ACCELERATED PUBLICATION

BIOTECHNOLOGY and BIOENGINEERING

Identification of Optimal Measurement Sets for Complete Flux Elucidation in Metabolic Flux Analysis Experiments

YoungJung Chang, Patrick F. Suthers, Costas D. Maranas

Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802; telephone: 814-863-9958; fax: 814-865-7846; e-mail: costas@psu.edu

Received 19 March 2008; revision received 26 March 2008; accepted 28 March 2008

Published online 15 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bit.21926

ABSTRACT: Metabolic flux analysis (MFA) methods use external flux and isotopic measurements to quantify the magnitude of metabolic flows in metabolic networks. A key question in this analysis is choosing a set of measurements that is capable of yielding a unique flux distribution (identifiability). In this article, we introduce an optimization-based framework that uses incidence structure analysis to determine the smallest (or most cost-effective) set of measurements leading to complete flux elucidation. This approach relies on an integer linear programming formulation OptMeas that allows for the measurement of external fluxes and the complete (or partial) enumeration of the isotope forms of metabolites without requiring any of these to be chosen in advance. We subsequently query and refine the measurement sets suggested by OptMeas for identifiability and optimality. OptMeas is first tested on small to medium-size demonstration examples. It is subsequently applied to a large-scale *E. coli* isotopomer mapping model with more than 17,000 isotopomers. A number of additional measurements are identified leading to maximum flux elucidation in an amorphadiene producing *E. coli* strain.

Biotechnol. Bioeng. 2008;100: 1039–1049.

KEYWORDS: metabolic flux analysis (MFA); isotopomers; incidence structure analysis; structural identifiability; integer programming; nonlinear optimization

Introduction

Metabolic fluxes are key descriptors of a cell's physiology (Nielsen, 2003) and targets of metabolic engineering for overproduction (Bailey, 1991). Metabolic flux analysis (MFA) is the gold standard method for the quantification of the fluxes (Stephanopoulos, 1999). These analysis methods infer intracellular fluxes using external flux and isotopic measurements (Sauer, 2006). The flux inference relies on a mathematical model describing the propagation of labeled atoms, for which a variety of techniques have been developed including positional enrichment (Sonntag et al., 1993), isotopomers (Schmidt et al., 1997; Zupke and Stephanopoulos, 1994), cumomers (Wiechert et al., 1999), bondomers (van Winden et al., 2002), and elementary metabolite units (EMUs) (Antoniewicz et al., 2007b).

A common practice in MFA methods is to make as many as possible measurements so as the mathematical model becomes over-determined and the fluxes can thus be reliably estimated by solving a least squares problem (Riascos et al., 2005; Yang et al., 2007). However, the nonlinearities in isotopic balances make it difficult to guarantee that even a seemingly encompassing set of measurements can uniquely determine all fluxes. Indeed, it was shown that a single labeling experiment is unlikely to uniquely determine the great majority of fluxes especially when large-scale isotope mapping reconstructions are employed (Suthers et al., 2007).

The task of analyzing the potential uniqueness of the flux distribution belongs to the class of *identifiability* problems. There have been several efforts to address this challenge including mathematical approaches for structural identifiability (Isermann and Wiechert, 2003; van Winden et al., 2001; Wiechert, 1995), and statistical approaches quantifying the confidence of flux estimate (Antoniewicz et al., 2006; Möllney et al., 1999). Alternatively, an integer programming approach (Rantanen et al., 2006) and a heuristic sequential approach (Ghosh et al., 2006) have been proposed to choose isotopic measurements for unique flux elucidation. However, current methods are not designed to choose from both external fluxes and isotopic measurements for large-scale isotopomer models while minimizing an appropriate cost function that quantifies the relative difficulty/expense of these measurements.

This experimental design task is mathematically equivalent to an NP-hard problem (Rantanen et al., 2006) due to the difficulty in ensuring identifiability in the absence of a priori knowledge on which measurements are to be made. In this article, we decide to bypass this difficulty by first

Correspondence to: C.D. Maranas

Contract grant sponsor: DOE

Contract grant number: DE-FG02-05ER25684

enforcing a relaxed identifiability condition to generate a candidate measurement set and subsequently tightening the relaxation if necessary. This relaxed identifiability condition is constructed by using *incidence structure analysis* inspired from linear systems structural controllability analysis (Lin, 1974; Shields and Pearson, 1976). We demonstrate the applicability of the analysis in choosing flux and isotopic measurements, and present the framework used for experimental designs resulting in the maximum identifiable system at the minimum relative cost.

The remainder of this article is organized as follows. In Materials and Methods Section, we define the MFA identifiability problem of interest and introduce incidence structure analysis as a way of analyzing the identifiability question. We next formulate the experimental design problem as an integer linear programming (ILP) problem OptMeas, and propose a solution procedure. In Results and Discussion Section, we demonstrate OptMeas and the solution procedure with metabolic models of increasing complexity. Finally, we summarize the main results and discuss further extensions of the current work.

Materials and Methods

Overview of Mathematical Analysis for MFA

For the systematic representation of the metabolites, isotopes and fluxes present in MFA models, we use the following sets throughout the article:

Sets:

$I = \{i\}$: metabolite pools
$I^N \subset I$: intermediate metabolites
$J = \{j\}$: unidirectional fluxes
$K_i = \{k\}$: isotopomers of metabolite $i \in I$

Each reversible reaction is split into forward and backward fluxes. We subsequently define parameters and state variables on these sets:

Parameters:

S_{ij} $(i,j) \in I^N \times J$: stoichiometry matrix

$\mathrm{IMM}^{j}_{i' \to i, k' \to k}$ $(k',k) \in K_{i'} \times K_i, (i',i,j) \in I \times I^N \times J$
: isotopomer mapping matrix (IMM)

Continuous variables:

$v_j \geq 0$ $j \in J$: flux values
$I_{ik} \in [0,1]$ $k \in K_i, i \in I$: isotopomer distribution vectors (IDVs)

Here, $S_{ij} > 0$ if flux j produces metabolite i and $S_{ij} < 0$ if j consumes i. IMM is nonnegative, and $\mathrm{IMM}^{j}_{i' \to i, k' \to k} > 0$ only if isotopomer k' of metabolite i' contributes to the formation of k of i via flux j. IMM takes fractional values for the indistinguishable isotopomers of symmetric molecules. Note that symbol I has subscripts i and k when it refers to isotopomer fractions.

The metabolic network under study is assumed to be at a steady-state yielding a number of overall metabolite and isotope balance equations. Constraint (1) imposes a flux balance for each intermediate metabolite:

$$\sum_{j \in J} S_{ij} v_j = 0, \quad i \in I^N. \tag{1}$$

Constraint (2) stipulates that the sum of the isotopomer fractions of any metabolite must be equal to 1

$$\sum_{k \in K_i} I_{ik} = 1, \quad i \in I \tag{2}$$

whereas the production and consumption terms for every isotopomer form of each intermediate metabolite must be equal to each other (see Suthers et al., 2007 for a derivation):

$$\sum_{j \mid S_{ij} > 0} \left(S_{ij} v_j \prod_{i' \in I} \sum_{k' \in K_{i'}} \mathrm{IMM}^{j}_{i' \to i, k' \to k} I_{i'k'} \right) + \sum_{j \mid S_{ij} < 0} S_{ij} v_j I_{ik} = 0, \quad k \in K_i, i \in I^N. \tag{3}$$

The product symbol Π in (3) signifies nonzero term-by-term multiplication of the isotopomer fractions that generate isotopomer k of metabolite i. With Equation (3), accounting for the presence of a stoichiometric coefficient that is equal to neither one nor zero requires the splitting of the corresponding metabolite by the introduction of "extra" metabolites. Split metabolites are interconverted by using "extra" unimolecular fluxes.

Given a set of measurements, a flux distribution is calculated by solving (1–3) and then tested for uniqueness. This identifiability question has been addressed before using local isolation of the solution (van Winden et al., 2001) and global analysis for unique solvability (Isermann and Wiechert, 2003). In this work, we focus on the *unique solvability for physiologically relevant metabolic flux bounds*. We embed this identifiability concept into the experimental design problem of choosing measurements. Solving directly the nonlinear system of Equations (1)–(3) to identify all flux distributions that are acceptable as solutions and testing for uniqueness is computationally intractable for large-scale metabolic maps (i.e., for over 100 reactions). Therefore, we chose to circumvent the effect of nonlinearities by simply tracking whether a particular variable (flux or isotopomer fraction) participates in any given balance equation as discussed next.

Incidence Structure Analysis

Unique flux distribution, given a set of measurements, is reached when the number of fluxes n^f that can be independently varied (i.e., the number of degrees of freedom) is equal to zero. For a linear system (1), n^f is

equal to $|C| - r_T$, where r_T is the rank of matrix S and $R = \{r\}$ and $C = \{c\}$ the rows and columns, respectively. However, assessing n^f for nonlinear systems (3) is not as straightforward (Cox et al., 2007). Therefore, in this article we adopt *incidence structure analysis* to approximate n^f while avoiding dealing directly with nonlinearities. Incidence structure analysis uses an *incidence matrix* which catalogues the occurrences of variables in equations of the original system. For example, consider a simple convergent network with each metabolite containing one carbon atom:

$$v_1\text{: A} \rightarrow \text{C}, \qquad v_2\text{: B} \rightarrow \text{C}, \qquad v_3\text{: C} \rightarrow$$

The balance equations around intermediate metabolite C are

$$\begin{array}{ll} \text{metabolic balance} & v_1 + v_2 - v_3 = 0 \\ \text{isotopic balance} & v_1 I_{\text{A1}} + v_2 I_{\text{B1}} - v_3 I_{\text{C1}} = 0 \end{array} \tag{4}$$

where I_{A1}, I_{B1}, and I_{C1} are the fractions of labeled isotopomers. Denoting the appearance of a variable in an equation by X (indicating a nonzero value), the incidence matrix of (4) is

	v_1	v_2	v_3	I_{A1}	I_{B1}	I_{C1}
metabolic balance	X	X	X			
isotopic balance	X	X	X	X	X	X

(5)

Each variable can be the output variable of an equation in which it participates. Conversely each equation can have up to one output variable. For instance, we can assign v_1 and v_2 to the metabolic and isotopic balances, respectively. This assignment implies that by measuring the labeling status of all three metabolites (I_{A1}, I_{B1}, and I_{C1}) and external flux v_3, we can calculate v_1 and v_2. The maximum number of such output assignment pairs is referred to as the *generic rank* r_G of the incidence matrix, which is equal to two for the above example.

The generic rank corresponds to the maximum cardinality matching of a bipartite graph (Fig. 1) which can be solved in polynomial time (Hopcroft and Karp, 1973).

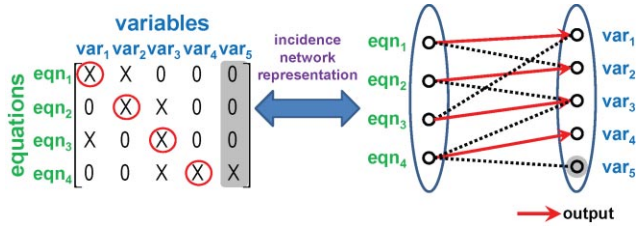

Figure 1. Two equivalent representations of the structure information in a system of equations. Incidence matrix (**left**) and bipartite graph (**right**) are equivalent ways of representing the structure information. Nonzero elements of the incidence matrix correspond to edges (both solid and broken) in the bipartite graph. The circled elements of the incidence matrix and the directed solid edges of the bipartite graph correspond to an example of maximum output variable assignments. The shaded column/node corresponds to a variable that needs to be known in advance to uniquely determine the system. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

However, we decided to use the following ILP formulation GenRNK that assigns columns to rows of an incidence matrix A (Georgiou and Floudas, 1989; Gupta et al., 1974), because it can be easily extended to account for using different cost values for experimental measurements:

$$\begin{aligned} \text{(GenRNK)}\ r_G = \max\ & \sum_{r\in R}\sum_{c\in C} y_{rc} \\ \text{s.t.}\ & \sum_{c\in C} y_{rc} - x_r = 0 \quad r \in R, \\ & \sum_{r\in R} y_{rc} - z_c = 0 \quad c \in C, \\ & y_{rc} \in \{0,1\}, x_r \in \{0,1\}, z_c \in \{0,1\}. \end{aligned}$$

In GenRNK, binary variable y_{rc} is equal to one if variable with index c is an output of equation in row r. Correspondingly, binary variables x_r and z_c model if row r and column c participate in any output assignments respectively. All unassigned columns must be measured to fully determine the system. The number of unassigned variables n^{sf} is therefore equal to $|C| - r_G$ (i.e., n^{sf}=4, in previously discussed example). It must be noted that n^{sf} may not be identical to n^f; instead it provides a valid lower bound for linear systems (Neumaier, 1997). For nonlinear systems, n^{sf} is only an approximation of n^f. A reconciliation procedure that closes the approximation gap between n^f and n^{sf} is discussed in Solution Strategy Section.

Application to MFA

The incidence matrix shown in Figure 2 is constructed by tracking the occurrences of variables v_j and I_{ik} in Equations (1)–(3). Note that one isotopomer balance equation for each intermediate metabolite is excluded in Equation (3) in order to eliminate inherent redundancy in the system. We denote

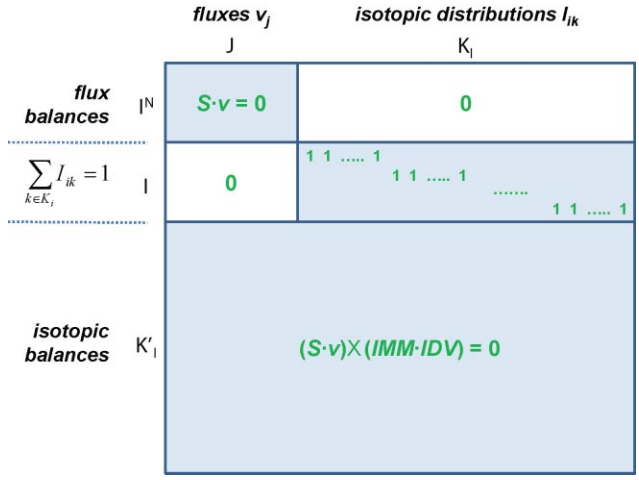

Figure 2. Generic incidence matrix of MFA for OptMeas. Here, K_I is the union of all the isotopomers and K'_I is the same as K_I except that one isotopomer balance is dropped for each metabolite. Note that Equations (1) and (2) are linear but (3) is nonlinear. The shaded submatrices correspond to structurally nonzero regions. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

the binary variable z_c in GenRNK for this incidence matrix as z_j for v_j and z_{ik} for I_{ik}. By solving GenRNK, we identify the measurements (unassigned variables) that fully determine the system. Additional restrictions arising in MFA analysis include the fact that every time a metabolite is deemed as measured all associated isotopomer fractions become specified. In addition, both measurement effort/time and potential to resolve fluxes are not the same for all measurement candidates. For example, external fluxes are considered to be easier to measure than isotopic distributions (Wiechert, 2001). This implies that the measurements must be carefully selected so as to eliminate any ambiguity in computing the remaining fluxes at a relative minimum of cost.

For the sake of clarity of presentation, we present the OptMeas formulation for the case of complete isotopic measurements. However, most current label measurement techniques do not directly measure isotopomer fractions but rather functions of isotopomer fractions, such as positional enrichment or multiplet peaks measured by NMR and mass distribution vectors (MDVs) by GC/MS. In Appendix A, we extend OptMeas to account for such partial IDV measurements.

OptMeas requires the introduction of an additional binary variable u_i that encodes whether metabolite i is measured. Here, $u_i = 1$ implies that the IDV of metabolite i is *not* measured and thus it remains a variable that can only be fathomed as the output of an equation. This new variable u_i is used to ensure that z_{ik} for all isotopomer fractions is either equal to zero or one depending on whether the corresponding metabolite is measured or not, respectively

$$z_{ik} \geq u_i, \quad k \in K_i, i \in I \tag{6}$$

OptMeas minimizes the sum of a weighted combination of all chosen measurements by using the following objective function

$$\sum_{j \in J} q_j(1 - z_j) + \sum_{i \in I} q_i(1 - u_i) \tag{7}$$

where q_j and q_i are relative weights. The relative weights used in this article are summarized in Table I. These weights are not exact estimates of the measurement costs but rather provide rough approximations of the order of their relative difficulty. These weights can be readily adjusted to better reflect different experimental settings. Figure 3 pictorially shows how OptMeas identifies the smallest measurement set that renders the incidence matrix of variables-equations of full column rank.

Table I. Relative measurement costs.

	Measurements	Relative weight
External fluxes	Liquid-phase	1
	Gas-phase	10
Isotopic distributions	IDV of source metabolite	1
	IDV of secreted liquid product	2
	IDV of gas product	3
	MDV of amino acid[a]	2–3
	MDV of R5P[b]	5
	IDV of small intermediary metabolite[c]	50

Intracellular fluxes are not part of candidate measurement set. More detailed discussion on the experiment cost of NMR analytes in terms of their abundance and signal strength can be found in Ghosh et al. (2006).

[a]The relative cost of available MDV measurement of amino acids (15 amino acids in the article) is assumed to be equal to two or three depending on the number of carbons and fragments to be analyzed. All other amino acids are not measured (e.g., Trp).

[b]Ribose 5-phosphate (R5P) is obtained relatively easy from the cellular ribonucleotide pools that are readily extracted.

[c]The complete characterization of IDV is measured only for small molecules (up to three carbons). IDV measurement of metabolites with more than four carbons is not allowed except if it is either easily analyzable substrate or product.

Solution Strategy

We first preprocess the network to detect redundancy and inherent unidentifiability, remove obviously unobservable fluxes and reduce the size of the model using the method of van Winden et al. (2001). This is followed by the step-wise procedure described next.

Step 0: Initialization. Construct OptMeas formulation for the processed model. This OptMeas is updated by introducing integer cuts in the course of the algorithm. We define set T containing the list of optimal solutions, and initialize it to be empty.

Step 1: Solve OptMeas. Solve the current realization of OptMeas using CPLEX 10 (ILOG, 2006) ILP solver and obtain (J^*, I^*) as optimal measurement choices for external fluxes and isotopic distributions.

Step 2: Remove linearly dependent flux measurements. Remove columns J^* from S. If the resulting matrix has full column rank, then continue with Step 3. Otherwise, introduce the following integer cut into OptMeas and return to Step 1:

$$\sum_{j \in J \setminus J^*} z_j - \sum_{j \in J^*} z_j < |J \setminus J^*|. \tag{8}$$

Step 3: Check for a unique flux elucidation. Test if the suggested measurement set (J^*, I^*) fully determines all fluxes in the network. This is accomplished by solving formulation TestUniq described in Appendix B. Essentially, TestUniq assesses if any fluxes can change values in the presence of the imposed measurement set (J^*, I^*). If (J^*, I^*) uniquely determines all fluxes, then move to the next step. Otherwise, go to Step 5.

Step 4: Check for solution optimality. Test if (J^*, I^*) is optimal by solving TestOpt given in Appendix B. Conceptually, TestOpt removes one measurement at a time from the set (J^*, I^*) and tests whether the value for this freed measurement variable is still locked at the same value. If so, then this measurement is redundant and can be removed from (J^*, I^*).

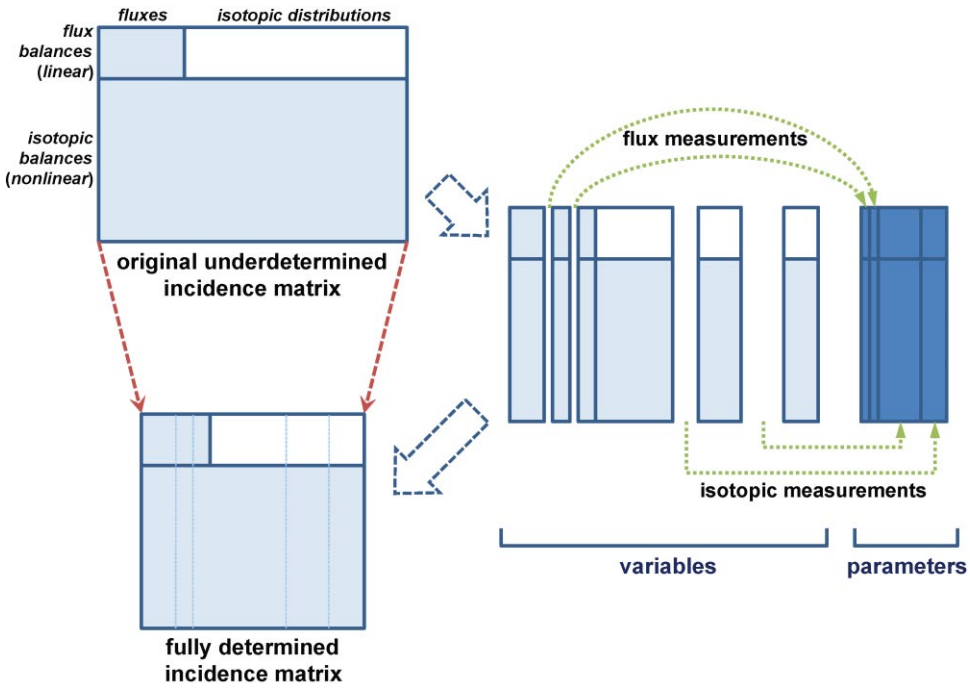

Figure 3. OptMeas modifies the structure of the incidence matrix through variable elimination. By measuring fluxes and/or isotopomer distributions, they cease to be variables implying that the corresponding columns in the incidence matrix can be removed, eventually producing a full-column rank matrix that is structurally nonsingular. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

Step 5: Termination criterion. If the current optimal solution has a higher relative cost than a predefined threshold, then terminate and report the current T as the final collection of all optimal measurement sets. Otherwise, include the current solution in T, introduce the following integer cut to OptMeas to remove (J^*, I^*) from the feasible region and go back to Step 1:

$$\sum_{j\in J\backslash J^*} z_j + \sum_{i\in I\backslash I^*} u_i - \left(\sum_{j\in J^*} z_j + \sum_{i\in I^*} u_i\right) < |J\backslash J^*| + |I\backslash I^*|. \quad (9)$$

Implementation

A set of Python codes were written for the automatic calculation of parameters (S_{ij}, $\mathrm{IMM}^j_{i'\to i,k'\to k}$, and q's) from user-provided input files (a list of reactions and their atom transitions and a list of available measurements and their relative costs). A Python script was also written to generate EMU networks from the input files utilizing EMU reduction (Antoniewicz et al., 2007b) and network decomposition with block decoupling (Young et al., 2007).

A C++ program making use of CPLEX 10 Concert technology (ILOG, 2006) was written to solve OptMeas while exploiting its sparse nature. The program was run on a Linux cluster box of four 2.6 GHz Pentium 4 processors with 8 GB memory. TestUniq and TestOpt were solved using GAMS/CONOPT 3 (Drud, 1994). GAMS/BARON 7.5 (Tawarmalani and Sahinidis, 2004) and the EMU representation were also used whenever possible to ensure global optimality during the verification process.

Results and Discussion

Small Network Example

OptMeas correctly predicted the minimal measurement sets for extensively studied small networks including a hypothetical analog network (Forbes et al., 2001; Ghosh et al., 2006), a branching network (Wiechert et al., 1999), and spiral networks (Isermann and Wiechert, 2003). As an illustrative example, we describe here a small metabolic network (see Fig. 4) adapted from Antoniewicz et al. (2006). This network has three free fluxes (eight flux variables minus five flux balance equations), but measuring all three external fluxes does not determine the system because of linear dependency. Because one external flux must be measured as a reference for other flux values, this implies that one or two of the remaining unresolved fluxes should be fixed through isotopic measurements.

We applied OptMeas to find all measurement sets at a minimum relative cost. Relative costs were assumed to be equal to one for external fluxes and I_A, two for I_D and I_E, and 50 for I_B and I_C. After the first iteration, OptMeas suggested measuring (v_1, v_6, v_8) at a total cost of three. The identifiability check in Step 2, though, detected that this

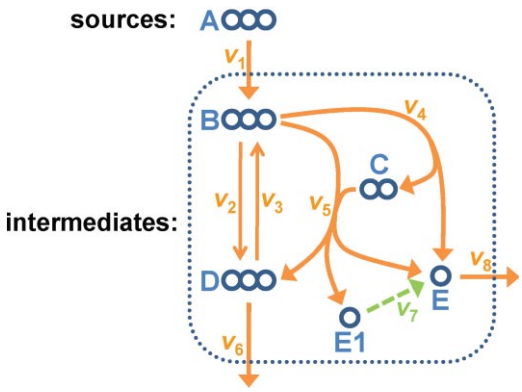

Figure 4. Illustrative example based on Antoniewicz et al. (2006). Metabolites are shown with a number of circles corresponding to the number of carbons. A dashed box separates the source and intermediary metabolites. External fluxes cross the dashed box while internal fluxes are entirely encircled within. Material flows and pseudo-fluxes are shown as solid and dashed arrows, respectively. E1 is a fictitious metabolite introduced to model the non-binary stoichiometry of reaction 5 under the mathematical formalism. E is directly measurable whereas E1 is not. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

measurement set is linearly dependent. In the second iteration, the identified measurement set was (v_1, I_A, I_D) at a total cost of four. TestUniq and TestOpt were solved globally and verified that this measurement set is indeed both optimal and fully determines the metabolic system. Through the use of integer cuts and successive iterations, we also identified (v_6, I_A, I_D) and (v_8, I_A, I_D) as alternate though still optimal solutions. Additional iterations revealed measurement sets with higher overall relative costs thus proving the existence of only three alternative optimal measurement sets.

We next used OptMeas to see if any of the full IDVs of A and D in the three suggested measurement sets could be replaced by less costly MDVs without affecting identifiability. We used a relative MDV measurement cost that is half of the corresponding IDV measurement cost. We found that the IDV of either A or D can be replaced by its MDV. Alternatively, if only the MDVs of both A and D are measured, then we need to measure two external fluxes. OptMeas found a total of nine such pairwise combinations (see Table II) that replace IDV by MDV measurements.

Table II. Optimal measurement sets identified by OptMeas for the illustrative small network.

Measurement set[a]	External flux	IDV	MDV	Relative cost
1	v_1	A	D	3
2	v_6	A	D	3
3	v_8	A	D	3
4	v_1	D	A	3.5
5	v_6	D	A	3.5
6	v_8	D	A	3.5
7	v_1, v_6	—	A, D	3.5
8	v_1, v_8	—	A, D	3.5
9	v_6, v_8	—	A, D	3.5

[a]All the measurement sets with total relative cost less than 4 are listed.

These included the (v_1, I_A, and the MDV of D) measurement set proposed in Antoniewicz et al. (2006).

Medium Scale Model for *E. coli* Metabolism

We next considered the metabolic model of the 1, 3-propanediol (PDO) producing *Escherichia coli* strain including 74 metabolites, 75 reactions, and 4,806 isotopomers (Antoniewicz et al., 2007c). All fluxes are numbered and metabolites are named in agreement with the nomenclature scheme used in Antoniewicz et al. (2007c). As in the original article, we did not account for the symmetry of glycerol to allow for a fair comparison of the obtained results. In Antoniewicz et al. (2007c), eight external fluxes ($v_{66\text{-}72}, v_{75}$), four IDVs (Gluc, Cit, Glyc, and CO_2), and partial measurements from 12 amino acids (23 distinct fragment MDVs) were measured to obtain the best-fit flux distribution in the model. According to the relative costs of Table I, this original measurement set has a total relative cost of 58. The large number of measurements is, in part, due to the fact that uncertainty in the measurements was taken into account.

We first analyzed the network to detect as many unidentifiable fluxes as possible, which were excluded from further investigations. These include the exchange rates of most bidirectional reactions and oxidative decarboxylations of malate (reactions 28 and 29) coupled with transhydrogenation reaction 64.

We next looked for new measurement sets with possibly a lower relative cost but with at least as good flux elucidation capability. It is important to stress that here we do not explicitly consider the effect of measurement imprecision. We identified 22 distinct alternate optimal measurement sets consisting of two external fluxes (O_2 and glucose uptake rates), four IDVs (Gluc, Cit, Glyc, and PDO), and fragment MDVs of three amino acids (Phe plus two interchangeable amino acids) at an overall cost of 21. We found that five (i.e., the oxygen uptake rate, the IDVs of Gluc, Cit, and Glyc, and the MDV of Phe) are essential independent of the cost structure, two (i.e., the glucose uptake rate and the IDV of PDO) are critical to maintain the optimal cost of 21, and two (i.e., the MDVs of two interchangeable amino acids) may differ. The measurement sets have similar flux elucidation capability as the original measurement set (data not shown) while substantially reducing the total relative cost and number of measurements (less than half of those of the original measurement set).

Closer inspection revealed that the oxygen uptake rate (v_{72}) is essential for the elucidation of oxidative phosphorylation (reactions 62 and 63). Phe measurement is critical because it is the only amino acid whose fragments encode the labeling information of E4P in the pentose phosphate pathway (see Fig. 5A). The additional fragment MDVs of select pairs of amino acids (see Fig. 5B) are needed to infer the sub-networks of glycolysis and citrate cycle. Note that a new measurement choice revealed by OptMeas is I_{PDO}

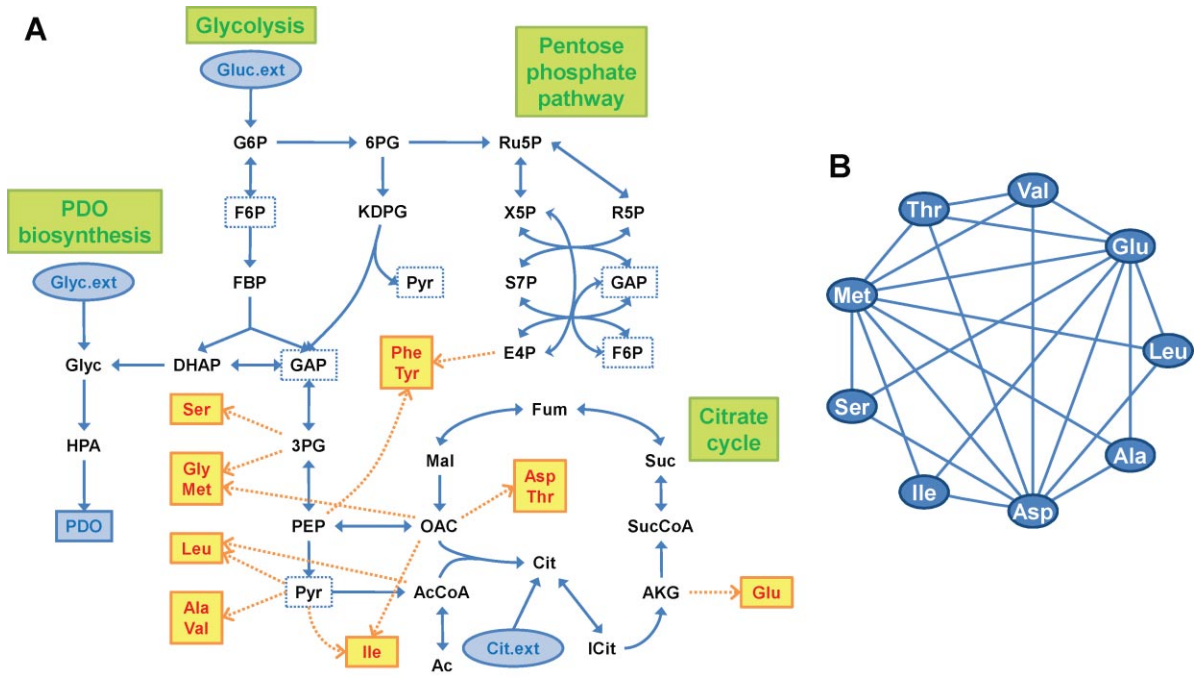

Figure 5. Network representation of the partial measurements for the medium-size *E. coli* network (Antoniewicz et al., 2007c). Panel (**A**) shows the biosynthesis of each amino acid whose fragment MDVs were measured. Source metabolites are circled, and the measured product (PDO) is boxed by a solid line. Measured amino acids are also boxed by solid lines and their direct carbon sources are indicated by dotted arrows. Some metabolites (in dotted squares) appear more than once to make the figure more readable. Panel (**B**) shows the complete set of alternatives for the two partial measurements (22 combinations in total) required in addition to the suggested core measurements. Only amino acid pairs connected with a line are valid measurement choices. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

absent in the original article (Antoniewicz et al., 2007c), which helps resolve the PDO biosynthesis pathway. We tested the effectiveness of I_{PDO} measurement by adding it to the original measurement set and solving TestUniq for various substrate labeling patterns (cf. Theorem 3 of Isermann and Wiechert (2003)). We found that I_{PDO} reduces the sum of all flux ranges (see Suthers et al., 2007) by about an order of magnitude.

In summary, in this example we demonstrated that a preliminary implementation of OptMeas can exhaustively generate minimal measurement sets without requiring any measurement to be pre-selected in advance for small to medium-size metabolic models. By rank-ordering different measurement choices with respect to their relative costs it provides a systematic way to decide on what needs to be measured. As a by-product of this analysis, OptMeas pinpoints all essential measurements that must be part of any measurement set for complete flux elucidation as well as catalogues the unidentifiable fluxes.

Large-Scale Model for *E. coli* Metabolism

Finally, we revisited the recently published large-scale isotopomer model of amorphadiene producing *E. coli* strain with 184 metabolites, 238 reactions, and 17,346 isotopomers (Suthers et al., 2007). The abbreviations for all metabolites and fluxes follow the conventions used in Suthers et al. (2007) and Reed et al. (2003). In Suthers et al. (2007), three external fluxes ($v_{AMDNt}, v_{BIOMASS\_EC\_ISO}, v_{EX\_glc-D}$), the IDV of Glc-D(e), and 22 distinct fragment MDVs of 13 amino acids were monitored. By using these measurements, Suthers et al. (2007) were able to tighten the ranges of flux values, but concluded that much ambiguity remained in flux elucidation, in part, due to the choice of substrate labeling.

To remedy this ambiguity in flux elucidation we explored computationally the potential effectiveness of performing additional labeling experiments. In addition to the thirteen monitored amino acids, we allowed the measurement of fragment MDVs of Arg and Tyr (Antoniewicz et al., 2007a). The biosynthesis of these fifteen measured amino acids is shown in Figure 6 on a map of central metabolism of the large-scale model. We modified the full model of the original article (Suthers et al., 2007) by treating reversible reactions H2Ot5, NH4t, O2t, PIt6, ACACT1r, HMGCOAS, and ADK1 as unidirectional in the direction of their net flux. In addition, we excluded the inherently unidentifiable and poorly resolved fluxes listed in Table III.

By solving OptMeas for the adjusted model using the relative costs of Table I, we identified four alternate optimal solutions. Common in all these alternate optimal solutions were six external fluxes ($v_{EX\_h}, v_{EX\_h2o}, v_{EX\_nh4}, v_{EX\_o2}, v_{EX\_pi}, v_{EX\_so4}$) and ATP maintenance ($v_{ATPM}$). These seven flux measurements are deemed to be *essential measurements*. In addition, a

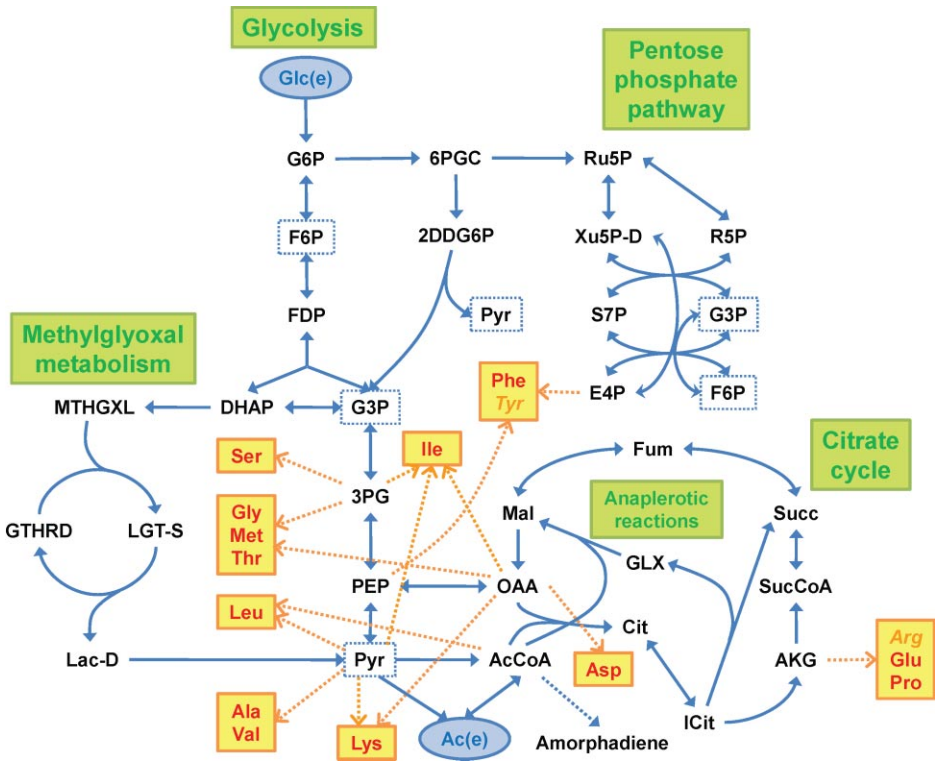

Figure 6. Biosynthesis of the amino acids measured in the large-scale *E. coli* model (Suthers et al., 2007). The 13 amino acids measured in Suthers et al. (2007) are shown in bold while two amino acids (Arg and Tyr) that can be measured in the current study are shown in italic. Metabolites are named using the conventions defined in *i*JR904 (Reed et al., 2003), and thus are slightly different from those in Figure 5A. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

single interchangeable measurement (i.e., IDV of acetate, IDV of external CO_2, MDV of Arg, or MDV of Tyr) is also needed for complete flux elucidation.

We found that if the glucose substrate is not uniformly labeled but rather contains glucose with only the first carbon labeled, then tracking the seven fluxes and the full IDV of acetate would allow us to resolve completely all remaining net fluxes with only a few exceptions for fluxes insensitive to any labeling scheme under stationary conditions. The identified additional measurements improved the objective value of TestUniq (sum of squared differences, SSD) from 22.1 of the original measurement set in Suthers et al. (2007) to 18.3.

We also tested other substrate labeling patterns for their potential to further improved flux resolution in conjunction with the proposed additional measurements. Specifically, we tried 25% [U-^{13}C] and 75% [n-^{13}C] glucose labeling, where n can be any integer from 1 to 6 meaning that a single glucose carbon atom is labeled at a time. The SSD of these patterns was always within the range 16.1–18.9 (18.3 when $n = 1$ as seen above) and the smallest SSD was achieved when $n = 5$. In contrast, using the proposed additional measurements, the 20% [U-^{13}C] and 80% non-enriched glucose resulted in a higher SSD value of 19.7. We believe this is due to the enhanced separation capability of the n-^{13}C towards the G3P produced by glycolysis and pentose phosphate cycle. This result shows that the substrate labeling can affect the range of fluxes that can be reliably determined (Möllney et al., 1999).

Summary

In this work, we introduced the OptMeas formulation and procedure to identify the minimal measurement sets that determine all the identifiable fluxes in a metabolic network. OptMeas makes use of incidence structure analysis to approximate the identifiability constraint. This scheme enables OptMeas to generate all the alternative sets of the minimal measurements required to determine the system and to do so very quickly (one solution of OptMeas required only a few minutes of CPU time even for the large-scale *E. coli* model). This good scalability enables the exhaustive identification of all alternative optimal solutions. We also introduced a systematic set of tests that queries and refines the solution of OptMeas for both optimality and feasibility.

A key novel aspect of OptMeas is that it does not require any measurement to be fixed or picked in advance. Thus, it can be applied to any experimental situation to rank-order different measurement choices. Although we specifically focused on ^{13}C isotopomer representation in the manuscript, OptMeas formulation is applicable to

Table III. Fluxes removed from the large-scale *E. Coli* model.

Category	Removed fluxes
Inherently unidentifiable[a]	*Exchange rates of all the reversible reactions*
	ALAR, ALATA_L, ASPT, ASPTA1, DAAD, VPAMT
	PGMT
	PPA
	AGPR
	ACONT, AKGD, CITL, CS, FRD2, FRD3, FUM, ICDHy, MDH, SUCD1i, SUCOAS
	SERAT
	FTHFD, GLYCL, MTHFD
	GLNS, GLUDy, GLUN
	GHMT2, PGCD, PSERT, PSP_L
	ENO, G1PP, GLCP, GLCS1, GLGC, HEX1, PDH, PFK, PGK, PGM, PPS, PYK
	ADK1
	ATPS4r, FDH2, LDH_D2, NADH6, NADH7, NADH8, NADH9, NADH10, NADH12, SUCD4, THD2, THD5
	ACKr, ACS, LDH_D, PTAr
	GLCpts, GLCt2
	VALTA
Poorly resolved[b]	ASNN, ASNS1, ASNS2
	ACGK, ACGS, ACODA, ACOTA, ARGSL, ARGSS, AST, CBPS, G5SADs, G5SD, GLU5K, OCBT, P5CD, SADH, SGDS, SGSAD, SOTA
	CYSS, TRPAS1
	GLUSy
	SERD_L
	FBP
	GLYOX, LGTHL, MGSA
	EDA, PGDHy, PGL
	CBMK2

Listed fluxes (grouped according to subsystem) were identified by solving FluxRange (Suthers et al., 2007) using 25% [U-^{13}C] and 75% [1-^{13}C] glucose substrate. FluxRange is solved using local optimization, so the list may not be complete. The reactions removed in the reduced model of Suthers et al. (2007) are underlined.

[a]Inherently unidentifiable fluxes are those that are not fully resolvable (degree of resolution <0.95) by measuring all external fluxes and the IDVs of all metabolites. Some linear dependencies such as the triplet (VPAMT, ALATA_L, VALTA) or the sextuplet (FRD2, FRD3, NADH7, NADH8, NADH9, NADH10) were additionally detected using local analysis.

[b]Poorly resolved fluxes are those that are not fully resolvable (degree of resolution <0.95) by measuring all external fluxes and the available isotopic measurements (see Table I).

any labeling scheme of any stable isotopes such as ^{2}H, ^{13}C, ^{15}N, and $^{17[18]}$O using any representation of isotopic distribution such as cumomers and EMUs. It must be noted that OptMeas for experimental design can integrate with other identifiability analysis methods.

OptMeas correctly identified all the optimal measurement sets for the small examples, including the branching network that cannot be determined using positional enrichment methods. When applied to the medium-scale *E. coli* model, OptMeas predicted that the O_2 uptake and Phe measurements are mandatory for flux elucidation. It also suggested an additional PDO measurement which we showed could improve flux elucidation capability. OptMeas was successfully scaled-up to the large-scale *E. coli* model with 17,346 isotopomers to suggest additional measurements that would be effective in tightening the range of flux estimate. OptMeas is particularly useful for the ability to pinpoint essential measurements, and these computationally derived insights can be used to plan measurement experiments.

The proposed solution procedure can greatly benefit by efficient global optimization algorithms as we demonstrated for the small examples. However, currently available state-of-the-art global optimization solvers are not capable of handling even the medium sized *E. coli* model using the most compact EMU representation. We are working on the possibility of tuning commercial global optimization solvers and other approaches to increase the size of the identifiability problem that is globally solvable. This extension is especially important for multiple labeled atoms because of the extremely large number of resulting isotopomers.

The major focus of this article has been the identification of the minimal measurement sets with full flux elucidation capacity under exact measurements. However, making redundant measurements at the cost of increased expense could be necessary to make the flux estimation more reliable in the presence of measurement errors or modeling uncertainties. Furthermore, substrate choices and their labeling patterns can also be used as experimental design variables in order to improve flux resolution. Finally, exploring more elaborate measurement cost models that account for savings due to the measurement of similar metabolites could further improve flux elucidation efficiency.

The authors would like to gratefully acknowledge useful discussions with the J.D. Keasling lab about the large-scale *E. coli* model and financial support work by DOE DE-FG02-05ER25684.

Appendix A: ILP Formulation OptMeas for the Identifiability in MFA

As noted in the article, most current label measurement techniques do not directly measure IDVs but rather functions of isotopomer fractions, such as MDVs by GC/MS. In this section, we extend OptMeas to account for such partial IDV measurements. In practice, an isotopic measurement of metabolite i provides independent measurements $p_{im}, m \in M_i(|M_i| < |K_i|)$, which are related to I_{ik} through the following measurement equations:

$$\frac{\sum_{k\in K_i} \alpha_{ikm} I_{ik}}{\sum_{k\in K_i} \beta_{ikm} I_{ik}} = p_{im}, \quad m \in M_i,\ i \in I \tag{10}$$

where α_{ikm} and β_{ikm} are parameters linking I_{ik} and p_{im}. For example, measurement linking equations between IDV and MDV of acetate with two carbon atoms are:

$$I_{\mathrm{Ac:00}} = p_{\mathrm{Ac:M0}}, \qquad I_{\mathrm{Ac:01}} + I_{\mathrm{Ac:10}} = p_{\mathrm{Ac:M1}},$$
$$I_{\mathrm{Ac:11}} = p_{\mathrm{Ac:M2}}.$$

Here, isotopomers are named based on the labeling status of each carbon (0 for ^{12}C and 1 for ^{13}C), and mass fractions by M followed by the total number of ^{13}C. There is a total of three MDV measurements, but only two of them are independent because $p_{\mathrm{Ac:M0}} + p_{\mathrm{Ac:M1}} + p_{\mathrm{Ac:M2}} = 1$.

The new system of equations with partial metabolite measurements contains variables v_j, I_{ik}, p_{im} and Equations (1)–(3) and (10). We denote the binary variable z_c in GenRNK as z^p_{im} for the columns corresponding to p_{im}. We additionally introduce binary variable u^p_i with associated relative cost parameter q^p_i to model the decision of making a partial measurement of metabolite i. Variable u^p_i is involved in the model in the same way as u_i, thus:

$$z^p_{im} \geq u^p_i, \quad m \in M_i,\ i \in I. \tag{11}$$

Obviously, there is no reason of measuring a metabolite i for both complete and partial measurements. Therefore, we introduce the following constraints into OptMeas:

$$u_i + u^p_i \geq 1, \quad i \in I. \tag{12}$$

The cost objective function of OptMeas is extended to account for partial measurements as follows:

$$\sum_{j\in J} q_j(1 - z_j) + \sum_{i\in I} (q_i(1 - u_i) + q^p_i(1 - u^p_i)). \tag{13}$$

The resulting modified OptMeas can be further extended using the same line of analysis presented here to account for multiple partial measurement options for isotopic distributions of metabolites.

Appendix B: NLP Sub-Problems for the Proposed Procedure

In the proposed procedure, the current measurement set (J^*, I^*) is tested if they under-determine (Step 3) or unnecessarily over-determine (Step 4) the MFA system. These tests are performed by solving nonlinear programming (NLP) problems described in this section. We first solve a forward problem for a specific substrate labeling pattern $\overline{I}_{ik}\ \forall i \in I\backslash I^N$ and flux distribution $\overline{v}_j$, and obtain all the other IDVs $\overline{I}_{ik}\ \forall i \in I^N$.

Uniqueness Test

In Step 3, in order to check if (J^*, I^*) uniquely determines all fluxes within flux bounds [v^{LO}, v^{UP}], the following NLP problem must be solved to global optimality:

$$\begin{aligned}
(\text{TestUniq}) \quad \max \quad & \sum_{j\in J\backslash J^*} (v_j - \overline{v}_j)^2 \\
\text{s.t.} \quad & \text{Equations (1)-(3)}, \\
& v_j = \overline{v}_j \quad j \in J^*, \\
& I_{ik} = \overline{I}_{ik} \quad k \in K_i, i \in I^*, \\
& v^{LO}_j \leq v_j \leq v^{UP}_j, \quad I_{ik} \in [0,1]
\end{aligned}$$

TestUniq is a variation of FluxRange of Suthers et al. (2007), and has an objective function the sum of squared differences (SSD) of all the fluxes that are not measured. Maximum SSD corresponds to the maximum difference in flux distributions that yield the same labeling pattern of the measured metabolites. If the maximum SSD is zero, then (J^*, I^*) uniquely determines all fluxes. If the maximum SSD is larger than zero then the imposed measurements do not uniquely determine the system. This numerical uniqueness check relies on both the global optimality certificate and Theorem 3 of Isermann and Wiechert (2003) which states that almost all realizations of substrate labeling produce the same identifiability results (with probability one) if the measurement is noise free.

Optimality Test

In Step 4, in order to check if the current solution corresponds to a minimal measurement set, we solve the following NLP problems for each $j^* \in J^*$ and $i^* \in I^*$:

$$\begin{aligned}
(\text{TestOpt}(j^*)) \quad \max \quad & (v_{j^*} - \overline{v}_{j^*})^2 \\
\text{s.t.} \quad & \text{Equations (1)-(3)}, \\
& v_j = \overline{v}_j \quad j \in J^*\backslash\{j^*\}, \\
& I_{ik} = \overline{I}_{ik} \quad k \in K_i, i \in I^*, \\
& v^{LO}_j \leq v_j \leq v^{UP}_j, \quad I_{ik} \in [0,1]
\end{aligned}$$

$$\begin{aligned}
(\text{TestOpt}(i^*)) \quad \max \quad & \sum_{k\in K_{i^*}} (I_{ik} - \overline{I}_{ik})^2 \\
\text{s.t.} \quad & \text{Equations (1)-(3)}, \\
& v_j = \overline{v}_j \quad j \in J^*, \\
& I_{ik} = \overline{I}_{ik} \quad k \in K_i, i \in I^*\backslash\{i^*\}, \\
& v^{LO}_j \leq v_j \leq v^{UP}_j, \quad I_{ik} \in [0,1]
\end{aligned}$$

Essentially these formulations fix all measurements, except for one at a time, and then explore if the "free" measurement variable is locked at its measured value due to the fixing of all the other measured variables. If so, then this measurement is redundant, otherwise it is indeed needed for fully resolving the network. Therefore, if the optimal value to TestOpt is larger than zero for some j^* or i^*, then the corresponding measurement is essential given measurements (J^*, I^*). Therefore, if all the optimal values to TestOpt are larger than zero, (J^*, I^*) contains no redundant measurement, and thus it is optimal. If the optimal value to TestOpt is zero for some j^* or i^*, then this indicates that the corresponding measurement is redundant.

References

Antoniewicz MR, Kelleher JK, Stephanopoulos G. 2006. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metab Eng 8(4):324–337.

Antoniewicz MR, Kelleher JK, Stephanopoulos G. 2007a. Accurate assessment of amino acid mass isotopomer distributions for metabolic flux analysis. Anal Chem 79(19):7554–7559.

Antoniewicz MR, Kelleher JK, Stephanopoulos G. 2007b. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. Metab Eng 9(1):68–86.

Antoniewicz MR, Kraynie DF, Laffend LA, González-Lergier J, Kelleher JK, Stephanopoulos G. 2007c. Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. Metab Eng 9(3):277–292.

Bailey JE. 1991. Toward a science of metabolic engineering. Science 252(5013):1668–1675.

Cox D, Little J, O'Shea D. 2007. Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra. New York: Springer.

Drud A. 1994. CONOPT—A large-scale GRG code. ORSA J Comput 6(2):207–216.

Forbes NS, Clark DS, Blanch HW. 2001. Using isotopomer path tracing to quantify metabolic fluxes in pathway models containing reversible reactions. Biotechnol Bioeng 74(3):196–211.

Georgiou A, Floudas CA. 1989. Optimization model for generic rank determination of structural matrices. Int J Control 49(5):1633–1644.

Ghosh S, Grossmann IE, Ataai MM, Domach MM. 2006. A three-level problem-centric strategy for selecting NMR precursor labeling and analytes. Metab Eng 8(5):491–507.

Gupta PK, Westerberg AW, Hendry JE, Hughes RR. 1974. Assigning output variables to equations using linear programming. AIChE J 20(2):397–399.

Hopcroft JE, Karp RM. 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. SIAM J Comput 2(4):225–231.

ILOG. 2006. ILOG CPLEX 10.1 User's Manual. Mountain View, CA: ILOG Inc.

Isermann N, Wiechert W. 2003. Metabolic isotopomer labeling systems. Part II. Structural flux identifiability analysis. Math Biosci 183(2):175–214.

Lin C-T. 1974. Structural controllability. IEEE Trans Automat Control 19(3):201–208.

Möllney M, Wiechert W, Kownatzki D, de Graaf AA. 1999. Bidirectional reaction steps in metabolic networks. IV. Optimal design of isotopomer labeling experiments. Biotechnol Bioeng 66(2):86–103.

Neumaier A. 1997. Scaling and structural condition numbers. Linear Algebra Appl 263(1–3):157–165.

Nielsen J. 2003. It is all about metabolic fluxes. J Bacteriol 185(24):7031–7035.

Rantanen A, Mielikäinen T, Rousu J, Maaheimo H, Ukkonen E. 2006. Planning optimal measurements of isotopomer distributions for estimation of metabolic fluxes. Bioinformatics 22(10):1198–1206.

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(9):R54.

Riascos CAM, Gombert AK, Pinto JM. 2005. A global optimization approach for metabolic flux analysis based on labeling balances. Comput Chem Eng 29(3):447–458.

Sauer U. 2006. Metabolic networks in motion: ^{13}C-based flux analysis. Mol Syst Biol 2:62.

Schmidt K, Carlsen M, Nielsen J, Villadsen J. 1997. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. Biotechnol Bioeng 55(6):831–840.

Shields RW, Pearson JB. 1976. Structural controllability of multiinput linear systems. IEEE Trans Automat Control 21(2):203–212.

Sonntag K, Eggeling L, de Graaf AA, Sahm H. 1993. Flux partitioning in the split pathway of lysine synthesis in *Corynebacterium glutamicum*: Quantification by ^{13}C- and ^{1}H-NMR spectroscopy. Eur J Biochem 213(3):1325–1331.

Stephanopoulos G. 1999. Metabolic fluxes and metabolic engineering. Metab Eng 1(1):1–11.

Suthers PF, Burgard AP, Dasika MS, Nowroozi F, van Dien S, Keasling JD, Maranas CD. 2007. Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. Metab Eng 9(5–6):387–405.

Tawarmalani M, Sahinidis NV. 2004. Global optimization of mixed-integer nonlinear programs: A theoretical and computational study. Math Program 99(3):563–591.

van Winden WA, Heijnen JJ, Verheijen PJT, Grievink J. 2001. A priori analysis of metabolic flux identifiability from ^{13}C-labeling data. Biotechnol Bioeng 74(6):505–516.

van Winden WA, Heijnen JJ, Verheijen PJT. 2002. Cumulative bondomers: A new concept in flux analysis from 2D [^{13}C,^{1}H] COSY NMR data. Biotechnol Bioeng 80(7):731–745.

Wiechert W. 1995. Algebraic methods for the analysis of redundancy and identifiability in metabolic ^{13}C-labeling systems. In: Schomberg D, Lessel U, editors. Bioinformatics: From nucleic acids and proteins to cell metabolism. Braunschweig, Germany: VCH. p 169–184.

Wiechert W. 2001. ^{13}C metabolic flux analysis. Metab Eng 3(3):195–206.

Wiechert W, Möllney M, Isermann N, Wurzel M, de Graaf AA. 1999. Bidirectional reaction steps in metabolic networks. III. Explicit solution and analysis of isotopomer labeling systems. Biotechnol Bioeng 66(2):69–85.

Yang J, Wongsa S, Kadirkamanathan V, Billings SA, Wright PC. 2007. Metabolic flux estimation: A self-adaptive evolutionary algorithm with singular value decomposition. IEEE-ACM Trans Comput Biol 4(1):126–138.

Young JD, Walther JL, Antoniewicz MR, Yoo H, Stephanopoulos G. 2007. An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. Biotechnol Bioeng 99(3):686–699.

Zupke C, Stephanopoulos G. 1994. Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. Biotechnol Prog 10(5):489–498.