Supporting Information

Mintz-Oron et al. 10.1073/pnas.1100358109

SI Text

Network Reconstruction Using Constraint-Based Modeling. We have recently developed a computational model building algorithm (MBA) that enables automatic reconstruction of tissue-specific metabolic network models by integrating a generic model with tissue-specific gene expression and proteomic data. The method was applied to reconstruct a model of liver metabolism in human that was shown to be able to improve upon metabolic flux predictions in liver in comparison with a generic human model (1). The implementation of the MBA is available upon request. The algorithm derives a consistent metabolic model from a generic model, based on network integration with various data sources. First, the method works by assembling an initial core set of reactions. Next, it uses a greedy heuristic search that is based on iteratively pruning reactions from the generic model, taken from an elimination list created in a random order, yet maintains the consistency of the pruned model. In each pruning step, a reaction is removed only if its removal does not prevent the activation of core reactions. Thus, a minimal set of reactions from a generic model that are needed to activate the reactions associated with this initial core set is added, obtaining a model that is consistent. That is, each of the core reactions can potentially carry a nonzero metabolic flux, within a global flux distribution, satisfying stoichiometric, mass-balance and reaction directionality constraints. To obtain priorities, we demand that reactions with low priority will appear earlier in the reaction elimination list, and thus are attempted to be eliminated before higher priority reactions. Using this scheme, for the global model reconstruction, (*i*) relaxation of *Arabidopsis* reactions directionality was prioritized over the addition of plant reactions (by assigning the plant reactions earlier than the reversed *Arabidopsis* reactions in the elimination list), and (*ii*) addition of plant reactions was prioritized over the addition of nonplant reactions (by assigning the nonplant reactions earlier than the plant reactions in the elimination list). For the compartmentalized and tissue-specific models, elimination of transport reactions was prioritized over elimination of enzymatic reactions (by assigning the transport reaction earlier than the enzymatic reactions in the elimination list).

As the resulting model depends on the chosen reaction scanning order, the algorithm is executed repeatedly for a number of times (1,000 in the results presented here because of observed convergence following this number of iterations) with different, random scanning orders. Each run results in a candidate model. All 1,000 candidate models are then processed to assign the noncore reactions with scores, representing the fraction of candidate models in which they appear. An aggregative model is built by considering the scores across all runs, starting with the core reactions and incrementally adding reactions according to their confidence score until a consistent, viable model is obtained.

Our model reconstruction approach does not assume an objective function, but rather aims to identify gap-filling reactions that would enable to activate a set of known core reactions extracted from KEGG (Kyoto Encyclopedia of Genes and Genomes) and Aracyc. As part of the definition of the reaction core, we include a production reaction for each biomass constituent. The inclusion of the biomass production reactions in the core results in a final model, which enables the production of all biomass compounds. We do not assume that biomass production rate is maximized in any step of the model reconstruction method.

Global Model Reconstruction. Here, the MBA algorithm is used to address the problem of gap-filling. The core reactions set is composed of Aracyc and KEGG-*Arabidopsis* reactions, as well as literature-based exchange reactions; the generic model is composed of the Plant Metabolic Network (PMN), and KEGG non-*Arabidopsis* reactions. Our goal is then to derive the most parsimonius consistent model, which includes maximal number of the *Arabidopsis*-specific reactions, and a set of additional reactions from other organisms. Several changes were introduced to the original MBA algorithm. We allow not only the addition of generic reactions but also relaxation of irreversibility of existing core reactions, in case this relaxation leads to a larger set of activated core reactions. Second, as reactions identified in plants are more likely to occur in *Arabidopsis* than reactions indentified in more distant organisms, each generic reaction is given an elimination priority according to its source organism. As a result, the addition of a plant originated reaction is prioritized over nonplant reaction. Thus, we attempt to bridge network gaps through (*i*) relaxation of irreversibilities of the model's reactions (first priority), and (*ii*) addition of enzymatic reactions from other organisms (second priority). Finally, because *Arabidopsis* is an autotrophic organism we maintain this property by defining a consistent model to be a model that not only satisfies activation of all core reactions but also allows production of all biomass compounds under minimal media.

Compartmentalized Model Reconstruction. We use our previously developed constraint-based modeling (CBM) for systematically predicting subcellular localization of enzymes in a metabolic network, based on a priori localization data for a subset of the enzymes, relying on a parsimony principle of minimal number of cross-membrane metabolite exchange (2). The input data of this method is a metabolic network, and the known localization of a subset of the network enzymes. Known localization data were collected from the SUBA (the *Arabidopsis* Subcellular Database) database (3), allowing experimental localization assignment for 49% of the model's reactions, aiming to predict the localization of the remaining reactions. Following the integration of the experimental data in the metabolic network, the rest of the reactions are duplicated to all compartments. To narrow down the list of potential localizations of these reactions, they are duplicated only to their predicted localizations, based on 10 prediction programs—TargetP (4), MitoProt2 (5), SubLoc (6), IPSort (7), Predotar (8), MitoPred (9), PeroxiP (10), WolfPSort (11), MultiLoc (12), and LocTree (13)—in case at least half of the programs assigned it with some localization. In all other cases, the reactions were duplicated to all compartments. Transport reactions were added to the network enabling metabolite exchange between the cytoplasm and all other compartments. The application of the network localization-prediction method on the large-scale *Arabidopsis* network involved the development of a heuristic variant following a similar approach to the MBA algorithm. Specifically, the core reaction set was assembled using experimental localization, and all remaining reactions represent the generic model. To retain the parsimony principle of a minimal number of cross-membrane metabolite transporters, elimination of transport reactions was prioritized over elimination of enzymatic reactions. In this scenario, in case it is possible, a gap will be filled with an enzymatic reaction rather than by activation of a transport reaction. Similar to the global model reconstruction process, we define a consistent model to be a model which

satisfies activation of all core reactions as well as allowing production of all biomass compounds under minimal media.

Tissue-Specific Model Reconstruction. Tissue-specific models were reconstructed for 10 tissues and developing stages: juvenile leaves, open flowers, flower buds, 10-d roots, 23-d roots, siliques, seeds, cotyledons, cell cultures grown in light, and cell cultures grown in dark. AtProteome, a high-density *Arabidopsis* proteome map (14), was integrated to the compartmentalized model to reconstruct a consistent functional model for each tissue using the MBA algorithm. In the AtProteome database a protein is represented by a set of detectable peptides that unambiguously identify that protein. In our analysis, to reduce noise level, a protein is assigned to a certain tissue if at least one-third of its peptides are detected in that tissue. This threshold was determined after careful manual inspection of the dataset and represents a compromise between reducing noise level but retaining high proteomics coverage. For each tissue, a core reactions set was composed of reactions from the compartmentalized model, identified in that tissue by the proteomics data. The generic model was composed from the set of the remaining compartmentalized model reactions. The MBA algorithm is then used to find the minimal set of reactions that should be added to the core reactions set to obtain a consistent model, prioritizing enzymatic reactions over transport reactions.

Models' Annotation. In this study we adapt an accepted standard for annotating metabolic network models, MIRIAM (minimum quality standard) (15), and provide the relevant annotation in Dataset S1, as well as within the System Biology Markup Language (16) files, enabling model use in common CBM software packages [e.g., Cobra Toolbox (17), OptFlux (18)]. The annotation tables include common identifiers for reactions and metabolites, metabolites formula, biological description of reactions and metabolites, reaction stoichiometry, reaction–gene associations, reaction EC value, metabolic pathway annotations of reactions, and confidence scores for reactions. The reported confidence values clearly discriminate novel predictions from experimental data. Toward this goal, we define for each reaction in the model a confidence score [similar to what was done in the reconstruction of the human network by Duarte et al. (19)]: (*i*) "inclusion confidence score" is denoted for each reaction in the generic model that denotes whether it was included in the core or predicted as part of the gap-filling procedure; (*ii*) "localization confidence score" is denoted for each reaction in the compartmentalized model, representing whether its subcellular localization is supported by experiments or computational prediction; and (*iii*) "tissue confidence score," representing whether its tissue assignment is supported by experimental or computational predictions.

Database Mapping. Mapping between Aracyc, PMN, and KEGG database compounds was done by matching of the following features: (*i*) compound name and synonym, (*ii*) chemical formula, and (*iii*) compound mapping to other databases (CAS, KNAPSACK, PubChem). Mapping between reactions was performed by matching of the following features: (*i*) reaction name and synonym, (*ii*) substrates and products participating in the reaction, (*iii*) reaction's EC number, and (*iv*) genes catalyzing the reaction. The resulting process can be defined as semi-automatic, as the extraction and comparison of data between databases can be largely automated. However, manual curation, in addition to comparison with Radrich et al. (20), who performed a similar procedure, was also necessary to solve discrepancies between them. The mapping process resulted in 71% and 55% mapped metabolites and reactions, respectively.

Cross-Validation Tests. To assess the performance of different approaches used in the reconstruction process, in each stage we applied a standard cross-validation test, repeating the relevant model reconstruction given only four-fifths of the data, aiming to predict the missing one-fifth. Such analysis tests the ability of the method to correctly predict the missing, left-out reactions. More specifically, various random subsets of the full core reaction sets were given as input, and the reaction-content of the model was then compared with the left-out core reactions. Hence, for the global model reconstruction, four-fifths of the known *Arabidopsis* reactions were given, aiming to predict the missing one-fifth. For the compartmentalized model reconstruction, four-fifths of the known enzyme localizations were given, aiming to predict the held-out localizations. Finally, for the tissue-model reconstruction step, four-fifths of each tissue proteomics data were given, aiming to predict the held-out proteomics data. Hyper-geometric *P* value, reflecting the enrichment of the tested reactions in the model, was computed, as well as precision and recall values for each stage, testifying for the predictive performance of the reconstruction approaches.

Blast Validation. To evaluate the existence of the added reactions set in the global reconstructed *Arabidopsis* network we perform a Blast search. Two lists of EC numbers are run against the *Arabidopsis* whole genome: (*i*) EC numbers representing reactions added to the cellular model by our gap-filling method, and (*ii*) EC numbers representing reactions not added to the global model. Each EC is represented by all its ORF sequences from all available organisms. Accordingly, two e-value lists are obtained, based on the best e-values obtained for each EC run. Wilcoxon rank sum test was then applied to evaluate the significance of the results.

Simulation of Known Metabolic Functions. To validate the generic model reconstruction, we followed the quality-assurance method presented in the reconstruction of the human metabolic network model by Duarte et al. (19). More specifically, we simulated 176 known metabolic functions found in various cell and tissue types. These simulations involved the search for a feasible flux distribution that produces a certain metabolite of interest, but allowed recycled cofactor pairs to enter and leave the system as needed. Only 2% of the simulations failed and were used to manually identify several missing reactions required to amend the model. The list of simulations is provided in Table S3.

Global View of the Models Generated for Different *Arabidopsis* Tissues. The network models derived for the various tissues and cell cultures can be used to examine their metabolic similarity. Toward this goal, we clustered the tissues and cultures based on the pathway annotation of their model's reaction content, and compared the resulting clustering to that obtained by using solely the proteomic data used as input for the network reconstruction procedure (Fig. S1). The results based on the two approaches showed an expected clustering between the two root tissues and the two flower tissues. However, the models-based clustering clearly separated the "dark" tissues and cultures (cell culture-dark, roots, and seeds) from the "light" (cell culture-light, flowers, leaves, and siliques), a separation that is not evident in the proteomic data-based clustering, also reported in ref. 14.

Model-Based Computational Design of Metabolic Engineering Strategies for Vitamin E Overproduction in Seeds. Here, we applied computational methods for predicting gene knockouts that are likely to increase tocopherols (vitamin E) content in *Arabidopsis* seeds, essential components of the human diet. More specifically, the reconstructed seed model was applied to predict reactions, the deletion of which caused accumulation of tocopherol. WT flux distribution was generated as described in predicting flux mea-

surements (main text), with the addition of a positive flux constraint, representing 1% of carbon accumulated in biomass in the Tocopherol O-methyltransferase reaction forcing tocopherol production. Then, the flux through each enzymatic reaction in turn was constrained to zero and a feasible flux distribution that undergoes a minimal redistribution with respect to the flux configuration of the WT was searched using the minimization of metabolic adjustment (MOMA) method. To assess robustness of our predictions several tests are performed.

Robustness to experimental error in flux measurements resulting in alternative possible WT flux distributions. The experimental error is taken into account based on the SD of the measured fluxes (21). Here, we used a variant of Flux Variability Analysis (22) to obtain a set of $2{\cdot}K$ alternative flux distributions for the WT (where K is the number of experimentally measured fluxes), that are both consistent with the measured flux rate for central metabolism and have minimal/maximal possible rates for each measured reaction in turn, in accordance with the SD in the pertaining measurements (the total sum of absolute flux through all other reactions in the model is further minimized as in ref. 23). Then, MOMA was applied to predict flux reroutes (including tocopherol production rate) following the knockout of each gene in the model, starting from each of the obtained flux distributions for the WT computed here. We now report only these gene knockouts showing robust result of increased tocopherol production for at least a half of the computed WT flux distributions.

Robustness to additional possible variation in the WT flux distribution. We apply a Flux Variability Analysis to obtain an additional set of $2{\cdot}M$ alternative flux distributions for the wild-type (where M is the number of model reactions), that is both consistent with the measured values for central metabolism and have extreme (maximal or minimal) rates in each reaction in the model in turn (with the total sum of flux through the remaining reactions minimized). Again, we report only on these gene knockouts showing robust result for at least a half of the computed WT flux distributions.

Robustness to the choice of WT tocopherol production rate. Alternative constraints on the WT tocopherol production rates were examined, including 0.1% and 0.01% of the carbon uptake rate. We report only on gene knockouts, showing robust result for all choices of this threshold.

Following the described robustness tests, a list of 71 reactions, the knockout of which is predicted to increase tocopherol production rate by at least 25% (following ref. 24) is presented in Table S6. The predicted knockouts cluster to several metabolic pathways, some of which suggest straightforward deletion strategies that involve the direct blocking of competing pathways, and others suggest novel potential targets with more complex and nonintuitive relation to vitamin E metabolism. Indeed, flux changes in chlorophyll, and tyrosine metabolism, involving competing pathways that use tocopherol precursors homogentisate acid and phytyldiphosphate, are known to affect tocopherol accumulation (25, 26), and were accordingly predicted by the model. In other studies, *Arabidopsis* tocopherol-deficient mutants have been shown to have elevation in fatty acid biosynthesis (27) and gluthatione biosynthesis (28), further supporting our model's predictions. On top of the predictions described above, which are experimentally supported, additional knockouts in the purine and pyrimidine metabolism, lysine and leucine metabolism, TCA cycle, and zeatin metabolism are also predicted to increase tocopherol level. Further experimental validation should be performed to reaffirm the correctness of the predictions.

1. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism. *Mol Syst Biol* 6:401.
2. Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T (2009) Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics* 25:i247–i252.
3. Heazlewood JL, Verboom RE, Tonti-Filippini J, Small I, Millar AH (2007) SUBA: The *Arabidopsis* Subcellular Database. *NucleicAcids Res* 35(Database issue):D213–D218.
4. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.
5. Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* 241:779–786.
6. Chen H, Huang N, Sun Z (2006) SubLoc: A server/client suite for protein subcellular location based on SOAP. *Bioinformatics* 22:376–377.
7. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18:298–305.
8. Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4:1581–1590.
9. Guda C, Fahy E, Subramaniam S (2004) MITOPRED: A genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* 20:1785–1794.
10. Emanuelsson O, Elofsson A, von Heijne G, Cristóbal S (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. *J Mol Biol* 330:443–456.
11. Horton P, et al. (2007) et al. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res* 35(Web Server issue):W585–W587.
12. Höglund A, Dönnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: Prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22:1158–1165.
13. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol* 348:85–100.
14. Baerenfaller K, et al. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 320:938–941.
15. Le Novère N, et al. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 23:1509–1515.
16. Hucka M, et al.; SBML Forum (2003) The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524–531.
17. Becker SA, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat Protoc* 2:727–738.
18. Rocha I, et al. (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst Biol* 4:45.
19. Duarte NC, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* 104:1777–1782.
20. Radrich K, et al. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol* 4:114.
21. Lonien J, Schwender J (2009) Analysis of metabolic flux phenotypes for two *Arabidopsis* mutants with severe impairment in seed storage lipid synthesis. *Plant Physiol* 151:1617–1634.
22. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264–276.
23. Blank LM, Kuepfer L, Sauer U (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol* 6:R49.
24. Tsegaye Y, Shintani DK, DellaPenna D (2002) Overexpression of the enzyme p-hydroxyphenylpyruvate dioxygenase in *Arabidopsis* and its relation to tocopherol biosynthesis. *Plant Physiol Biochem* 40:913–920.
25. Lee K, et al. (2007) Overexpression of *Arabidopsis* homogentisate phytyltransferase or tocopherol cyclase elevates vitamin E content by increasing gamma-tocopherol level in lettuce (*Lactuca sativa* L.). *Mol Cells* 24:301–306.
26. Holländer-Czytko H, Grabowski J, Sandorf I, Weckermann K, Weiler EW (2005) Tocopherol content and activities of tyrosine aminotransferase and cystine lyase in *Arabidopsis* under stress conditions. *J Plant Physiol* 162:767–770.
27. Sattler SE, Gilliland LU, Magallanes-Lundback M, Pollard M, DellaPenna D (2004) Vitamin E is essential for seed longevity and for preventing lipid peroxidation during germination. *Plant Cell* 16:1419–1432.
28. Kanwischer M, Porfirova S, Bergmüller E, Dörmann P (2005) Alterations in tocopherol cyclase activity in transgenic and mutant plants of *Arabidopsis* affect tocopherol content, tocopherol composition, and oxidative stress. *Plant Physiol* 137:713–723.

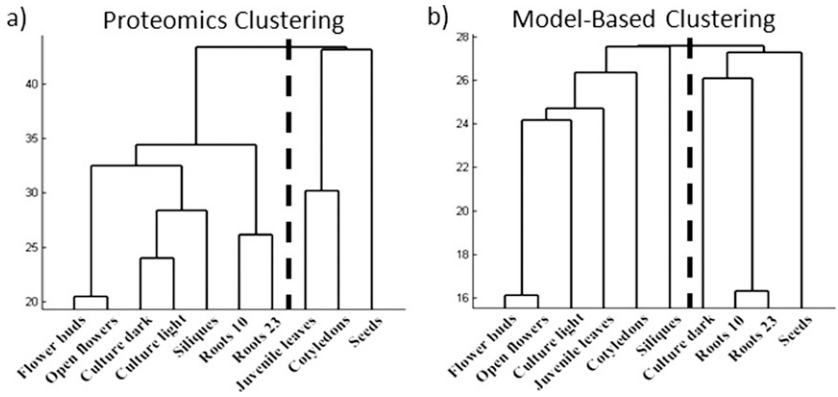

Fig. S1. Metabolic similarity between different tissues and cell cultures based on clustering of (*A*) tissue-specific proteomics data and (*B*) the reconstructed tissue and culture-specific models. Tissue model-based clustering separates the "dark" tissues and cultures from the "light," a separation that is not evident in the proteomic data-based clustering.

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)
Table S3 (DOCX)
Table S4 (DOCX)
Table S5 (DOCX)
Table S6 (DOCX)
Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)