together with genome-scale models (GEMs) that represent the biochemical reactions encoded in an organism's genome, have enabled radical modification of microbial central metabolism (e.g. Gleizer et al. 2019) and substantial improvements in bioproduct yields (e.g. Park et al. 2007, Lee et al. 2007). These methods can, e.g. allow bioengineers to predict the behavior of their system and identify interventions, such as gene knock-outs or knock-ins, that will help them modify the organism's phenotype (Burgard et al. 2003, Tepper and Shlomi 2009). However, many metabolic engineering applications require the modification not of microorganisms, but of multicellular eukaryotes like plants. Most GEMs of plants to date (e.g. Poolman et al. 2009; Dal'Molin et al. 2010ab; Saha et al. 2011; Arnold and Nikoloski 2014), have treated plants, which are composed of multiple tissues with substantial functional diversity, as single-tissue aggregated metabolic networks. This has motivated the creation of "multi-tissue" GEMs to investigate source-sink dynamics and resource allocation, with the earliest efforts in this space focusing on the interplay between mesophyll and bundle-sheath cells in C4 photosynthesis (Dal'Molin et al. 2010b; Shaw and Cheung 2020).

Re-engineering of plant metabolism on the scale seen in microbial systems has not, to date, been possible and predictive modeling has been neither validated in detail nor applied to successful plant metabolic engineering. This is partly due to the ease and high throughput of microbial transformation relative even to model plant systems. In addition to the greater experimental demands, the metabolic modeling of these systems is also substantially harder. There is, consequently, a relative lack of MFA datasets with which to compare the predicted flux maps from FBA in plants. This contrasts with the availability of rich multiomic datasets combining flux estimates with transcript and protein data for a number of different genotypes and growth conditions in systems like Escherichia coli (Ishii et al. 2007). The challenges involved in generating ^{13}C-MFA flux maps for plants make improvement of plant FBA flux predictions an attractive path towards replicating the biotechnological successes seen in microbes.

An appealing approach to improving the quality of plant FBA predictions is the integration of additional network-wide data from transcriptomic and proteomic datasets. Gene expression data—particularly transcript data—are substantially easier to generate than ^{13}C-MFA flux maps. Previous attempts at integrating gene expression datasets into metabolic flux predictions have been reviewed elsewhere (Machado and Herrgård 2014; Vijayakumar et al. 2017). Methods developed before 2014 were evaluated on the basis of their ability to improve upon parsimonious FBA (pFBA; Lewis et al. 2010) in terms of their predictions' agreement with MFA-estimated fluxes in microorganisms and were found to not do so reliably (Machado and Herrgård 2014). A key limitation of these studies was a lack of comparison of FBA predictions against ^{13}C-MFA-derived flux estimates. This lack of comparison against ^{13}C-MFA is shared by the plant FBA literature, in which we are aware of only a small number of evaluations under heterotrophic conditions in green algae (Boyle et al. 2017), Arabidopsis cell cultures (Williams et al. 2010; Cheung et al. 2013), and Brassica napus embryos (Hay and Schwender 2011). Since then, several studies have developed algorithms benchmarked by their ability to make predictions in agreement with empirical flux maps derived from MFA studies (Tian and Reed 2018; Pandey et al. 2019; Ravi and Gunawan 2021). These studies have focused on unicellular organisms or animal tissues modeled in isolation. Their application to FBA in more complex systems is limited by the large number of resource-intensive MFA datasets needed to calibrate them (Tian and Reed 2018) or their need for a reference expression dataset paired with an assumed-correct flux map (Pandey et al. 2019; Ravi and Gunawan 2021).

To improve the accuracy of FBA in multicellular systems, particularly plants with their complex metabolic networks, we developed a method that integrates tissue-atlas data from multi-tissue systems into the flux-minimization procedure employed in pFBA. This method incorporates evidence from gene expression datasets into FBA metabolic flux predictions by applying weights to individual reactions according to the relative transcript or protein expression of the gene(s) assigned to those reactions between different modeled tissues. The method is evaluated on its ability to make predictions in agreement with MFA flux maps. We demonstrate substantial improvements in the agreement of our FBA-predicted fluxes with flux estimates from a ^{13}C-MFA study on Arabidopsis thaliana rosette leaf central metabolism (Ma et al. 2014). Finally, we show that multiple gene expression datasets, when used as inputs, result in similar improvements in agreement and that this result generalizes across different MFA flux maps. This approach has particular potential for plant and animal systems for which there are only a limited number of experimental flux maps.

2 Materials and methods

2.1 Overview of approach

Our method makes two key assumptions: (i) Metabolic flux maps predicted from pFBA (Lewis et al. 2010), minimizing the sum total of flux through the network, are more likely to reflect real flux maps than ones not subject to this constraint, and (ii) A reaction present in two tissues A and B catalyzed by an enzyme encoded by a gene that is highly expressed in A and poorly expressed in B is likely to carry higher flux in tissue A.

We incorporate assumption 1 by making the objective function of our FBA optimization the minimization of total flux, the same as pFBA (Lewis et al. 2010). This is represented mathematically as finding the minimum value of the linear combination of all fluxes in the network, with each flux v_i multiplied by a corresponding coefficient c_i:

$$\min \sum_{j2\mathrm{Reactions}} c_j v_j \quad (1)$$

Where Reactions is the list of all reactions j in the network, v_j is the flux through a reaction j, and c_j is the coefficient—hereafter referred to as a penalty weight since it represents a penalization of the likelihood of using a reaction j to carrying flux. When c_j takes a value of 1 for all reactions, our method reduces to pFBA, which can be seen as the limiting case of gene expression having no influence in predicting network flux patterns. We incorporate assumption 2 by calculating, for each reaction in our network model, a coefficient derived from the relative expression of genes encoding the enzyme(s) that catalyze that reaction between the different tissues in the gene expression dataset. The association between reactions and genes is captured by the gene-protein-reaction (GPR) terms in the model. This results in reactions mapped to relatively highly expressed genes receiving small values of c_j and reactions mapped to minimally expressed genes receiving large ones. This use of the coefficient vector to account for relative expression evidence is related to the approach taken in Jenior et al. (2020). However, among other differences in implementation, the two methods differ in their assumed relationship between gene expression and flux and their application. Our method compares gene expression across tissues within a multi-tissue model to generate more accurate flux predictions, rather than comparing the expression of genes to the most expressed gene in a dataset as a proxy for transcriptional investment and a way of generating context-specific models.

2.2 Model construction and dataset selection

The A.thaliana core metabolism model developed in Arnold and Nikoloski (2014) was used as the basis for a multi-tissue diel model. This model was chosen due to its rich GPR annotation and focus on central metabolism. The core model was duplicated six times to create leaf, stem, and root versions of the model for both day and night, which were interconnected by transporters allowing the movement of specific compounds and metabolites. The substrates, products, and constraints applied to the model can be found in the Supplementary Methods. The model used in this study can be found in Supplementary Dataset S2.

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

^{13}C-MFA flux maps were obtained in planta in A.thaliana by Ma et al. (2014), and these were used as the empirical best estimates of flux distributions. Although there are not any other ^{13}C-MFA flux maps available of autotrophic A. thaliana leaves Szecowka et al. (2013) provide estimates of select fluxes in autotrophic A. thaliana leaf central metabolism, which we used for additional confirmation of our method's efficacy. The pairing of fluxes in both flux studies to the FBA network is described in Supplemental Dataset S1.

We searched the literature for high-quality, high-coverage RNA-seq, and quantitative proteomic tissue atlases and found two suitable datasets meeting these criteria: Klepikova et al. (2016) and Mergner et al. (2020). The proteomic dataset from Mergner et al. (2020) is a mass spectrometry-based quantitative proteome that reports IBAQ (Intensity-Based Absolute Quantification) values, which are an accurate measure of protein abundances (Krey et al. 2014). For bioinformatic processing details, see Supplementary Methods. For dataset IDs, growth conditions, and key parameters from each study, see Supplementary Tables S4 and S5.

2.3 Penalty weight vector calculation

We calculated the expression weight for each gene in each tissue on the basis of how the expression of a reaction in a particular tissue, as measured by transcriptomic or proteomic abundance, compared with the expression of that same gene in the other tissues.

$$W_{it} = \frac{\mathrm{Max}(E_i)}{E_{it}} \quad (2)$$

where W_{it} is the expression weight for a given gene i in a tissue t, E_i is the list of expression values of gene i for each tissue, E_{it} is the expression of gene i in tissue t, and Max() is the maximum value from a set of one or more elements. Note that although the transcriptomic and proteomic datasets used in this study report absolute quantities, our method is applicable as long as relative amounts of RNA or protein across tissues are available. Many GPRs in the model consist of multiple genes that represent isozymes or members of protein complexes. The former are denoted by OR terms and the latter by AND terms in the GPR formulation. This results in many reactions having more than one expression weight due to being mapped to multiple genes. We combine these multiple weights into a single penalty weight value for each reaction by averaging the expression weights of isozymes and taking the "worst" (i.e. largest, most penalizing value) when genes form subunits of a protein complex. As an example, the penalty weight for a reaction R in the leaf subnetwork of our model with a GPR of the form (Gene1 OR Gene2) AND (Gene3), correspond-ing to a protein complex made of the product of Gene 3 and the product of Genes 1 or 2, would be represented by:

$$c_{R;lf} = SF \cdot \mathrm{Max}\left(\frac{W_{gene1;lf} + W_{gene2;lf}}{2}; W_{gene3;lf}\right) \cdot 1 + 1 \quad (3)$$

where $c_{R,lf}$ represents the overall penalty weight in the leaf (lf) for reaction R, SF (or the scaling factor) is a coefficient that modulates the magnitude of the calculated penalty weights and $W_{gene1,lf}$, $W_{gene2,lf}$, and $W_{gene3,lf}$ are the penalty weights for the individual genes Gene1, Gene2, and Gene3. Note that in the present implementation of this method, stoichiometric coefficients in GPR terms are ignored. When one or more genes contained in a GPR for a reaction/tissue combination are all more highly expressed than the same genes in the other tissues, the scale for that reaction/tissue combination will be 1. For reaction/tissue combinations that have no corresponding GPR, we explored setting the penalty weights to 1 or a value calculated from the median penalty weight assigned to reactions in the same tissue (for details, see Supplementary Methods).

2.4 Optimization

The optimization performed in this paper is a variation on pFBA, which finds the flux map(s) satisfying imposed constraints with minimum total flux through the network (Lewis et al. 2010). The minimization of total flux (Equation 1) is subject to the following constraints:

$$Sv = 0 \quad (4)$$

$$LB_j \le v_j \le UB_j \quad (5)$$

$$v_{\mathrm{biomass}(\mathrm{tissue})} = v_{\mathrm{fixed\ biomass}_{\mathrm{tissue}}} \quad (6)$$

Where S is the stoichiometric matrix of the metabolic network being modeled, v is the vector of all fluxes, LB and UB are the vectors of all upper and lower bound constraints, and $v_{\mathrm{biomass(tissue)}}$ and $v_{\mathrm{fixed\ biomass(tissue)}}$ are the biomass flux for a given tissue and the defined biomass constraint for that tissue, respectively. Equation (4) represents the steady state of all internal metabolites, Equation (5) represents the bounds and reversibility constraints, and Equation (6) represents the definition of biomass accumulation rates. All optimizations were done in the COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox in MATLAB (Heirendt et al. 2019) using the Gurobi^{TM} optimizer version 8.1.1 (Gurobi Optimization, LLC 2019).

2.5 Error evaluation

We assume that the ^{13}C-MFA fluxes reported in Ma et al. (2014) are the true in vivo metabolic fluxes and therefore regard the discrepancy between FBA-predicted fluxes and these ^{13}C-MFA fluxes as a measure of error. Biomass accumulation (i.e. the difference in dry weight between a timepoint t and another timepoint t_{+1}) was not reported in Ma et al. (2014), but is the basis for the flux through the biomass equation in FBA. To allow a comparison between our FBA-predicted fluxes and the MFA-estimated fluxes in Ma et al. (2014), we set an arbitrary biomass flux of 0.01 g/h through the leaf, stem, and root biomass reactions in both the day and night, similar to the approach taken in de Oliveira Dal'Molin et al. (2015). We then normalized our fluxes by multiplying them by the ratio of the measured leaf CO_2 uptake from Ma et al. (2014) and the net leaf CO_2 uptake in our FBA flux map. A weighted average error for each FBA-predicted flux map was obtained using the following expression:

$$X_{j2\mathrm{Measured}} \left(\frac{|v_j^p - v_j^m|}{v_j^m}\right) P_{j2\mathrm{Measured}} \frac{v_j^m}{v_j^m} A \quad (7)$$

where v_j^p and v_j^m are the FBA-predicted and MFA-estimated fluxes of a flux j and A is the normalization factor previously described. We calculated weighted average errors rather than just average errors because small absolute differences between FBA-predicted and MFA-estimated flux values can correspond to extremely large % error values when the MFA-estimated fluxes are small. We quantified the maximum/minimum weighted average errors of each flux map using flux variability analysis (FVA; Mahadevan and Schilling 2003). Additional details can be found in the Supplementary Methods.

3 Results

3.1 The application of gene expression penalty weights reliably reduces discrepancies between FBA-predicted and MFA-estimated fluxes

Predicted flux maps were generated for a multi-tissue diel model of A.thaliana's central metabolism using FBA in which the sum of all the metabolic and transport fluxes required for steady-state growth is minimized, with each flux being multiplied by a penalty weight that was derived from the relative expression of the gene(s) involved in conducting that flux (see Section 2). Penalty weights for each reaction were calculated from RNA-seq (Klepikova et al. 2016; Mergner et al. 2020) and proteomic (Mergner et al. 2020) datasets using the relative expression of each gene in the different tissues. The weighted average % error between these flux maps and ^{13}C-MFA estimates from Ma et al. (2014) were used to quantify the

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

Table 1. Weighted average % error values calculated from weighted versus unweighted flux maps for transcriptomic and proteomic datasets from Klepikova et al. (2016) and Mergner et al. (2020).[a]

Dataset	Light level	Weighted average error (%)	
		No gene expression weights	With penalty weights
Mergner et al. Transcriptome	High	169–180	14.7–17.1
	Low	93.8–103	14.9–18.1
Mergner et al. Proteome	High	169–180	10.9–13.4
	Low	93.8–103	8.74–10.9
Klepikova et al. Transcriptome	High	169–180	14.8–17.4
	Low	93.8–103	19.3–21.7

[a]Values represent the lowest and highest possible error values given the results of FVA. Weighted average error values were calculated from flux maps generated using a scaling factor of 1.

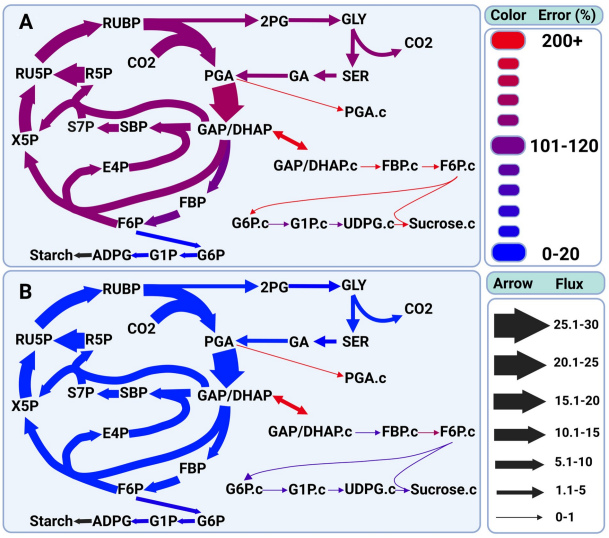

Figure 1 Percent errors of specific reactions in central metabolism before (A) and after (B) gene expression weight application. The error values in (A) are the lowest possible given FVA results and the values in (B) are the highest possible given FVA results. We see substantial decreases in errors associated with central carbon assimilation, as well as starch and sucrose synthesis. Since the ^{13}C-MFA estimated fluxes from Ma et al. (2014) do not feature the flux from ADPG to Starch, this flux lacks an estimated error and is therefore shown in black. Flux values are relative to the lowest flux in the network.

accuracy of these FBA predictions, as compared with the accuracy of flux maps generated by pFBA (Lewis et al. 2010) alone. The flux maps arrived at after the application of either transcriptomic or proteomic penalty weights show greater agreement, as measured by the weighted average % error, with ^{13}C-MFA estimates than the results from pFBA alone (Table 1). These reductions in error are substantial and statistically significant at a ¼ .01; they are consistent across comparisons against two different flux maps (high- and low-light conditions) and are sustained across a range of assumed ratios of starch to sucrose production and carboxylase to oxygenase fluxes through RuBisCO (vo/vc). Marked reductions in error are seen whether one uses the transcriptomic or proteomic tissue-atlas datasets from Mergner et al. (2020) or the transcriptomic dataset from Klepikova et al. (2016), so that the improvement in flux predictions is not dependent on the values obtained in a specific gene expression dataset or type.

We wanted to confirm that these reductions in error are in fact dependent on penalty weights calculated from gene expression data and not an artifact of the weighting procedure itself. Indeed, previous studies have used the application of randomized weights as a method of exploring different possible flux modes in a plant metabolic network (Cheung et al. 2015). We found that substituting the leaf for the root proteomic dataset, and vice-versa, resulted in no reduction in weighted average error (Supplemental Table S1) compared with pFBA. Neither did randomization of the penalty weight vector and subsequent optimization. The mean of the weighted average errors of 50 high-light condition flux maps generated with independent randomized penalty weight vectors at a scaling factor of 1 was 201%, versus the unweighted error value of 169%–180% for that condition.

3.2 Increases in agreement between FBA-predicted and MFA-estimated fluxes are broadly distributed across central metabolism

Although there is variation among individual fluxes in the degree to which omic data integration improves agreement between predicted and experimentally derived values, the reduction in weighted error as a result of penalty weight application is distributed broadly across the fluxes for which ^{13}C-MFA estimates are available. If, for example, the improvement were due to a substantial decrease in one or a small number of high-flux reactions and a negligible decrease or even increase in error for other reactions (Fig. 1) the overall finding would be less striking and potentially less broadly applicable. The reductions in error are consistent not only across metabolic subsystems within a single FBA flux map, but also across alternative stoichiometric network structures. Initial pFBA-derived solutions for a model identical to that used to generate the other predictions except with unconstrained uptake and discharge of protons from root tissue show similar reductions in error (Supplementary Table S2). Upon application of penalty weights, this model converges to a similar value of weighted average error and linear correlation as other model configurations.

3.3 Error reductions are a function of the scaling factor parameter and are improved by the application of a tissue-specific median weight for reactions lacking gene protein reaction terms

The magnitude of the penalty weights calculated and applied by the present method depends on the magnitude of the scaling factor term, (Equation 3). The increased agreement between the FBA-predicted and MFA-estimated flux maps only manifests in the majority of cases for scaling factors of 0.05–0.1 or greater (Fig. 2). We also note that the relationship between the scaling factor value and the improved agreement is monotonic—i.e. we do not see erratic increases and decreases as we increase the scaling factor value and, by extension, the strength of the assumed relationship between flux and gene expression. The necessity of a non-negligible scaling factor, the consistency of error improvement as the scaling factor is increased, and the similarity in the pattern of error improvement across multiple datasets as seen in Fig. 2, all suggest that real biological signal related to the partitioning of metabolic activity across the plant's tissues is being extracted from the gene expression datasets. Finally, we observe that the flux maps generated using penalty weight derived from Mergner et al. (2020) proteomic dataset have noticeably better weighted average errors than flux maps generated using transcriptomic dataset (Table 1 and Fig. 2). This is consistent with the closer relationship between measured protein levels and metabolic fluxes than between transcripts and fluxes. It is also consistent with at least one other study's attempts at integrating gene expression data into FBA in E.coli (Tian and Reed 2018).

Although the method presented does not involve fitting the Scaling Factor parameter using goodness-of-fit to the 13C-MFA fluxes, in Fig. 1 and Tables 1 and 2, we show results from a Scaling Factor of 1 because it falls in the plateau of low average error values we see in Fig. 2. To further explore the usefulness of a Scaling Factor of 1, we used the fluxes reported in Szecowka et al. (2013)

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

for illuminated A.thaliana leaves estimated by kinetic flux profiling. The FBA-derived flux map generated using vo/vc and starch: sucrose synthesis constraints from that study without any omic weighting has a weighted average error of 108%; this error drops to 8, 6, and 9% when protein or transcript weights from Mergner et al. (2020) or transcript weights from Klepikova et al. (2016), respectively, are applied with a Scaling Factor of 1 (Supplementary Table S6 and Dataset S5).

In our initial formulation of the algorithm for generating gene expression-derived penalty weights, the weight of all reactions with no associated GPR was set to 1, since this is the implicit value of the coefficient for all reactions in a standard pFBA optimization. Since this runs the risk of introducing a systematic bias against using reactions that have associated GPRs, we attempted to counteract this risk by assigning all reactions lacking a GPR a penalty weight corresponding to the median penalty weight of all weighted reactions in the tissue in which those reactions are found. Comparing the results with and without the tissue-specific median penalty weights for reactions without GPRs, we see modest improvements in the weighted average errors from a scaling factor of 1 onwards when using the transcriptomic and proteomic datasets from Mergner et al. (2020; Fig. 2), though the effect is not large, indicating that our method is robust to including or omitting the tissue-specific median weight correction.

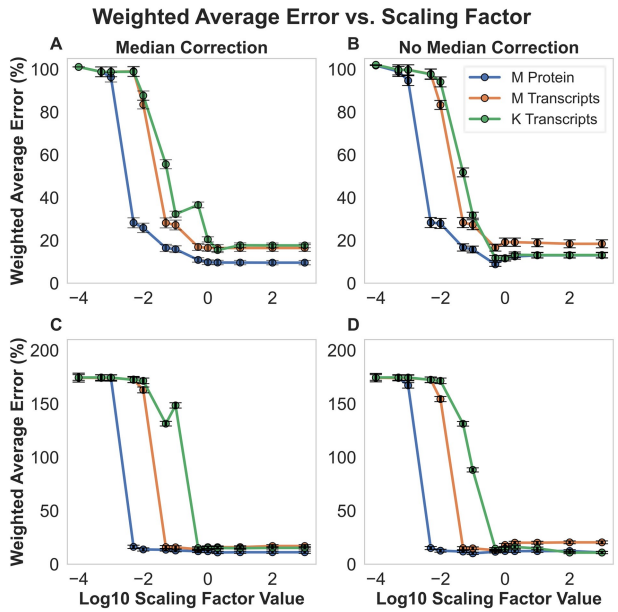

Figure 2 Weighted average errors of FBA predictions compared with MFA-estimated flux maps as a function of scaling factor value, light-level, and application of a tissue-specific median weight correction. Panels show weighted average errors of flux maps generated using (A) low-light constraints and a tissue-specific median correction applied, (B) low-light constraints and without a tissue-specific median correction applied, (C) high-light constraints and with a tissue-specific median correction applied, and (D) high-light constraints and without a tissue-specific median correction applied. "M Protein" and "M Transcripts" refer to flux maps generated using proteomic- and RNA-seq-derived weights from Mergner et al. (2020). "K Transcripts" refers to flux maps generated using RNA-seq derived weights from Klepikova et al. (2016). Upper and lower bars on each point represent the highest and lowest possible weighted average errors given FVA results, and the points themselves represent the average of these values.

3.4 Changes in the carbon and energy economy upon application of gene expression weights

In addition to improving quantitative agreement between the FBA-predicted and MFA-estimated flux maps, the gene expression weighting procedure also generates flux maps that present a substantially different picture of carbon and energy metabolism in Arabidopsis leaves.

In both high and low light FBA-predicted fluxes there is a substantial decrease in leaf mitochondrial electron transport chain (ETC) activity and overall flux in mitochondria-localized reactions in the light relative to nighttime ETC activity and overall flux (Supplementary Table S3). MFA and other recent work further points to low TCA cycle fluxes in photosynthesizing leaves (Tcherkez et al. 2005; Xu et al. 2021; 2022). This decrease in mitochondrial activity goes hand-in-hand with a predicted decrease in the use of unusually high fluxes related to proline metabolism to indirectly support the consumption of excess reductant produced via the light reactions of photosynthesis. Alongside this decrease in mitochondrial activity is a decrease in the ratio of cyclic electron flow (CEF) to linear electron flow (LEF) in the chloroplast (Table 2). Although reliable empirical measurements of this CEF/LEF ratio are difficult to obtain, previous studies have shown that a C3 plant like Arabidopsis relying on CEF to bring the ratio of ATP/NADPH produced up to that needed for normal growth would have a CEF amounting to 13% of LEF (Kramer et al. 2004). Due to the presence of other balancing mechanisms, such as the malate valve (Selinski and Scheibe 2019), this 13% value would represent an upper bound on stoichiometrically predicted values for CEF/LEF. Application of gene expression data decreases the CEF/LEF ratios in all but one FBA-predicted flux map to values much closer to the expected 13% upper bound than are predicted using conventional pFBA (Table 2).

Ma et al. (2014) reported MFA-derived estimates of %vpr, or the rate of photorespiratory CO_2 release via glyoxylate decarboxylation as a % of CO_2 assimilation, as well as the ratio of RuBisCO

Table 2. Measures of carbon and energy utilization derived from the predicted flux maps with and without penalty weights applied.

Dataset used for weighting	Light	RuBisCO flux net CO_2 assimilation	Photorespiratory CO_2 loss/net CO_2 assimilation (%)	CEF/LEF (%)	% of leaf daytime CO_2 assimilation going to biomass
None	High	2.86	62	24	43
	Low	1.85	26	31	54
Mergner et al. Protein	High	1.29	26	20	18
	Low	1.17	14	15	26
Mergner et al. Transcripts	High	1.20	25	21	18
	Low	1.15	14	27	33
Klepikova et al. Transcripts	High	1.30	27	17	19
	Low	1.25	15	14	31
Reference values	High	1.28[a]	28[a]	13[b]	56%[c]
	Low	1.17[a]	16[a]		

[a]Ma et al. (2014).
[b]Kramer et al. (2004).
[c]Weraduwage et al. (2015). The superscript letters b and c reference values are not associated with a particular light level.

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

carboxylation flux to net CO_2 assimilation in the leaf. The unweighted flux predictions for the high and low light conditions disagree substantially with these estimates (Table 2). However, the application of gene expression weights consistently brings estimates of these parameters into close agreement with MFA-derived values. The integration of gene expression also changes the predicted efficiency with which Arabidopsis converts atmospheric CO_2 into biomass (Table 2). For comparison with these predicted efficiencies, we used the empirical A.thaliana biomass, leaf area, and gas exchange data reported in Weraduwage et al. (2015) to calculate that 56% of the net CO_2 assimilation in illuminated leaves ends up in incorporated into biomass, which is closer to the value in our unweighted flux predictions than our weighted flux predictions, although it should be noted that these data were gathered from a hydroponic system.

4 Discussion

^{13}C-MFA is broadly accepted as being the most reliable method for estimating metabolic flux maps in vivo due to its ability to make use of substantial amounts of isotopic labeling data to arrive at well-supported flux maps in small- to medium-scale networks (Antoniewicz 2015). However, the technique's utility is limited by the substantial experimental effort that goes into the generation of each individual flux map. FBA, with its requirement of much less experimental data, has become the method of choice for more exploratory or predictive metabolic modeling studies. The implicit assumption is usually that the predictions of FBA—or at least the range of its predictions in cases where a unique solution is not provided—agree with those we would arrive at if we were able to conduct a ^{13}C-MFA study. This makes our optimization procedures when performing FBA and validation of FBA models against MFA results of vital importance. The method presented here, by bringing FBA-predicted fluxes into line with MFA estimates represents a step in the direction of higher-confidence FBA flux maps.

One limitation, as well as motivation, for this study is the lack of a large set of ^{13}C-MFA datasets in plants and other multi-tissue eukaryotic systems. Systems like E.coli have multiomic datasets consisting of transcriptomic, proteomic, and fluxomic measurements (Ishii et al. 2007) that have been utilized to empirically infer the relationship between gene expression and metabolic fluxes. This empirical training can then be used to more accurately predict fluxes in new contexts (Tian and Reed 2018). The sparsity of ^{13}C-MFA data in more complex systems makes such an approach currently impossible.

A noteworthy theoretical aspect of the present approach is its simplicity, the only variable parameter being a single scaling factor that controls the magnitude of the penalty weights. That the assumption of a consistent value relating the relative abundances of transcripts or proteins in different tissues to the "preference" of an organism to partition flux among particular reactions can result in substantial improvement in error was of great interest in light of the complexity of the relationship between measures of gene expression—transcriptomic and proteomic abundances—and flux. Particularly when making biotechnological interventions in a system to modify its metabolism, there is often an assumed strong linear relationship between transcription, translation, and, ultimately, metabolic flux, but the reality is rarely so simple. Although moderate correlations between transcript and protein abundances have been demonstrated across many systems, the degree of correlation varies across systems and experimental contexts (Maier et al. 2009; Liu et al. 2016). The correlation between these data types and rates of central metabolic reactions, which carry the large majority of total metabolic flux, is weaker still (Kuile and Westerhoff 2001). Some previous studies found that changes in the gene expression related to individual reactions typically do not correlate well with changes in fluxes (Schwender et al. 2014; Tian and Reed 2018), with some central metabolic fluxes in particular showing a negative correlation between changes in gene expression and flux. In both cases, gene expression data related to reactions were compared within the same cell type or tissue; in our study, we instead compare intertissue abundances, mirroring the long-standing practice in the literature of inferring relative metabolic activity in different tissues by their transcript and protein investment in relevant pathway steps. It may be that only by considering gene expression on an intertissue basis in the context of the entire complex stoichiometric network underlying metabolism can predictive gains from including gene expression evidence be properly realized.

Future work should aim to expand the number of available datasets, and the experimental conditions and genotypes for which they are gathered, in order to enable more thorough evaluation of methods like the one presented in this article. Indeed, evaluating the presented method requires ^{13}C-MFA fluxes, multi-tissue omic data, and a GEM all for the same biological system, which, to our knowledge, is currently only available for A.thaliana. Building on the work of Ma et al. (2014), experimental improvements and refinements of the underlying network architecture of central carbon metabolism have been introduced in the context of ^{13}C-MFA in Camelina sativa (Xu et al. 2021; 2022) and Nicotiana tabacum (Chu et al. 2022). In this study, Ma et al. (2014) flux maps are used without change and we adopted a highly curated A.thaliana GEM from which to construct the whole-plant model. This approach precluded the possibility of our reanalyzing the MFA-estimated flux map or biasing the construction of a purpose-built GEM, making the MFA-to-FBA comparison more favorable. However, in future studies, a combination of MFA network refinements, expanded datasets, and further improvements in the flux estimation procedures holds promise for improving the fidelity of the ^{13}C-MFA comparison data. On the FBA side, the use of more detailed growth and composition measurements for FBA along with more detailed representation of different tissue types will potentially allow for more biologically accurate and representative FBA flux map predictions. These improvements in both MFA-estimation and FBA-prediction of flux maps, along with an expansion in the number of available ^{13}C-MFA datasets against which to compare FBA predictions, will allow for more extensive validation of the method described in this paper as well as other methods aiming to incorporate omic datasets into flux prediction.

A distinct aspect of the proposed method is its demonstrated ability to bring FBA-predicted fluxes in line with MFA-estimated fluxes across multiple input datasets, model architectures, and using multiple independent gene expression datasets. Our hope is that methods for incorporating transcriptomic and proteomic data may advance this field to the point where FBA-predicted flux maps can be used with high confidence for practical engineering goals. This, combined with the automated reconstruction of GEMs from genom-ic and biochemical databases (Saha et al. 2014) suggests a future with rapid turnaround from the initial identification of an organism of interest to metabolic flux predictions and rational genetic engineering to achieve biotechnological aims.

Acknowledgements

We would like to thank Dr Doug Allen for permission to adapt Fig. 3 from Ma et al. (2014) for use in Fig. 1 in this publication. Figures 1 and 3 were created with BioRender.com.

Supplementary data

Supplementary data is available at Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by the Office of Science (BER), U.S. Department of Energy [grant number DE-SC0018269 to J.A.M.K. and Y.S-H.]. This work was supported, in part, by the NSF Research Traineeship Program [grant

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023

number DGE-1828149 to J.A.M.K.]. This publication was also made possible by a predoctoral training award to J.A.M.K. from [grant number. T32-GM110523] from National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

References

Antoniewicz MR. Methods and advances in metabolic flux analysis: a mini-review. J Ind Microbiol Biotechnol 2015;42:317–25.

Arnold A, Nikoloski Z. Bottom-up metabolic reconstruction of Arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiol 2014;165:1380–91.

Boyle NR, Sengupta N, Morgan JA et al. Metabolic flux analysis of heterotrophic growth in Chlamydomonas reinhardtii. PLoS One 2017;12:e0177292.

Burgard AP, Pharkya P, Maranas CD et al. OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 2003;84:647–57.

Chen X, Alonso AP, Allen DK et al. Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. Metab Eng 2011;13:38–48.

Cheung CYM, Williams TCR, Poolman MG et al. A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. Plant J 2013;75:1050–61.

Cheung CYM, Ratcliffe RG, Sweetlove LJ et al. A method of accounting for enzyme costs in flux balance analysis reveals alternative pathways and metabolite stores in an illuminated Arabidopsis leaf. Plant Physiol 2015;169:1671–82.

Chu KL, Koley S, Jenkins LM et al. Metabolic flux analysis of the non-transitory starch tradeoff for lipid production in mature tobacco leaves. Metab Eng 2022;69:231–48.

de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW et al. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiol 2010a;152:579–89.

de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW et al. C4GEM, a genome-scale metabolic model to study C4 plant metabolism. Plant Physiol 2010b;154:1871–85.

de Oliveira Dal'Molin CG, Quek LE, Saa PA et al. A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Front Plant Sci 2015;6:1–12.

Gleizer S, Ben-Nissan R, Bar-On YM et al. Conversion of Escherichia coli to generate all biomass carbon from CO_2. Cell 2019;179:1255–63.e12.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual Version 8.1, 2019. URL: https://www.gurobi.com/documentation/8.1/refman/index.html (July 2022, date last accessed).

Hay J, Schwender J. Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (Brassica napus L.) embryos. Plant J 2011;67:526–41.

Heirendt L, Arreckx S, Pfau T et al. Creation and analysis of biochemical constraint-based models using the COBRA toolbox v.3.0. Nat Protoc 2019;14:639–702.

Ishii N, Nakahigashi K, Baba T et al. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 2007;316:593–7.

Jenior ML Moutinho TJ Jr, Dougherty BV et al. Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. PLoS Comput Biol 2020;16, e1007099.

Klepikova AV, Kasianov AS, Gerasimov ES et al. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J 2016;88:1058–70.

Kramer DM, Avenson TJ, Edwards GE et al. Dynamic flexibility in the light reactions of photosynthesis governed by both electron and proton transfer reactions. Trends Plant Sci 2004;9:349–57.

Krey JF, Wilmarth PA, Shin J-B et al. Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. J Proteome Res 2014;13:1034–44.

Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS Lett 2001;500:169–71.

Lee KH, Park JH, Kim TY et al. Systems metabolic engineering of Escherichia coli for L-threonine production. Mol Syst Biol 2007;3:149.

Lewis NE, Hixson KK, Conrad TM et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 2010;6:390.

Liu Y, Beyer A, Aebersold R et al. On the dependency of cellular protein levels on mRNA abundance. Cell 2016;165:535–50.

Ma F, Jazmin LJ, Young JD et al. Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proc Natl Acad Sci USA 2014;111:16967–72.

Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol 2014;10:e1003580.

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

Maier T, Güell M, Serrano L et al. Correlation of mRNA and protein in complex biological samples. FEBS Lett 2009;583:3966–73.

Mergner J, Frejno M, List M et al. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 2020;579:409–14.

Orth JD, Thiele I, Palsson BØ et al. What is flux balance analysis? Nat Biotechnol 2010;28:245–8.

Pandey V, Hadadi N, Hatzimanikatis V. Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. PLoS Comput Biol 2019;15:1–23.

Park JH, Lee KH, Kim TY et al. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. Proc Natl Acad Sci USA 2007;104:7797–802.

Poolman MG, Miguet L, Sweetlove LJ et al. A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol 2009;151:1570–81.

Ratcliffe RG, Shachar-Hill Y. Measuring multiple fluxes through plant metabolic networks. Plant J 2006;45:490–511.

Ravi S, Gunawan R. DFBA—predicting metabolic flux alterations using genome-scale metabolic models and differential transcriptomic data. PLoS Comput Biol 2021;17:e1009589.

Saha R, Chowdhury A, Maranas CD et al. Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotechnol 2014;29:39–45.

Saha R, Suthers PF, Maranas CD et al. Zea mays irs1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS One 2011;6:e21784.

Schwender J, König C, Klapperstück M et al. Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Front Plant Sci 2014;5:1–16.

Selinski J, Scheibe R. Malate valves: old shuttles with new perspectives. Plant Biol J 2019;21:21–30.

Shaw R, Cheung CYM. Multi-tissue to whole plant metabolic modelling. Cell Mol Life Sci 2020;77:489–95.

Szecowka M, Heise R, Tohge T et al. Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell 2013;25:694–714.

Tcherkez G, Cornic G, Bligny R et al. In vivo respiratory metabolism of illuminated leaves. Plant Physiol 2005;138:1596–606.

Tepper N, Shlomi T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 2010;26:536–43.

Tian M, Reed JL. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics 2018;34:3882–8.

Vijayakumar S, Conway M, Lió P et al. Seeing the wood for the trees: a Forest of methods for optimization and omic-network integration in metabolic modelling. Brief Bioinform 2017;19:1218–35.

Weraduwage SM, Chen J, Anozie FC et al. The relationship between leaf area growth and biomass accumulation in Arabidopsis thaliana. Front Plant Sci 2015;6:1–21.

Williams TCR, Poolman MG, Howden AJM et al. A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. Plant Physiol 2010;154:311–23.

Xu Y, Wieloch T, Kaste JAM et al. Reimport of carbon from cytosolic and vacuolar sugar pools into the Calvin-Benson cycle explains photosynthesis labeling anomalies. Proc Natl Acad Sci USA 2022;119:e2121531119.

Xu Y et al. The metabolic origins of non-photorespiratory CO2 release during photosynthesis: a metabolic flux analysis. Plant Physiol 2021;186:297–314.

Downloaded from https://academic.oup.com/bioinformatics/article/39/5/btad186/7114031 by guest on 19 July 2023