Planning isotopomer measurements for estimation of metabolic fluxes

Ari Rantanen[1], Taneli Mielikäinen[1], Juho Rousu[2] and Esko Ukkonen[1]

[1]Department of Computer Science and HIIT BRU, University of Helsinki, Finland

[2]Department of Computer Science, Royal Holloway, University of London, UK

Firstname.Lastname@cs.Helsinki.FI

Abstract: Flux estimation by using isotopomer information of metabolites is currently the only method that can give quantitative estimates of the activity of metabolic pathways. However, the measurement of isotopomer distributions of intermediate metabolites is costly and tedious with current technologies. In this paper we study the question of finding the smallest subset of metabolites to measure that ensure an adequate level of the isotopomer information. We study the computational complexity of this optimization problem in the case of the so-called positional enrichment data, give exact and fast heuristic solutions and evaluate empirically the efficacy of the proposed methods.

1 Introduction

The goal of metabolic flux analysis is to discover the steady state conversion velocities of metabolites to each other through chemical reactions catalyzed by the enzymes of an organism. Information about reaction rates, or fluxes, constitutes an important aspect of the physiological state of the cell that can be harnessed in many different applications ranging from pathway optimization in metabolic engineering [SAN98] and from characterization of the physiology of an organism [Kel01] to more efficient drug design for human diseases such as cancer [BSCL04].

The most accurate information about the fluxes can be obtained by conducting isotopomer tracer experiments where the cell is fed with a mixture of natural and ^{13}C-labeled nutrients. The fate of the ^{13}C atoms can be observed by measuring the resulting NMR [SGH$^+$99] or mass spectrum [CN99, WH99] of metabolic products and intermediates. From the measurements one obtains information about the fluxes of the alternative pathways producing a metabolite. This methodology has been successfully applied in numerous cases to explicitly solve key fluxes in specific metabolic networks and experimental conditions of interest [MdGW$^+$96, SAN98, Szy95].

A popular general method for estimating the flux distribution of an arbitrary metabolic network is based on iteration where a candidate flux distributions are generated iteratively until the fluxes fit well-enough with the measured data [SCNV97, WMI$^+$99, WMPdG01]. Recently Rousu *et al.* [RRM$^+$03] proposed a general direct flux estimation method that

first propagates the measurement information in the metabolic network and then augments the stoichiometric constraints to the fluxes with generalized isotopomer balances. Rantanen *et al.* [RRP⁺05] improved the propagation of isotopomer information by introducing a partition of metabolite fragments to sets having equal isotopomer distributions in all steady states.

Developing measurement techniques for intermediate metabolites and conducting the measurements using a developed technique are nontrivial, laborious and costly processes. This experimental burden could be eased by concentrating the measurements to non-redundant subsets of metabolites that alone give the most information about the fluxes. In this paper we formalize this problem and derive algorithms for finding such optimal subsets of metabolites to measure. The methods are facilitated by the recent flow analysis method of Rantanen *et al.* [RRP⁺05], that enable us to discover redundancies within sets of measurements of metabolites.

The structure of the paper is the following. Section 2 gives a short introduction to flux estimation using isotopomer information. Section 3 introduces the concept of fragment equivalence. In Section 4 we motivate the problem of selecting optimal set of metabolites to measure, define a measurement optimization problem at hand, study its computational complexity and give heuristic and exact algorithms for solving it. Section 5 presents the results of experiments conducted with metabolic model of central carbon metabolism of *Saccharomyces cerevisiae*. A summary of the related work is given in Section 6, together with discussion on possible future directions.

2   Metabolic flux estimation using ^{13}C isotopic tracers

A metabolic network is composed of a set $\mathcal{M} = \{M_1, \ldots, M_m\}$ of metabolites and a set $\mathcal{R} = \{\rho_1, \ldots, \rho_n\}$ of reactions that perform their interconversions.

For the purposes of ^{13}C isotopic tracing, only carbon atoms are of interest. Thus, we represent a k-carbon metabolite as a set of carbon locations $M = \{c_1, \ldots, c_k\}$. Fragments of metabolites are simply subsets $F = \{f_1, \ldots, f_h\} \subseteq M$ of the metabolite. A fragment F of M is denoted as $M|F$. In a slight abuse of notation, we also denote by F and M the corresponding physical pools of molecules that have the required molecular structure.

Isotopomers—different isotopic versions of the molecule $M = \{c_1, \ldots, c_k\}$—are represented by binary sequences $b = (b_1, \ldots, b_k) \in \{0, 1\}^k$ where $b_i = 0$ denotes a ^{12}C and $b_i = 1$ denotes a ^{13}C in location c_i. Molecules that belong to the b–*isotopomer* of M are denoted by $M(b)$. Isotopomers of metabolite fragments $M|F$ are defined in an analogous manner: a molecule belongs to the $F(b)$–*isotopomer* of M, denoted $M|F(b_1, \ldots, b_h)$, if it has ^{13}C in all locations f_j that have $b_j = 1$, and ^{12}C in other locations of F.

The *isotopomer distribution* $D(M)$ of metabolite M gives the relative abundances $0 \leq P_M(b) \leq 1$ of each isotopomer $M(b)$ in the pool of M such that

$$\sum_{b \in \{0,1\}^{|M|}} P_M(b) = 1.$$

178

Isotopomer distribution $D(M|F)$ of fragment $M|F$ is defined analogously.

Reactions are pairs $\rho_j = (\alpha_j, \lambda_j)$ where $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{mj}) \in \mathbb{Z}^m$ is a vector of *stoichiometric coefficients*—denoting how many molecules of each kind are consumed and produced in a single reaction event—and λ_j is a carbon mapping describing the transition of carbon atoms in ρ_j. Bidirectional reactions are modeled as a pair of reactions. (For simplicity of presentation, we assume in this paper that the reactions have *simple stoichiometries* $\alpha_{ij} \in \{-1, 0, 1\}$ and that the carbon mappings λ_j are bijections. Both of these restrictions, however, can be lifted without great conceptual difficulty.) Metabolites M_i with $\alpha_{ij} < 0$ are called *reactants* and with $\alpha_{ij} > 0$ are called *products* of ρ_j. We make the modeling assumption that whenever two carbons originate from the same reactant and are transferred to the same product, the atoms remain together physically in that reaction. (This assumption can be removed by modeling each reaction not satisfying the assumption by a sequence of reactions satisfying the assumption.) A *pathway* from fragments $\{F_1, \ldots, F_p\}$ to F' is a sequence of reactions that define a bijective (composite) mapping from all carbons of $\{F_1, \ldots, F_p\}$ to all carbons of F'.

It will be useful to distinguish between the subpools of a metabolite produced by different reactions and pathways. Therefore, we denote by $M_{ij}, j > 0$, the subpool of M_i produced ($\alpha_{ij} > 0$) or consumed ($\alpha_{ij} < 0$) by reaction ρ_j. By M_{i0} we denote the subpool of M_i that is related to the external inflow ($\beta_i < 0$) or external outflow ($\beta_i > 0$) of M_i. We call the sources of external inflows *external nutrients*. We denote by $I_i = \{M_{ij}|\alpha_{ij} < 0\}$ the inflow and by $O_i = \{M_{ij}|\alpha_{ij} > 0\}$ the outflow subpools of M_i. Subpools of fragments are defined analogously.

In flux estimation, the quantities of interest are the *velocities* $v_j \geq 0$ of the reactions ρ_j, giving the number of reaction events of ρ_j per time unit. If the velocities v_j of reactions $\rho_j \in \mathcal{R}$ and the sizes of metabolite pools stay constant over time, we say that the metabolic network is in a *steady state*. In such a state the metabolite balance

$$\sum_{j=1}^{n} \alpha_{ij} v_j = \beta_i \tag{1}$$

holds for each metabolite M_i, which tells that the rate of production and consumption of the intermediate metabolite is the same. Here β_i is the measured external inflow ($\beta_i < 0$) or external outflow ($\beta_i > 0$) of metabolite M_i. If in addition the isotopomer distributions of metabolites remain constant, the system is in a *isotopomeric steady state*. In such a state, the rate of production and consumption of each metabolite M_i satisfies the isotopomer balance

$$\sum_{j=1}^{n} \alpha_{ij} v_j P_{M_{ij}}(b) = \beta_i P_{M_{i0}}(b) \tag{2}$$

for any $b \in \{0, 1\}^{|M_i|}$.

The isotopomer distributions of the outflow subpools M_{ij} are always identical to the distribution of the whole metabolite pool M_i as we assume that reactions sample uniformly their reactant pools. If, however, the pathways leading to a *junction metabolite*—a metabolite with more than one producer—manipulate the carbons of the metabolite differently, then

the isotopomer distributions of the inflows ($\alpha_{ij} < 0$) often have differences. Also, (2) together with (1) gives linearly independent equations that constrain the fluxes, ideally so that a single (correct) flux distribution can be pinpointed. The stoichiometric linear system (1) alone can be underdetermined, hence the additional equations (2) are used.

Applying (2) suffers from two difficulties [RRM$^+$03]:

1. The (mass spectrometric and NMR) measurements do not in general come in the form of fully determined isotopomer distributions, but as a set of linear *isotopomer constraints*

$$d_{i,h} = \sum_b s_{b,i,h} P_{M_i}(b), \tag{3}$$

for the metabolites M_i, where the coefficients $s_{b,i,h} \in \mathbb{R}$ depend on the measurement technique and the metabolite. For example, the constraints for a metabolite given by a mass spectrometric measurement depend on the fragmentation pattern of a metabolite in the tandem mass spectrometer and how many of the produced fragments have sufficiently high frequency to exceed the detection limit. NMR technology gives more direct access to relative frequencies of some isotopomers. However, in general some isotopomers cannot be uncovered and the sensitivity is lower than that of a mass spectrometer.

Because of these practical hindrances, instead of (2), we will have to resort to weaker form of balances

$$\sum_{j=1}^{n} \alpha_{ij} v_j d_{i,h} = \beta_i d_{i0,h}, \tag{4}$$

(for all metabolites M_i) that at worst only contain the metabolite balance equations (1) and in the best case, meet (2).

2. With current technologies, not all metabolites can be measured, so all isotopomer frequencies are not available. This calls for methods that can be used to derive isotopomer frequencies or their combinations from measurements made for other metabolites in the network. In [RRM$^+$03], a general methodology was proposed, where measurements of the form of (3) can be propagated in between two junction metabolites to obtain the constraints of the form of (4) for the junction metabolite with as many linearly independent equations as possible. The method relies on the fact that in individual reactions—and in general pathways with no junction metabolites—from isotopomer constraints (3) of reactants one can compute isotopomer constraints to products, and vice versa, by using vector space operations and the carbon maps.

3 Fragment equivalence

In [RRP$^+$05], a flow analysis method was developed that extends the scope of propagation to make it possible to propagate information through the junction metabolites. The idea

is to find fragments $M|F$ and $M_i|F'$ that are *equivalent* (denoted by $M|F \equiv M_i|F'$) in the sense that in all isotopomeric steady states their isotopomer distributions are the same, no matter what kind of labellings are used for the external nutrients. Intuitively, source fragment $M|F$ and target fragment $M_i|F'$ are equivalent if all pathways producing F' from the fragments of external nutrients go through F, in all pathways from F to F' carbons of F stay intact and all pathways have the same carbon maps between F and F'. Now, because F is always a precursor of F', carbons of F travel always intact to F' and the fragments F are similarly oriented when reaching F' via every pathway (carbon maps are the same), the isotopomer distribution of F' is equal to the one of F — regardless of fluxes and the labellings of nutrients.

The basic example of such equivalence is that of reactant and product fragments of a single reaction, which easily follows from the assumed intactness of fragments in a single reaction:

Lemma 1. *Let $\rho_j = (\alpha_j, \lambda_j)$ be a reaction with a reactant M and a product M_i. If fragment $M|F$ satisfies $\lambda_j(F) \subset M_{ij}$, then $M|F \equiv M_{ij}|\lambda_j(F)$.*

If M_i is not a junction metabolite, that is, it has only one inflow, then $M_{ij} = M_i$ in Lemma 1, and we have $M|F \equiv M_i|\lambda_j(F)$. By the transitivity of the equivalence relation $\equiv$, the result can be applied to an *unbranched pathway* i.e., to a pathway that contains no junction metabolites. Here, intactness of the fragment is ensured by requiring that the image of the fragment belongs to a single metabolite in each prefix of the pathway:

Lemma 2. *Let $M, M_{i_h}, 1 \leq h \leq r$, be the metabolites and $\rho_{j_h}, 1 \leq h \leq r$, the reactions of a pathway. Let M be a reactant of ρ_{j_1} and let ρ_{j_h} be the sole producer of M_{i_h}, and denote by $\Lambda^h = \lambda_{j_h} \circ \cdots \circ \lambda_{j_1}$ the composite carbon mapping of the pathway $(\rho_{j_1}, \ldots, \rho_{j_h})$. If for some fragment F of M, $\Lambda^h(F) \subseteq M_{i_h}$ for each $1 \leq h \leq r$, then $M|F \equiv M_{i_r}|\Lambda^r(F)$.*

For source fragment $M|F$ and target fragment $M_i|F'$ having several pathways in between them, it is further required that F must be always be a precursor of F' and that the composite carbon mappings of the pathways are the same.

Lemma 3. *Let $R_1, \ldots, R_p$ be the set of unbranched pathways connecting M and M_i with associated composite carbon mappings $\Lambda^1, \ldots, \Lambda^p$ and let M_{ik} denote the subpool of M_i produced by R_k. For fragment $F = (f_1, \ldots, f_h)$ of M, if $M|F \equiv M_{ik}|\Lambda^k(F)$ for $1 \leq k \leq p$, $\Lambda^1(f_t) = \cdots = \Lambda^p(f_t)$ for each $1 \leq t \leq h$ and every pathway producing F' from some fragments of external nutrients contains F then $M|F \equiv M_i|F'$.*

The equivalence relation defined with above lemmas is symmetric and transitive. If we further require that every fragment is equivalent to itself, the equivalence partitions a metabolic network to *equivalence classes* of fragments with equivalent isotopomer distributions. The equivalence classes of fragments can be efficiently (in polynomial time) found by constructing the fragment flow graph of a metabolic network and applying dominator tree analysis [LT79] to it [RRP+05]. Briefly, in a dominator tree constructed from the fragment flow graph fragment F' is a descendant of F if and only if all pathways from the
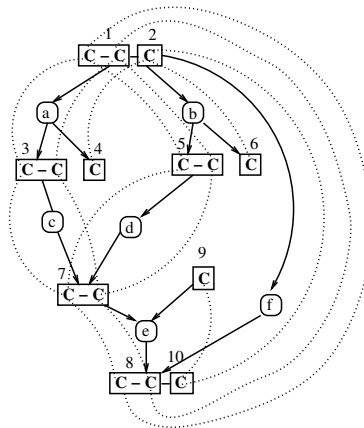
Figure 1: Example of the equivalence classes of fragments in a metabolic network consisting of six reactions $a - f$. The carbon maps are shown by dotted lines (one dotted line between a substrate carbon c and a product carbon c' for each reaction producing c' from c). The equivalence classes are $\{1, 3, 5, 7, 8\}, \{2, 4, 6\}, \{9\}$ and $\{10\}$. Fragments 3 and 5 are equivalent to fragment 1 as there exists only single reactions, a and b, producing 3 and 5 from 1 (Lemma 1). Fragments 7 and 8 are equivalent to fragment 1 as in every pathway producing 7 and 8 fragment 1 is transferred to 7 and 8 intact and every pathway has the same carbon maps from 1 to 7 from 1 to 8 (Lemma 3).

fragments of external nutrients to F' contain F and carbons of F travel always intact and similarly oriented to F'. An example of the equivalence classes is given in the Figure 1. For flux estimation, these equivalences serve in several roles: First, fragment isotopomer distributions of subpools of the junction can differ only if the fragment is not equivalent with any fragment in its reactants. It is only those fragment isotopomers that can potentially induce linearly independent constraints to the fluxes. This gives us possibilities to remove redundant equations from the flux system to be solved.

Second, we can use interchangeably any isotopomer measurement of two equivalent fragments, thus enabling us to form balance equations in junctions where adjacent metabolites are poorly or not at all measured. As by Lemma 3 the equivalence classes may extend beyond junctions, the technique improves the propagation methods described in [RRM$^+$03].

In this paper, these equivalences are utilized in yet another way. Namely, they turn out to be powerful tools for selecting a small subset of metabolites to be measured so that the fluxes of a metabolic network can still be discoved from the measurements.

4 Measurement optimization in the positional enrichment case

Using the linear system consisting of equations (1) and (4) for solving the fluxes ρ_j is not straightforward. There are several problems. First of all, the system should be of full rank to give point solution instead of just some linear constraints for the fluxes. Full rank means

that there should be sufficiently many linearly independent equations. In extreme case it is possible that full rank is not achievable at all (for example, if one-carbon metabolite has large number of producers.)

The independence of equations depends on the actual values of the isotopomer abundances which in turn depend in an intricate fashion on the isotopomer abundances of the external nutrients and on the actual fluxes in the steady state under consideration. Hence the quality of the equation system obtained depends on the measurements performed as well as on the ^{13}C labeling patterns used for external nutrients.

On the other hand, the measurements are expensive and tedious. In (tandem) mass spectrometer it is necessary to separate metabolites in the cell extract and develop for each metabolite or metabolite group a specific experimental protocol. In NMR it might be necessary to develop specific experiments for different metabolite groups or to separate metabolites prior to the NMR experiments to obtain sufficient signal resolution as structurally similar compounds often overlap in NMR spectra. Therefore there is a need to design optimized measurement strategy that minimizes the experimental effort. Here the equivalence concept of Section 3 can be utilized: measuring more than one representative of an equivalence class does not add new information; the distribution observed for one member of the class can be used also in association with the others to write equations (4).

In the rest of the paper we will consider the measurement optimization problem in a special case (positional enrichment) satisfying rather strong but experimentally justified conditions. Even in this case the optimization problem turns out to be computationally hard but still tractable in practice. Solutions to the problem given below can be used to guide iterative experiment planning process towards small set of metabolites to measure in its early stages. Our computational techniques can also be generalized for less constrained situations. Hence the present exercise is of wider interest.

4.1 Optimization problems

In the so-called positional enrichment measurements of ^{13}C, one observes the ^{13}C isotopomer distribution of an individual carbon c_h of a metabolite M which simply is the relative abundance of ^{13}C in c_h. NMR is often [FS98, MdGW$^+$96, SNP$^+$99, SPB$^+$99, WSdGM97] able to deliver such data for (some) individual carbon locations of some metabolites. Thus the availability of positional enrichment data for the carbons of some metabolites in our network is a reasonable first approximation of what can be measured.

Formally, we assume that one ^{13}C positional enrichment measurement gives for a metabolite $M = \{c_1, \ldots, c_k\}$ the ^{13}C labeling frequencies $P_{M|c_h}(1)$ and $P_{M|c_h}(0)$ for each carbon c_h of M. Here

$$P_{M|c_h}(1) = \sum_{b \in \{0,1\}^{|M|}:b_h=1} P_M(b),$$

i.e., the full isotopomer distribution of M marginalized to $b_h = 1$.

With the positional enrichment data available for all metabolites M, we can infer by

Lemma 1 the positional enrichment frequencies $P_{M_{ij}|c_h}(1)$ for all subpools M_{ij} of all junction metabolites M_i. Then the generalized isotopomer balances (4) get the form

$$\sum_{j=1}^{n} \alpha_{ij} v_j P_{M_{ij}|c_h}(1) = \beta_i P_{M_{i0}|c_h}(1). \tag{5}$$

Note here, that the corresponding equation for $b_h = 0$ is linearly dependent of equations (1) and (5) as $P_{M|c_h}(0) = 1 - P_{M|c_h}(1)$.

Based on the positional enrichment data, we can thus write $|M_i|$ new equations, in addition to the mass balance (1), for each junction metabolite M_i. This is the strongest system of equations we can hope for to get based on positional enrichment measurements. (Note, however, that there is no guarantee that this system would allow solving all the fluxes: the system may be underdetermined in some junctions and overdetermined in some others.)

Now it should be clear that because of the equivalence of carbons of different metabolites, measuring all metabolites may sometimes be redundant. Already some subset would allow us to write the full set of $|M_i|$ equations (5) for each junction M_i. This leads to the following optimization problem:

Problem 1 (Positional enrichment measurement (PEM) minimization problem). Given a metabolic network $G = (\mathcal{M}, \mathcal{R})$, find a smallest set of metabolites to measure for positional enrichment data such that, noting the equivalences of the carbons, it is possible to write a full set of $|M_i|$ equations (5) for each junction metabolite M_i of the network G.

4.2   Solution by set covering techniques

By its combinatorial nature, the PEM minimization is a variant of the well-known set cover problem. To see this, observe that an equation (5) can be written as soon as we know $P_{M|c_h}(1)$ for all subpools M_{ij} of M_i. This is the case when we have measured for each M_{ij} some carbon c_t of some metabolite M_r such that $M_r|c_t \equiv M_{ij}|c_h$. We now say that measuring M_r *covers* a subpool carbon $M_{ij}|c_h$ of a junction M_i if $M_{ij}|c_h \equiv M_r|c_t$ for some carbon c_t of M_r. Then we get the following technical formulation of PEM minimization:

Lemma 4. *A set K of metabolites is a solution of the PEM minimization if K is a smallest set such that it covers all subpool carbons $M_{ij}|c_h$ of all junction metabolites M_i.*

Then the NP-hardness and the inapproximability of our problem does not come as a surprise:

Theorem 1. *PEM minimization problem is not polynomial time approximable within a factor $c \log |\mathcal{M}|$ for some constant $c > 0$, unless $P = NP$.*

Proof sketch (details omitted). A polynomial-time approximation-preserving reduction from the set cover problem shows the NP-hardness [GJ79] and the inapproximability [ACV+99]. $\qquad\square$

The greedy approximation algorithm of the set cover problem canimmediately be applied to PEM minimization. The algorithm constructs a small set of metabolites to be measured by adding a new metabolite M_k to the set if M_k covers the largest amount of not yet covered junction carbons $M_{ij}|c_h$. This is repeated until all junction carbons $M_{ij}|c_h$ have a covering metabolite selected. When selecting the next metabolite to measure the number of new covered junction carbons can be weighted by the measurement cost of the metabolite.

The minimization algorithm has a well known performance guarantee, by factor $1 + \ln \delta$ where δ is the size of the set to be covered [ACV+99]. In our case $\delta \leq \Sigma_{m \in \mathcal{M}} |M|$ because the number of equivalence classes to be covered can not be larger than the number of different carbon locations of the metabolites in the network. Thus, we get the following theorem:

Theorem 2. *The greedy set cover algorithm finds for the PEM minimization problem a solution which is within a factor of* $1 + \ln(\Sigma_{M \in \mathcal{M}} |M|)$ *from the optimum.*

4.3 Solution by integer linear programming

Most metabolic network models used in flux estimation contain only a few dozens of metabolites and reactions. Thus, methods for obtaining the optimal metabolite sets might be of interest, even if they have exponential worst-case time complexity. One versatile approach is to model the problem as a mixed integer linear program (MILP), i.e., as a minimization of some linear objective function in a polytope, possibly requiring that some variables in the optimal solutions are integral (see [Mar01] for further details).

The objective function to minimize is the sum of costs of the metabolites that provide the maximum positional enrichment information (see Problem 1). Let $m_1, \ldots, m_{|\mathcal{M}|}$ be indicator variables whether or not the metabolite $M_i \in \mathcal{M}$ is measured. Let w_i be the cost of measuring the metabolite M_i. Thus, the objective function to be minimized is

$$\min_{m_1, \ldots, m_{|\mathcal{M}|}} \sum_{i=1}^{|\mathcal{M}|} w_i m_i$$

with the constraints $m_M \in \{0, 1\}$ for each metabolite $M \in \mathcal{M}$. The other constraints are as follows.

The requirement of Problem 1 that we have to write a balance equation for every carbon $c \in M_i$ can sometimes lead to sets of measured metabolites that are too laborious to measure in practice. In that case we can try to find less expensive solutions by requiring for each junction M_i only $k_i \leq |M_i|$ balance equations. The cost of this relaxation is that we might lose some independent balances constraining the fluxes in (5).

Let $x_{i,c}$ be the indicator variable that the balance equation can be written for a carbon $c \in M_i$. Then the constraints can be stated as

$$\sum_{c \in M_i} x_{i,c} - k_i \geq 0.$$

185

The balance equation can be written for a carbon $c \in M_i$ if for all corresponding carbons c_{ij} in the inflow subpools M_{ij} and for carbon c there is a measured metabolite M' with equivalent carbon c'. Let $R_{i,c}$ be the set of reactions with the carbon c as a product. Let $E_{i,c,j}$ be the set of indices p of metabolites that have a carbon equivalent with the inflow subpool carbon c_{ij} produced by the reaction ρ_j and let $E_{i,c}$ be the set of indices p of metabolites that have a carbon equivalent with c. Now

$$\sum_{p \in E_{i,c,j}} m_p - x_{i,c} \geq 0$$

and

$$\sum_{p \in E_{i,c}} m_p - x_{i,c} \geq 0$$

must hold.

Combining these constraints we obtain the following mixed integer linear program (we assume $k_i = 0$ for each non-junction metabolite M_i):

$$
\min_{m_1,\ldots,m_{|\mathcal{M}|}} \sum_{i=1}^{|\mathcal{M}|} w_i m_i
$$

$$
\text{s.t.} \sum_{c \in M_i} x_{i,c} - k_i \geq 0 \qquad \forall M_i \in \mathcal{M}
$$

$$
\sum_{j \in E_{i,c,j}} m_p - x_{i,c} \geq 0 \qquad \forall \rho \in R_{i,c}, c \in M_i \in \mathcal{M}
$$

$$
\sum_{j \in E_{i,c}} m_p - x_{i,c} \geq 0 \qquad \forall c \in M_i \in \mathcal{M}
$$

$$
m_i \in \{0,1\} \qquad \forall M_i \in \mathcal{M}
$$

$$
x_{i,c} \in \{0,1\} \qquad \forall c \in M_i \in \mathcal{M}
$$

The number of variables in the program is

$$|\mathcal{M}| + \sum_{M_i \in \mathcal{M}} |M_i|$$

and the number of inequalities is

$$|\mathcal{M}| + \sum_{M_i \in \mathcal{M}} \sum_{c \in M_i} (|R_{i,c}| + 1).$$

If the number of integer variables is too high, the integer linear program can be relaxed to linear program, i.e., the requirement of the variables being binary can be relaxed to the requirement that the variables have their values in the unit interval $[0,1]$. In fact, the constraints $x_{i,c} \in \{0,1\}$ can be relaxed to $x_{i,c} \in [0,1]$ without affecting the solution since the cost of the solution is minimized when all variables $x_{i,c}$ are either 0 or 1; the values

of $x_{i,c}$ can be replaced by $\lceil x_{i,c} \rceil$ without violating the constraints or increasing the cost of the solution. If the constraints $m_i \in \{0, 1\}$ are relaxed, then the obtained solutions are not necessarily integral, but the solution can be transformed to (possibly suboptimal) integral solution by randomized rounding techniques. For example, the following procedure can be applied:

1. Find the optimal solution for the linear program. If the values of all variables m_i are integral, output the solution and halt.

2. Choose one variable m_i with non-integral value randomly with probability proportional to its value and replace the occurrences of the variables m_i in the linear program by the constant 1 and go to the step 1.

4.4   Isotopomer data of full metabolites

It is possible to modify the above techniques of PEM minimization for other types of measurement data. In the other end of the spectrum is the full isotopomer data, i.e., the distributions $P_M(b)$ for full metabolites M. Similar set cover and MILP algorithms can be used in this case as well.

Note that, in the PEM case, the equivalence classes are the largest possible (because they correspond to the smallest possible, one-carbon, fragments) and hence their number is the smallest possible which leads to a small number of measurements. For the full metabolite isotopomer data the sets are smaller and hence a larger number of measurements may be necessary. However, full metabolite measurements give more more data per measurement and therefore allow writing more equations (4).

Finally we note that the relevant cases in which positional enrichment data is available only for some carbons of some metabolites in the network or more than one measurement per equivalence class is required can be handled with straight forward modifications to the techniques given above. For example, by setting high costs to metabolites overlapping in NMR spectrum one can test whether the separation of these metabolites is necessary or is it possible to derive the same isotopomer information from some other metabolites easier to measure. Techniques can also be used to find out how many equations (4) per each junction can be written given a set of metabolites whose positional enrichments are measurable.

5   Computational experiments

We tested the method for selecting the minimum set of metabolites to measure with the model of central carbon metabolism of *Saccharomyces cerevisiae* containing glycolysis, penthose phosphate pathway and citric acid cycle. Carbon mappings were provided by the ARM project (`http://www.metabolome.jp/`). The network consisted of 35 metabolites and 37 reactions of which five were bidirectional. Cofactor metabolites were

excluded from the analysis. The only carbon source of the network was glucose. There were five external products in the model. Eleven of the metabolites were produced by more than one reaction and thus formed junctions. Visualizations of the metabolic network used in the experiments and the fragment equivalence classes discovered are available at `http://www.cs.helsinki.fi/group/sysfys/`.

We assumed that either positional enrichment or full isotopomer information was available for every metabolite in the network. We also assumed that the effort needed to measure the isotopomer information is equal for every metabolite. We either required that a balance equation should be written for each carbon of a junction metabolite or settled (*number of producing reactions - 1*) of equations for each junction. To inspect the effect of improved partition of fragments to the equivalence classes introduced in [RRP$^+$05] to the metabolite selection we also tested our methods using equivalence classes of [RRM$^+$03] that will not go through junctions. The minimum sets of metabolites to measure were discovered by using greedy set cover heuristics and MILP programs with guaranteed optimal solution. MILP programs were solved by using publicly available lp_solve 5.1.1.3 package (`http://groups.yahoo.com/group/lp_solve/`).

The results of the tests are summarized in the Table 1. Solutions to the MILP programs were computed instantaneously with a PC with 2,4 GHz Pentium 4 processor, except for the fifth problem of the Table 1 that took 50 seconds to finish. From the results we see that the number of metabolites whose isotopomer distributions are needed to construct the balance equations is surprisingly low. According to our tests it is enough to measure only one fourth of all metabolites in the model. (The results are somewhat optimistic as the symmetries of the metabolites were not taken into account.) This can be taken as encouraging news to experimentalist who wants to estimate the fluxes of the metabolic network using method of [RRM$^+$03]. In our experiments the differences between the optimal solutions given by ILP solver and greedy heuristics were nominal. The sets of metabolites suggested by different experiments contained mostly the same metabolites, but there were also some variations.

6   Discussion

In this article we have introduced a new experimental planning problem in which one wants to maximize the amount of isotopomer information useful for metabolic flux estimation while minimizing the experimental effort needed to measure this information. The problem has great practical value as the measurement of isotopomer distributions of intermediate metabolites is time consuming, not to mention the time and money needed in the development of the measurement techniques. Thus the computational methods that can help experimentalist to concentrate on the minimal set of useful metabolites are well appreciated. We have studied the computational complexity of the introduced problem and given heuristic and exact algorithms for solving different variations of it. Our experiments suggests that this approach can find compact sets of metabolites whose isotopomer information will produce as many generalized balance equations tying the flux distribution as isotopomer information about every metabolite in the model would do.

measurement type	# of balances	algorithm	equivalences	# of measured metabolites
full	max	greedy	flow	9
full	max	MILP	flow	8
pos	max	greedy	flow	9
pos	max	MILP	flow	8
pos	min	MILP	flow	8
pos	max	greedy	no flow	11
pos	max	MILP	no flow	10

Table 1: Sizes of the minimum sets of metabolites whose isotopomer information is required to measure in different settings. First column (measurement type) indicates whether full isotopomer distribution (full) or positional enrichment information (pos) was assumed to be available, second column tells whether maximum amount (max) or at least *(# of producing reactions - 1)* isotopomer balance equations were required for each junction. Third column indicates whether greedy algorithm or MILP program was used to obtain the solution and fourth column whether a new flow analysis technique introduced of [RRP+05] and sketched in Section 3 or a previous method [RRM+03] was used in the construction equivalence classes. Finally, the last column gives the sizes of minimum sets of metabolites out of 35 whose isotopomer information is required.

To the authors' best knowledge the problem of selecting an optimal set of metabolites to measure for flux estimation has not been systematically studied before. In other areas of bioinformatics set cover techniques have been used in experimental design to select primers for polymerase chain reaction (PCR) experiments [DI99, PRWZ96], perturbation experiments to discriminate among the set of hypothetical gene interaction networks [ITK00] and microarray probes that allow one to recognize the targets in the sample [KRS+04].

The selection of the optimal set of metabolites to measure covers only one half of the experimental design of isotopomer tracing experiments. The other half is the selection of the labeling mixture of external nutrients. Together with the actual flux distribution the labeling of nutrients induce the isotopomer distributions of every metabolite in the network thus having strong effect to the linear independence of isotopomer balance equations. The selection of optimal labeling in the context of iterative flux estimation methods [SCNV97] is studied by Möllney *et al.* [MWKdG99] and by Araúzo-Bravo and Shimizu [ABS03].

An interesting future work is to look for methods that combine the contribution of this article with techniques that propose optimal labellings of nutrients in the direct flux estimation context of [RRM+03]. From experimental point of view it would also be useful to develop an experimental design method that tries to falsify the given model of metabolic network by proposing a cost effective set of measurements of metabolite fragments whose isotopomer distributions should be equal if the model were correct.

References

[ABS03]   M. Araúzo-Bravo and K. Shimizu. An improved method for statistical analysis of metabolic flux analysis using isotopomer mapping matrices with analytical expressions. *Journal of Biotechnology*, 105:117–133, 2003.

[ACV⁺99]   G. Ausiello, P. Crescenzi, Kann V., A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer, 1999.

[BSCL04]   L. Boros, N. Serkova, M. Cascante, and W-N. Lee. Use of metabolic pathway flux information in targeted cancer drug design. *Drug Discovery Today: Therapeutic Strategies*, 1(4):435–443, 2004.

[CN99]   B. Christensen and J. Nielsen. Isotopomer Analysis Using GC-MS. *Metabolic Engineering*, 1:E8–E16, 1999.

[DI99]   K. Doi and H. Imai. A Greedy Algorithm for Minimizing the Number of Primers in Multiple PCR Experiments. *Genome Informatics*, 10:73–82, 1999. Presented at the Genome Informatics Workshop 1999 December 14-15, 1999, Garden Hall, Yebisu Garden Place, Tokyo, Japan.

[FS98]   B. Follstad and G. Stephanopoulos. Effect of reversible reactions on isotope label redistribution Analysis of the pentose phosphate pathway. *European Journal of Biochemistry*, 252:360–371, 1998.

[GJ79]   M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W.H. Freeman and Company, 1979.

[ITK00]   T. Ideker, V. Thorsson, and R. Karp. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Pacific Symposium on Biocomputing 5*, pages 302–313, 2000.

[Kel01]   J. Kelleher. Flux estimation Using Isotopic Tracers: Common Ground for Metabolic Physiology and Metabolic Engineering. *Metabolic Engineering*, 3:100–110, 2001.

[KRS⁺04]   G. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert. Optimal robust nonunique probe selection using Integer Linear Programming. *Bioinformatics*, 20:186–193, 2004.

[LT79]   T. Lengauer and R. Tarjan. A Fast Algorithm for Finding Dominators in a Flowgraph. *ACM Transactions on Programming Languages and Systems*, 1:121–141, 1979.

[Mar01]   A. Martin. General Mixed Integer Programming: Computational Issues for Branch-and-Cut Algorithms. In Michael Jünger and Denis Naddef, editors, *Computational Combinatorial Optimization: Optimal and Provably Near-Optimal Solutions*, volume 2241 of *Lecture Notes in Computer Science*, pages 1–25. Springer, 2001.

[MdGW⁺96]   A. Marx, A. de Graaf, W. Wiechert, L. Eggeling, and H. Sahm. Determination of the fluxes in the central metabolism of Corynebacterium glutamicum by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnology and Bioengineering*, 49:111–129, 1996.

[MWKdG99]   M. Möllney, W. Wiechert, D. Kownatzki, and A. de Graaf. Bidirectional reaction steps in metabolic networks IV: optimal design of isotopomer labeling systems. *Biotechnology and Bioengineering*, 66:86–103, 1999.

[PRWZ96]   W. Pearson, G. Robins, D. Wrege, and T. Zhang. On the primer selection problem in polymerase chain reaction experiments. *Discrete Applied Mathematics*, 71:231–246, 1996.

[RRM$^+$03]   J. Rousu, A. Rantanen, H. Maaheimo, E. Pitkänen, K. Saarela, and E. Ukkonen. A method for estimating metabolic fluxes from incomplete isotopomer information. In *Computational Methods in Systems Biology, Proceedings of the First International Workshop, CMSB 2003*, volume 2602 of *Lecture notes in Computer Science*, pages 88–103, 2003.

[RRP$^+$05]   A. Rantanen, J. Rousu, E. Pitkänen, H. Maaheimo, and E. Ukkonen. Flow analysis of metabolite fragments for flux estimation. In *Proceedings of the 3rd International Workshop on Computational Methods in Systems Biology, CMSB 2005, Edinburgh*, 2005.

[SAN98]   G. Stephanopoulos, A. Aristidou, and J. Nielsen. *Metabolic engineering: Principles and Methodologies*. Academic Press, 1998.

[SCNV97]   K. Schmidt, M. Carlsen, J. Nielsen, and J. Viladsen. Modeling Isotopomer Distributions in Biochemical Networks Using Isotopomer Mapping Matrices. *Biotechnology and Bioengineering*, 55:831–840, 1997.

[SGH$^+$99]   T. Szyperski, R. Glaser, M. Hochuli, J. Fiaux, U. Sauer, J. Bailey, and K. Wütrich. Bioreaction Network Topology and Metabolic Flux Ratio Analysis by Biosynthetic Fractional ^{13}C Labeling and Two-Dimensional NMR Spectrometry. *Metabolic Engineering*, 1:189–197, 1999.

[SNP$^+$99]   K. Schmidt, L. Nørregaard, B. Pedersen, A. Meissner, J. Duus, J. Nielsen, and J. Villadsen. Quantification of Intracellular Metabolic Fluxes from Fractional Enrichment and ^{13}C–^{13}C Coupling Constraints on the Isotopomer Distribution in Labeled Biomass Components. *Metabolic Engineering*, 1:166–179, 1999.

[SPB$^+$99]   J. Shen, K. Petersen, K. Behar, P. Brown, T. Nixon, G. Mason, O. Petroff, G. Shulman, R. Shulman, and D. Rothman. Determination of the rate of the glutamate/glutamine cycle in the human brain by in vivo ^{13}C NMR. *Proceedings of the National Academy of Sciences of the United States of America*, 96(14):8235–8240, 1999.

[Szy95]   T. Szyperski. Biosynthetically directed fractional ^{13}C-labelling of proteinogenic amino acids. *European Journal of Biochemistry*, 232:433–448, 1995.

[WH99]   C. Wittmann and E. Heinzle. Mass Spectrometry for Metabolic Flux Analysis. *Biotechnology and Bioenginering*, 62:739–750, 1999.

[WMI$^+$99]   W. Wiechert, M. Möllney, N. Isermann, M. Wurzel, and A. de Graaf. Bidirectional Reaction Steps in Metabolic Networks: III. Explicit Solution and Analysis of Isotopomer Systems. *Biotechnology and Bioengineering*, 66:69–85, 1999.

[WMPdG01]   W. Wiechert, M. Möllney, S. Petersen, and A. de Graaf. A Universal Framework for ^{13}C Metabolic Flux Analysis. *Metabolic Engineering*, 3:265–283, 2001.

[WSdGM97]   W. Wiechert, C. Siefke, A. de Graaf, and A. Marx. Bidirectional Reaction Steps in Metabolic Networks: II. Flux Estimation and Statistical Analysis. *Biotechnology and Bioengineering*, 55:118–134, 1997.