Journal of Process Control 21 (2011) 1483–1492

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Identification of minimal metabolic pathway models consistent with phenotypic data

Zita I.T.A. Soons*, Eugénio C. Ferreira, Isabel Rocha

IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

ARTICLE INFO

Article history:
Received 1 December 2010
Received in revised form 29 May 2011
Accepted 30 May 2011
Available online 28 June 2011

Keywords:
Elementary modes
Generating vectors
Controlled random search
Model reduction
Metabolism
Escherichia coli

ABSTRACT

In metabolic systems, the cellular network of metabolic reactions together with constraints of (ir)reversibility of enzymes determines the space of all possible steady-state phenotypes. Analysis of large metabolic models, however, is not feasible in real-time and identification of a smaller model without loss of accuracy is desirable for model-based bioprocess optimization and control. To this end, we propose two search algorithms for systematic identification of a subset of pathways that match the observed cellular phenotype relevant for a particular process condition. Central carbon metabolism of *Escherichia coli* was used as a case-study together with three phenotypic datasets obtained from the literature. The first search method is based on ranking pathways and the second is a controlled random search (CRS) algorithm. Since we wish to obtain a biologically realistic subset of pathways, the objective function to be minimized is a trade-off between the error and investment costs. We found that the CRS outperforms the ranking algorithm, as it is less likely to fall into local minima. In addition, we compared two pathway analysis methods (elementary modes versus generating vectors) in terms of modelling accuracy and computational intensity. We conclude that generating vectors have preference over elementary modes to describe a particular phenotype. Overall, the original model containing 433 generating vectors or 2706 elementary modes could be reduced to a system of one to three pathways giving a good correlation with the measured datasets. We consider this work as a first step towards the use of detailed metabolic models to improve real-time optimization, monitoring, and control of biological processes.

1. Introduction

Most mathematical models used for optimization and control of biotechnological processes are relatively simple and generally ignore the complex interactions between the extracellular environment and the thousands of intracellular enzymes and metabolites. The lack of this information in bioreactor monitoring and control can have a profound impact on biological systems and lead to poor bioreactor control performance. Nevertheless, the use of methods based on large models in process monitoring and control is nowadays limited due to their complexity and the lack of appropriate methodologies. The challenge of the development of a large-scale modelling strategy that predicts cellular phenotypes is not yet solved and is addressed here in the view of bioprocess control.

Genome-scale stoichiometric models are currently the best approximation to a representation of the metabolic capabilities of the cell. However, stoichiometric models represent an infinite number of possible phenotypes and systems biology tools need to be applied such that the simulation matches the phenotypes in given conditions. Also, most tools in systems biology are designed for steady-state applications, whereas the aim of process control requires a dynamic approach. Although dynamics are not addressed explicitly in this work, the model is formulated such that it can be easily extended as such. Moreover, as a consequence of the complexity of the models, the computational intensity is high. The model simulations are therefore too slow for some applications, such as online monitoring and control. Several model reduction approaches can be used to simplify models for use in process control, like the use of lumped reactions, sensitivity analysis tools [1], singular perturbation theory [2], and elimination of the dynamics of some processes based on their time scales [3].

Tools that have the potential to solve some of the above problems may come from metabolic pathway analysis. Metabolic pathway analysis is the discovery and analysis of meaningful routes in metabolic networks. It is becoming increasingly important for assessing network properties and linking the cellular phenotype to the corresponding genotype. Amongst several concepts elementary mode (EM) analysis [4], extreme pathway analysis (EP) [5], and the concept of generating vectors (GVs) [6] are promising tools. The first two tools have been evaluated by several authors, amongst

* Corresponding author. Tel.: +351 253 604 422; fax: +351 253 604 429.
E-mail addresses: zita@deb.uminho.pt, zita.soons@gmail.com (Z.I.T.A. Soons), ecferreira@deb.uminho.pt (E.C. Ferreira), irocha@deb.uminho.pt (I. Rocha).

0959-1524/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2011.05.012

Nomenclature

EP	extreme pathway
EM	elementary mode
GV	generating vector
RMSE	root mean squared error

Enzymes

acont	aconitase
ackr	acetate kinase
acs	acetyl-CoA synthetase
act	acetate reversible transport via proton seaport
adk	adenylate kinase
akgd	2-oxogluterate dehydrogenase
ATPm	ATP maintenance requirement
ATPs4r	ATP synthase
citl	citrate lyase
CO2t	CO_2 transporter via diffusion
cs	citrate synthase
eno	enolase
fba	fructose-bisphosphate aldolase
fbp	fructose-bisphosphatase
fum	fumarase
g6pdh	glucose 6-phosphate dehydrogenase
gadp	glyceraldehyde-3-phosphate dehydrogenase
icdhyr	isocitrate dehydrogenase (NADP)
cl	isocitrate lyase
mdh	malate dehydrogenase
pgdh	6-phosphogluconate dehydrogenase
O2t	O_2 transport (diffusion)
pdh	pyruvate dehydrogenase
pgm	phosphoglycerate mutase
pfk	phosphofructokinase
pgi	glucose-6-phosphate isomerase
pgk	phosphoglycerate kinase
pox	pyruvate oxidase
ppc	phosphoenolpyruvate carboxylase
ppck	phosphoenolpyruvate carboxykinase
pps	phosphoenolpyruvate synthase
ptar	phosphotransacetylase
pts	phosphotransferase system
pyk	pyruvate kinase
rpe	ribulose 5-phosphate 3-epimerase
rpi	ribose-5-phosphate isomerase
sucd	succinate dehydrogenase
tala	transaldolase
thd5	NAD transhydrogenase
tkt1	transketolase
tkt2	transketolase

Metabolites

2PG	2-phosphoglycerate
3PG	3-phosphoglycerate
6PG	6-phosphogluconate
13DPG	1,3-bisphosphoglycerate
ACE	acetate
ACCOA	acetyl-coenzyme A
ACP	acetyl-phosphate
AKG	α-ketoglutarate
CIT	citrate
COA	coenzyme A
DHAP	dihydroxyacetonephosphate
E4P	erythrose-4-phosphate
F6P	fructose-6-phosphate
FDP	fructose-1,6-biphosphate
FUM	fumarate
G3P	glyceraldehyde-3-phosphate
G6P	glucose-6-phosphate
GLC	glucose
GLY	glyoxylate
ICIT	isocitrate
MAL	malate
OAA	oxaloacetate
PEP	phosphoenolpyruvate
PYR	pyruvate
R5P	ribose-5-phosphate
RU5P	ribulose-5-phosphate
S7P	sedoheptulose-7-phosphate
SCA	succinyl-coenzyme A
SUC	succinate
X5P	xylulose-5-phosphate

others [7,8]. Llaneras and Pico [9] provide a comprehensive review on these three and other concepts to generate and characterize the flux space. EMs analysis identifies all minimal functional pathways inherent to a metabolic network. EPs analysis identifies the minimal set of independent pathways through the network, which is a subset of the EMs. It is stated that, in large networks, the number of EMs can be several-fold greater than the number of EPs. However, since the computation of EPs requires splitting up all reversible internal reactions into forward and backward reactions, whereas the EMs analysis allows for reversible reactions, the number of reactions in the network for EPs analysis increases and the resulting number of EPs may not necessarily be much smaller than the number of EMs for a network containing reversible reactions. GVs are in turn a subset of the EPs. If one allows reversible reactions in the computation of the GVs, their number is lower than the number of EMs or EPs.

Pathway analysis for large metabolic networks has the problem of combinatorial explosion of possible routes across the networks. In many situations, particularly concerning EMs, many more pathways exist than necessary to construct all admissible flux distributions. Therefore, some of them can be taken as a generator set of the whole admissible region. Thus, it may not be necessary to use the full set of pathways for specific applications. Of particular interest is the subset of pathways describing a set of measured phenotypic data. The importance of these lays in the fact that the internal fluxes are not independently distributed but strictly constrained by external fluxes through the pathways at steady state [10]. A challenging task is how to select these pathways to describe a physiological state of interest. In literature, several approaches are described. Trinh et al. [11] give an overview of this problem.

Several authors concluded that EMs analysis is the preferred choice for finding possibly important routes in the majority of the applications [7,8,12]. The conclusion arises because EPs and GVs do not represent the complete set of simplest (genetically independent) routes within the metabolic network under investigation. Here, we evaluate these tools from a different perspective: for selecting a number of pathways that describe a particular phenotype. We compare the use of EMs, the largest set, and GVs, the smallest subset, with the aim of bioprocess optimization and control. We evaluate the modelling accuracy and computational intensity.

In recent years, several approaches that combine the use of EMs with experimental data have been used to predict cellular phenotypes and maximum production capabilities. Provost and Bastin [13] achieved a model reduction by deriving a dynamic model based on EMs. The model is based on the elimination of intracellular rates

to obtain a macroscopic model connecting substrates and products. The dynamic model, compatible with the underlying metabolic network, is built on these macro-reactions. The basic assumption is that the main dynamics is contained in the extracellular metabolites and that the intracellular metabolites are at quasi-steady state. This approach of combining EMs with experimental data is the basis for the model in this work, subject to several modifications [14].

In the sequel, two methods will be compared to select a limited number of pathways matching the phenotype in given conditions. The first method is based on ranking and the second on a controlled random search (CRS) algorithm. An overview of the work is shown in Fig. 1.

2. Model

2.1. Model based on pathway analysis

We adopt a model based on pathway analysis that is built from a stoichiometric model. The first step in our modelling approach (Fig. 1) is therefore the definition of the stoichiometric matrix N ($n_m \times n_v$) for the central carbon metabolism based on the metabolites, reactions, compartments (internal or external), and the corresponding enzyme directions (we refer to Section 2.2 for more details on the biological content and to Supplement 1 for the stoichiometric matrix itself):

$$N = \begin{bmatrix} N_s \\ N_\xi \end{bmatrix} \tag{1}$$

being N_s the $n_s \times n_v$ stoichiometric matrix containing the intracellular metabolites and N_ξ the $n_\xi \times n_v$ external stoichiometric matrix. The quasi-static approximation for mass balances of the internal metabolites is mathematically expressed by:

$$\frac{ds}{dt} \cong 0 \Rightarrow \begin{bmatrix} N_s \\ N_\xi \end{bmatrix} v - \begin{bmatrix} o \\ v_m \end{bmatrix} = o \tag{2}$$

where s denotes the concentrations of the intracellular metabolites, v is the vector of the specific reaction rates (called metabolic fluxes), and v_m is the specific uptake and excretion rates of the measured extracellular species.

Next, on the basis of a set of macro-reactions, the dynamical model of a bioreactor can be established as [2]:

$$\frac{d\xi}{dt} = Kr(t) + u(t) \tag{3}$$

where ξ is the vector with the concentrations of the external metabolites per biomass weight, $r(t)$ the vector of specific macro-reaction rates, and $u(t)$ the net exchange of the metabolites with the outside of the reactor. The stoichiometric matrix K of the macro-pathways reads [13]:

$$K = N_\xi \cdot E \tag{4}$$

where E is the $n_v \times n_{EMs}$ elementary modes matrix or the $n_v \times n_{GVs}$ generating vectors matrix. Each column of E represents one pathway (EM or GV). We chose to compute both using METATOOL 5.1.0 [15].

In continuous cultivations, there is no accumulation of metabolites and Eq. (3) can be simplified to:

$$K \cdot r = -u \tag{5a}$$

The exchange rates with the environment then become equal to the measured specific rates:

$$v_m = -u \tag{5b}$$

Any steady state flux pattern can be expressed as a non-negative linear combination of pathways. In overdetermined systems the matrix K is not-invertible. Some authors applied the Moore–Penrose inverse to calculate the rates r through the pathways [16]. An issue, not addressed in some literature on the calculation of the pathway rates, is that of reversibility of pathways. A pathway is considered reversible if all its reactions are reversible. Conversely, pathways containing one or more irreversible reactions are irreversible. As a consequence, the pathway rates of the irreversible modes should be greater than or equal to zero.

Poolman et al. [16] tackled this by simply removing the columns of K that lead to negative rates in irreversible modes and recalculating the assignment. Schwartz and Kanehisa [17] took into account the reversibility constraints in a quadratic programming problem to calculate the pathway rates. Another method, based in linear programming, is the concept of the α-spectrum [18,19]. The α-spectrum encloses all possible solutions, but is not intended to find a reduced set of modes. In this work we compute the rates r using a non-negative least squares algorithm [20], since all pathways are irreversible. Note that our approach can easily be extended with reversible pathways, in case they are present, by splitting up these reversible pathways into two irreversible pathways.

2.2. Stoichiometric model

The network model was reconstructed to represent *Escherichia coli* growing on glucose minimal media. The starting point was the reduced model for the central carbon metabolism of *E. coli* [21]. This model was modified in the following way: biomass formation was modelled by acknowledging the metabolic drain from the central metabolic pathways [22]; and oxidative phosphorylation was lumped [23]. The resulting model contains glycolysis, pentose phosphate pathway, TCA cycle, anaplerotic reactions, biomass formation, oxidative phosphorylation, maintenance energy, and membrane transport reactions. Energy requirements for biomass formation and energy production are also included. The model contains 45 metabolites and 48 reactions, of which 23 are irreversible. The number of degrees of freedom is 10 (there are linearly dependent balance equations). The metabolic network is shown in Fig. 2.

The distinction between balanced and unbalanced metabolites in the computation of the pathways is based on the classification as intracellular and extracellular metabolites. Therefore, the following specific rates have been defined for the extracellular metabolites:

$$v_m = [q^m_{GlcEX}, \mu, q^m_{AceEX}, q^m_{CO_2}, q^m_{O_2}] \tag{6}$$

where q^m_{GlcEX} is the specific glucose consumption rate, μ is the specific growth rate, q^m_{AceEX} is the specific acetate production or consumption rate, $q^m_{CO_2}$ the specific carbon dioxide evolution rate, and $q^m_{O_2}$ the specific oxygen uptake rate. Note that, in continuous cultivation, the uptake and excretion rates are equal to the exchange rates (Eq. (5)); in batch cultivation, in accordance to the quasi-static approximation, i.e. constant specific uptake and excretion rates during exponential growth, these specific rates are equal to the specific consumption or accumulation of glucose, biomass, and acetate.

2.3. Experimental data

We expected that all observed phenotypes could be represented by a non-negative linear combination of the pathways. However, often, the opposite case is encountered [10]. As a measure of the ability of the stoichiometric model to represent a particular phenotype, we computed the sum of squared errors of the specific rates in Eq. (6) normalized by the specific growth rate (SSE/μ) for 21 datasets. The specific rates from the model are computed using non-negative least squares as explained in Section 2.1.

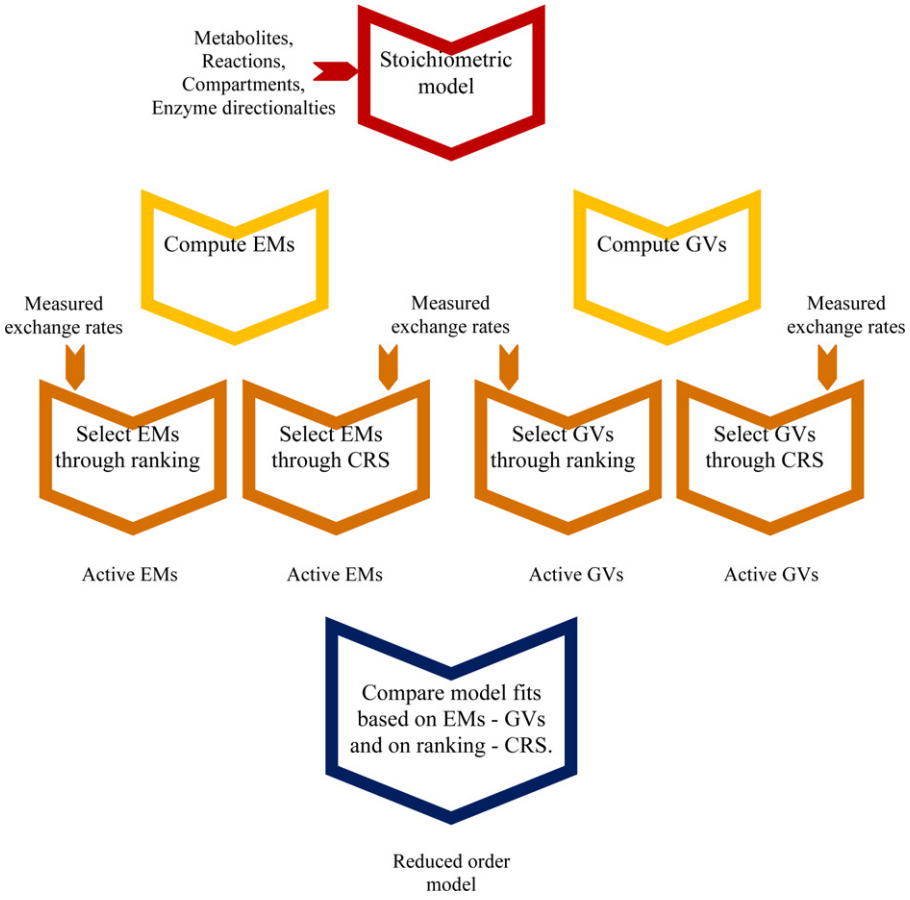

Fig. 1. Procedure for selection of a subset of pathways.

As a result, Fig. 3 shows that SSE/μ is nonzero for several datasets. This can happen if the network model is somewhat simplified, misses important pathways, or if the experimental data contain measurement errors. Here, it was found that, in most cases, the carbon balance was not closed. We refer to Supplement 2 for the computation of the carbon balance and for a figure showing the carbon balance of the datasets. Therefore, before attempting to select pathways, the experimental data should be carefully considered. As an alternative means to seek for inconsistencies between model and measurements, the calculability analysis proposed in [24] can be used. We chose three datasets showing different phenotypes to illustrate our method of selecting pathways. Those datasets represent a continuous cultivation at a specific growth rate of 0.1 h^{-1} [25], in which acetate formation was absent; a continuous cultivation with a dilution rate of 0.4 h^{-1} [26], where acetate was formed but in very small amounts; and a batch cultivation [27], where acetate formation was considerable. The carbon recovery was 108% for dataset [25], 78% for dataset [26], and 100% for dataset [27].

3. Selection of pathways

3.1. Objective function

Computation of the pathways using METATOOL gives 2706 EMs, of which 1622 is biomass producing modes and 433 GVs. A way to select the pathways that describe a particular phenotype with biomass production could be based on the biomass yield on glucose and oxygen, e.g. by selecting the pathway with the highest yield [23] or by selecting a single pathway or a linear combination of pathways close to those two experimentally measured yields. Supplement 3 shows an example of such method for our model using a particular dataset. Song and Ramkrishna [10] chose a fixed number of GVs based on yield analysis using quadratic programming.

Here, however, we aim to select a realistic subset of pathways (corresponding to a subset of K in Eq. (3)) that matches an observed phenotype considering all possible pathways (for instance also non-biomass producing modes), and also model size, and efficiency of the pathways. Hence, we reduce the number of pathways on the basis of an objective function that takes into account these aspects.

In general, increasing the model size (or the number of selected pathways) is likely to improve the estimation errors. On the other hand, our assumption is that only a small number of pathways are active under defined process conditions. Besides, in our search for a biologically meaningful subset, we think that the more efficient pathways, in terms of investment in enzymes, are more likely to be active in practice. Carlson [28] supports the view that inexpensive pathways (in investment in enzymes) have preference during nutrient limitation.

The proposed objective function therefore is a trade-off between the error *RMSE* and the investment costs required to establish the selected pathways *IC*:

$$J = RMSE + c \cdot IC \tag{7}$$

where c is a factor to weight the importance of investment costs against the actual error. The choice of the weighting factor c influences the selection of the number and index of the pathways. Since the main contribution for the objective function should be the *RMSE*, we chose the following value: $c = 5 \times 10^{-4}$. The effects of different values of factor c are discussed in Section 4.3. *RMSE* denotes

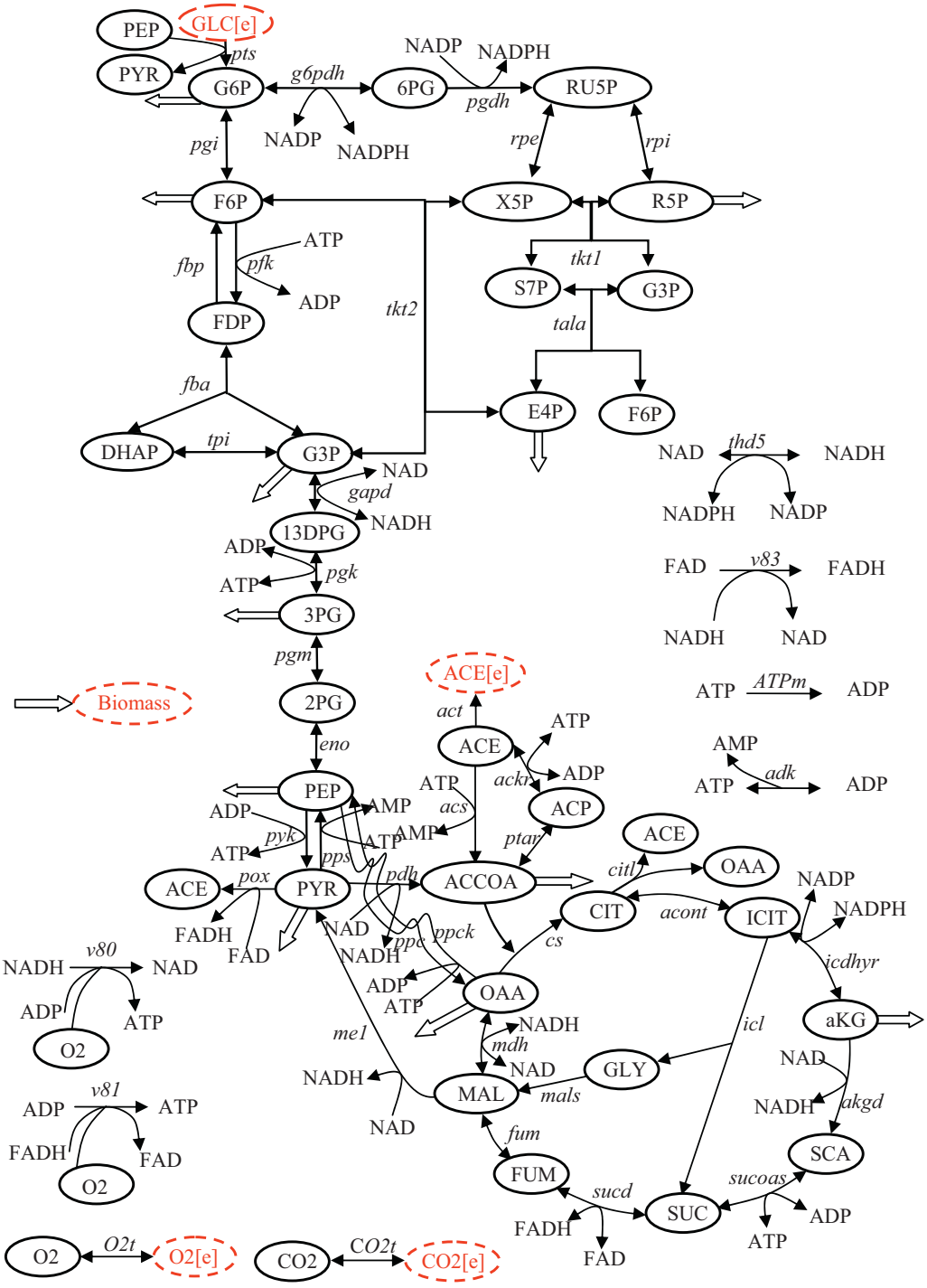

Fig. 2. Metabolic network for the central carbon metabolism of *E. coli* with glucose as the sole carbon source. The external metabolites are shown in a dotted red oval.

the weighted root mean squared error between the simulated (v_{ms}) and experimental (v_{me}) uptake and excretion rates:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum \left(\frac{v_{me} - v_{ms}}{W} \right)^2} \tag{8}$$

where n is the number of measurements and w is the weights, computed as the absolute value of v_m. The weight for acetate is set to unity to avoid dividing by zero when acetate is absent.

Model size (the number of selected pathways) and efficiency (the investment required to establish the pathways, that is, to produce the enzymes) are simultaneously evaluated in one criterion: investment costs *IC*. It is calculated as the number of nonzero components of the selected pathways E_k times their rates r_k:

$$IC = K(E_k \cdot r_k) \tag{9}$$

The pathways are normalized for glucose uptake. In general, Eq. (9) penalizes the selection of more pathways and inefficient pathways. We are now left with the problem of selecting the best subset of overall pathways K and computing their corresponding rates, which minimizes the objective function Eq. (7). Hereto we compare two algorithms: ranking (Section 3.2) and a controlled random search (Section 3.3).

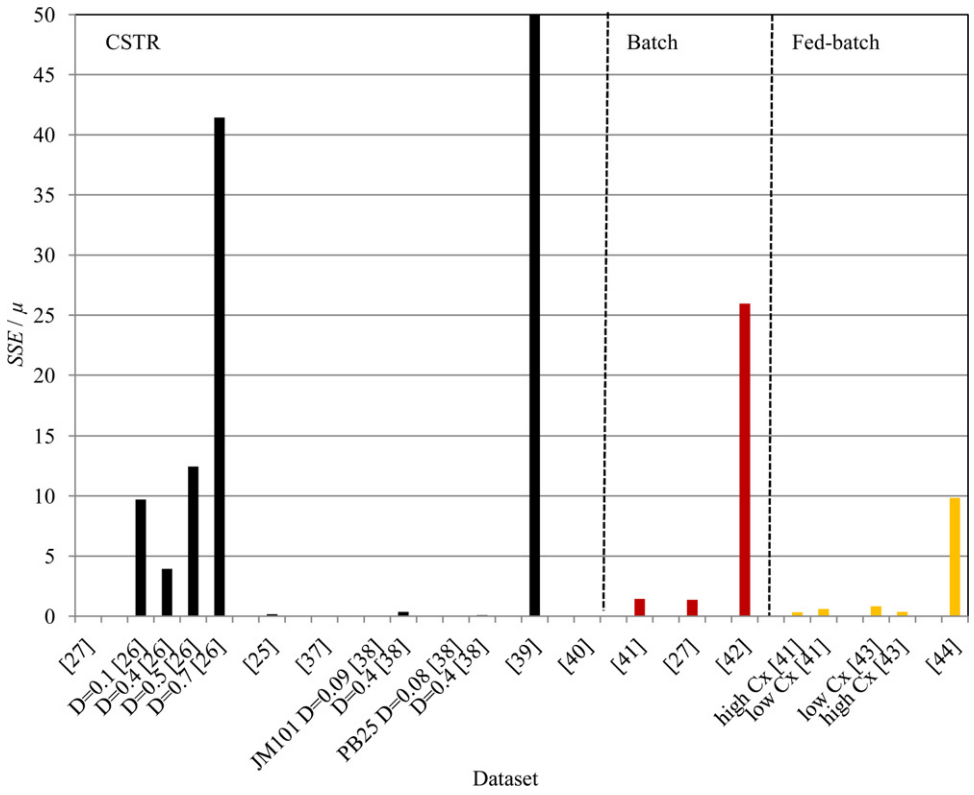

Fig. 3. Measure of the ability of the stoichiometric model to represent the experimental datasets from literature by a non-negative linear combination of all pathways in K (Eq. (5)). The measure consists of the sum of squared errors (*SSE*) in specific rates normalized by the specific growth rate (μ). The dilution rate (D) is given in h^{-1} [25–27,37–44].

3.2. Pathway ranking

The approach "ranking of EMs" was developed based on the idea of adding one column of K (representing the macro-reaction of one EM) to the current subset of K to give the largest improvement of the objective function until a minimal objective function is found. The algorithm uses this approach to expand the model, starting with a single term which minimizes the objective function. As an example, the ranking algorithm would select three out of the 433 original pathways for the given dataset in Fig. 4. Supplementary 4 contains the MATLAB code for the ranking and controlled random search algorithms.

3.3. Controlled random search algorithm

Price [29] developed a CRS procedure, which searches for global minima in an iterative procedure. A drawback of this method is the computational time; nevertheless, it is more efficient than a pure random search. This procedure is applied here to select a limited number of pathways and their index from K on the basis of the objective function. Using a particular phenotypic dataset, the specific rates v_{me} are calculated from the rates in literature and the following steps are performed:

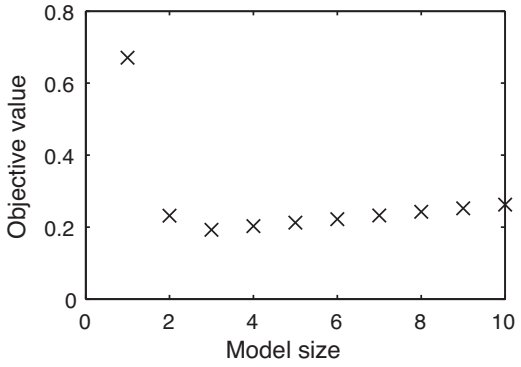

Fig. 4. Selection plot of the generating vectors on dataset [27] using the pathway ranking algorithm. Three GVs give the minimal objective value J (Eq. (7)).

1. First, a set of NT trial points (potential solutions) is randomly generated with $p = 2$ elements (number of pathways and their column index k) from the search domain V. The search domain V is defined by specifying limits on each of the p variables. In this work, the constraints were set to a maximum of ten pathways and the maximum index defined by the number of calculated pathways. Each trial point specifies which columns of K are used and then the rates r are computed using the non-negative least-squares algorithm. The error of this estimation is used to compute the objective function in Eq. (7) for that trial point and the results are stored in matrix A.
2. Then the search starts by generating for each iteration a new trial point as follows: a new point (TP) is generated by choosing $p+1$ random distinct points RP_1, RP_2, …, RP_{p+1} from the set of NT stored points and computing the centroid G of the n points RP_1, …, RP_p minus the last point RP_{p+1}:

$$TP = 2 \cdot G - RP_{p+1} \tag{10}$$

Provided that the new trial point TP satisfies the constraints, the goal function is evaluated (J_{TP}). In the original CRS algorithm, the centroid is given by:

$$G = \frac{1}{p}\sum_{j=1}^{p} RP_{p(j)} \tag{11}$$

In this work, we used an algorithm based on the CRS2 algorithm [30] that showed improved convergence properties by including the current best point TP_{min}:

$$G = \frac{1}{p+1}\left(TP_{\min} + \sum_{j=1}^{p} RP_{p(j)}\right) \quad (12)$$

3. The stored point M in matrix A with the greatest objective function value (J_M) is determined. J_{TP} is then compared with J_M. If $J_{TP} < J_M$, M is replaced by TP in A.

Step two and three are repeated until the stop criterion is satisfied (all penalties J in the stored matrix A are identical, the maximum J is smaller than a certain value, or the maximum number of function evaluations is reached).

4. Results and discussion

4.1. Ranking and CRS of EMs for dataset [26]

Both the ranking and the CRS algorithm reduced the original model containing 2706 EMs to a system based on one and the same mode for biomass growth; and another pathway for the original model of 433 GVs. The corresponding macro-reaction for the EM that connects the extracellular substrates and the end-products is:

$$\text{EM 1123}: \quad 1.00\,GlcEX + 1.75\,O_2 \rightarrow 0.0938\,Biomass + 2.03\,CO_2 + 0\,AceEX$$

with a rate of 4.59 h^{-1}. The original set of generating vectors did not contain this mode and another similar overall pathway is selected:

$$\text{GV 166}: \quad 1.00\,GlcEX + 1.76\,O_2 \rightarrow 0.0935\,\text{Biomass} + 2.04\,CO_2 + 0\,AceEX$$

with a rate of 4.61 h^{-1}. Simulation of the reduced model gives an appropriate match with the data ($J = 0.352$ in Eq. (7), Table 1). Note that acetate formation, which was almost 1000 times lower than the specific glucose consumption, is fitted to be zero in the reduced model. Application of the algorithms to a different dataset, in which acetate is produced in significant amounts, would lead to the selection of different mode(s), as is shown in the next section.

The selected EM is a pathway with an incomplete glycolysis (without *pgi*, but active pentose phosphate pathway) and an incomplete TCA cycle (without *mdh*, but instead the collective action of *me1*, *pps*, and *ppc* completes the TCA cycle). Although the (ir)reversibilities of the enzymes allow flux through those pathways, it is not likely that they are physiologically meaningful. The selected GV seems more biologically relevant, containing a full glycolysis and TCA cycle. With the aim of selecting biologically relevant pathways under given environmental conditions, it would be meaningful, before selecting pathways, to take into account more constraints imposed on the cell. As such, in addition to the imposed (ir)reversibilities of the enzymes, additional constrains can be assigned on the flux directionalities of reactions based on measurements in particular conditions; and subsequently, only the pathways that satisfy these additional constrains would be computed. For instance, these constrains can be assigned based on thermodynamic grounds (using intracellular metabolite concentrations) or on C13 labelling experiments.

4.2. Controlled random search for dataset [25]

The CRS algorithm was intended to find global minima. However, this is not guaranteed and the algorithm was run several times to find the best solution. All obtained sets of GVs have close objective function values, but not all the same. Convergence to global instead of local optima could be improved by increasing the number of trial points NT in the search (at the expense of increasing computational time). As an example, Fig. 5 shows that the iterative search converges to a set of two GVs. The corresponding macro-reactions are the following, with rates r of 7.4 and 3.34 h^{-1}:

$$\text{GV 88}: \quad 1.00\,GlcEX + 1.02\,O_2 \rightarrow 0.0717\,\text{Biomass} + 1.24\,CO_2 + 0.865\,AceEX$$

$$\text{GV 183}: \quad 1.00\,GlcEX + 3.1.0\,O_2 \rightarrow 0.0685\,\text{Biomass} + 3.10\,CO_2 + 0\,AceEX$$

Simulation of the reduced model gives an appropriate match with the data (Fig. 5D and Table 1), for the GVs (as well as for the EMs).

4.3. Comparison of methods and discussion

An overview of the results on the selection of a subset of pathways for the two selection methods and the two pathway modelling approaches for the three datasets is shown in Table 1. It can be observed that we obtained nearly the same *RMSE* for the reduced models compared to the original model containing all pathways. So, in given environmental conditions, selection of few pathways does not harm modelling accuracy. Besides, a presumably more realistic description of the metabolism is obtained compared to full models. However, in order to verify whether the pathways selected in this fashion are also active *in vivo*, additional experimental data would be required.

In addition, Table 1 shows that the CRS algorithm outperformed the ranking algorithm for a model size larger than one pathway. If the optimal model size is one, both methods select the same pathway, at the expense of a slightly larger computational time for the CRS. The ranking algorithm is based on expanding the model from a single and fixed "best" pathway with the "next best" pathway and so on. However, in reality the set of pathways giving the best objective value J does not necessarily include the best pathway. So, the ranking algorithm may not find the best set. In other words, the ranking approach has the problem of falling in local minima. Conversely, the iterative search procedure CRS may find the global minima. However, this is not guaranteed and the algorithm should be run several times to find the best solution. These findings are in line with other authors. Crampin et al. [31] already stated that the non-orthogonality of the matrix K means that the optimal subset of size $K+1$ is not necessarily the optimal subset K plus the "next best term" and that the selection process must therefore be iterative. Also Judd and Mees [32] stated: "It appears that finding the optimal model of size K is NP-hard – related to the feasible basis extension problem. If this is the case, we cannot expect to obtain the optimal solution easily."

If we compare the three datasets, the best results were obtained using dataset [25], followed by datasets [26,27]. Similar results were observed using the full pathway model (Fig. 3). As mentioned before, inspection of the experimental data suggested that the measurements of [26] are probably contaminated with an error. The respiration quotient ($RQ = CER/OUR$) was 0.69; when, for growth under aerobic conditions RQ is expected to be about 1.

Comparing the reduced model based on EMs and GVs, the *RMSE* values were the same (not shown). This is in line with our expectations, because all EMs can be reconstructed from the GVs by a non-negative linear combination. The objective values J may be slightly higher for GVs, because the best EM may not be present in the set of GVs, leading to slightly increased investment costs in the objective function. On the other hand, the total number of EMs can be several-fold greater than the number of GVs, especially for networks with many reactions and therefore the selection of

Table 1
Overview of the results for the selection methods "ranking" and "controlled random search" using the pathway modelling methods "elementary modes" and "generating vectors" on three datasets from literature.

Dataset	Pathways	Selection method	J	Reduced RMSE	Original RMSE	Reduced model size	Original model size
CSTR [25]	EMs	Ranking	0.062	0.045 (0.060)[a]	0.050 (0.058)	1	2706
		CRS	0.062	0.045 (0.060)	0.050 (0.058)	1	2706
	GVs	Ranking	0.066	0.049 (0.058)	0.050 (0.058)	3	433
		CRS	0.064	0.049 (0.058)	0.050 (0.058)	2	433
CSTR [26] $D = 0.4\,h^{-1}$	EMs	Ranking	0.352	0.335 (0.970)	0.348 (0.961)	1	2706
		CRS	0.352	0.335 (0.970)	0.348 (0.961)	1	2706
	GVs	Ranking	0.352	0.336 (0.969)	0.348 (0.961)	1	433
		CRS	0.352	0.336 (0.969)	0.348 (0.961)	1	433
Batch [27]	EMs	Ranking	0.164	0.147 (0.619)	0.160 (0.610)	3	2706
		CRS	0.164	0.148 (0.619)	0.160 (0.610)	2	2706
	GVs	Ranking	0.177	0.160 (0.610)	0.160 (0.610)	3	433
		CRS	0.163	0.148 (0.615)	0.160 (0.610)	2	433

[a] The RMSE values in parentheses are without weights in Eq. (8).

active pathways from a larger set of pathways becomes harder and more time-consuming; moreover, the computational intensity of the calculation of the EMs itself is a challenge for large metabolic networks, whereas GVs allow computation in polynomial time [6]. In summary, although EMs give the smallest objective values with the least pathways, we prefer GVs to EMs for the selection of a subset of pathways (from the many for large networks) that describe a particular phenotype.

The effect of changing the importance of investment costs against the sum of squared errors (factor c in the objective function Eq. (7)) depends on the datasets. The effect of efficiency is relatively more important for datasets with small *RMSE*, as can be seen in Table 1. In these cases, where the model is able to represent the dataset fairly (in contrast to cases with high experimental errors), the factor c was chosen such that efficiency does affect the choice of the pathways, though *RMSE* remains the main term contributing to the objective value. Although the selected pathways are different for different choices of c, the final impact of slight changes in c is minor, because there are many similar pathways (in particular elementary modes) that lead to similar results, as can be seen from Fig. S3 as well.

We used data from three particular experiments (3×5 measured metabolites) to illustrate the results. Interestingly, the approach is flexible enough to be of use, at a higher computational cost, if more datasets are simultaneously taken into account (e.g. replicate studies). In this way, the selection of pathways becomes less dependent on a single instance of measurements and therefore more reliable. Also, with the aim of developing a model that represents a wider range of phenotypes, the algorithm should be run on datasets from different conditions and sequentially merging the selected pathways into one set. For instance, representing the presence or absence of acetate formation or varying yields (which involves fluxes through different pathways).

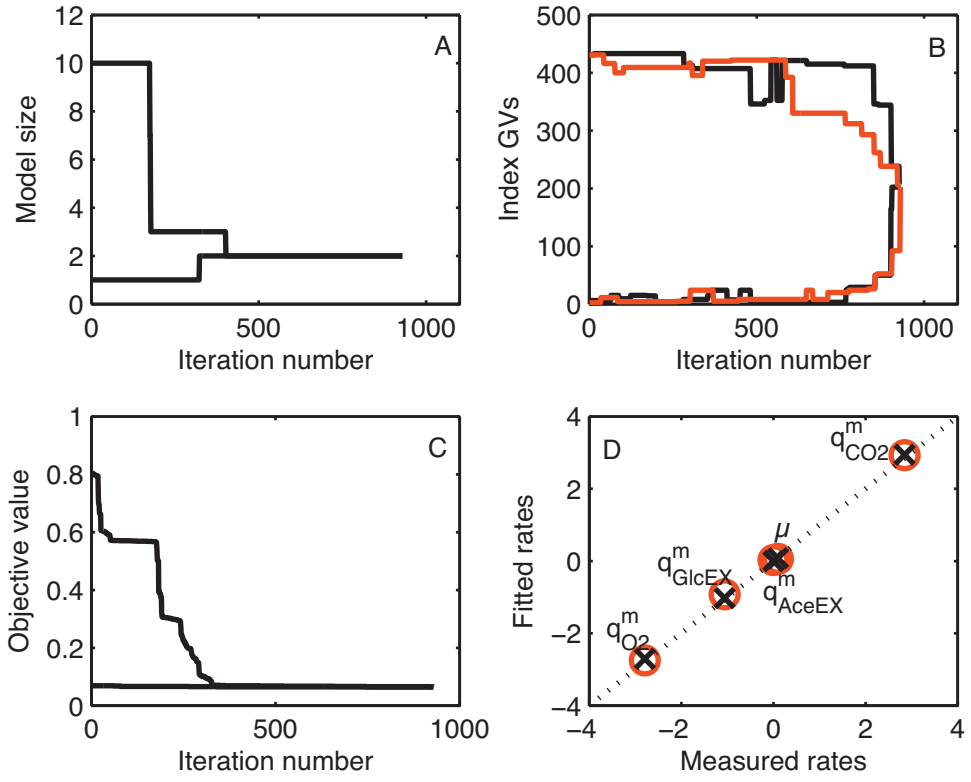

Fig. 5. Controlled random search to select a subset of GVs for dataset [25]. The lines in A, B, and C represent the maximum and minimum values within the dataset. (A) Model size, (B) index of the two selected GVs, (C) objective values J (Eq. (7)), (D) measured versus estimated specific rates for the reduced models based on GVs (X) and on EMs (O).

As mentioned before, most models for online optimization and control are relatively simple, often based on lumped reactions. The use of genome-scale stoichiometric models, on the other hand – currently the best representation of the cell – is often limited by the fact that these models do not allow a direct simulation of phenotypes and their use for the computation of EMs or GVs is impossible or too slow for online applications. They also face the challenge of estimating appropriate dynamics. The development of a reduced model in this work is an attempt to more accurately represent the capabilities of the cell in given conditions, while still being suitable for online applications.

The next step will be online monitoring of the evolution of the metabolites ξ (biomass, glucose, and acetate) through estimation of the pathway fluxes ($r(t)$ in Eq. (3)). One way would be to extend the (controlled random) search with the simultaneous selection of reaction mechanisms. A challenge, however, is the independent estimation of the reaction mechanisms and its model parameters. An alternative way is the use of observer-based techniques. The best way is currently under study. Ultimately we aim to optimize and control these fluxes in real-time towards enhanced production of target metabolites, for example through online optimization of feed profiles [33].

Another application where the selection of pathways could be used is for the redirection of the central metabolism towards the high-efficiency production of biochemicals [34,35]. Based on this analysis, those authors proposed targets for metabolic engineering towards an improved yield. Nevertheless, some genetic modifications such as gene deletions cause a decrease in growth rate, resulting in a decrease in productivity. To counterbalance the effect of genetic modifications on productivity, other technologies based on optimization of pathways and real-time control might be more efficient. We think that improvements in the production of biochemicals can be obtained by redirecting the metabolism of the cells towards the desired pathways by manipulating the environment of the cells using dynamic models based on pathway analysis.

In modelling metabolic systems through EMs, usually some metabolites are considered “external” in the sense that they are available for uptake or can be secreted from the cell. Those metabolites are the sources of the network and their concentrations are assumed to be buffered. Internal metabolites have to be balanced with respect to production and consumption at steady state. In many cases, there are biochemical reasons to treat a metabolite as balanced or unbalanced (based, for example on known membrane transporters). Often, however, the classification is ambiguous, since the buffered condition can also be assumed for metabolites that are not secreted. Given that the higher the number of external metabolites the larger the number of EMs, Dandekar et al. [36] propose a classification method of metabolites as external or internal that minimizes the number of EMs.

Ultimately, we wish to optimize and control in real-time towards the desired pathways. For this purpose, it is important to capture the essential process dynamics. In our next step of building a dynamic model, therefore, we will investigate methods to classify a metabolite as unbalanced or balanced based on time-scale separation [14], rather than on the physical presence of the metabolite, since the various reactions operate on different time scales, from milliseconds to hours or days.

The development of methodologies to improve real-time process optimization, monitoring, and control based on large-scale metabolic models has the potential to raise process efficiency and productivity. This work is a first step towards the use of metabolic models in real-time by presenting a two-step methodology to capture a large metabolic network by only a small number of pathways under defined process conditions.

Role of funding source

The funding sources had no involvement in the presented work.

Acknowledgments

We would like to thank Rui Oliveira (UNL) and Kiran Patil (EMBL Heidelberg) for the helpful discussions on elementary modes and the manuscript. The authors are grateful to the FCT – Portuguese Science Foundation for the financial support obtained under the scope of the MIT-Portugal program (MIT-Pt/BS-BB/0082/2008) and for the postdoctoral research grant of Zita Soons (SFRH/BPD/44180/2008).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jprocont.2011.05.012.

References

[1] I. Smets, K. Bernaerts, J. Sun, K. Marchal, J. Vanderleyden, J.F. Van Impe, Sensitivity function-based model reduction. A bacterial gene expression case study, Biotechnol. Bioeng. 80 (2002) 195–200.
[2] G. Bastin, D. Dochain, On-line Estimation and Adaptive Control of Bioreactors, Elsevier, Amsterdam, 1990.
[3] J.E. Haag, A.V. Wouwer, P. Bogaerts, Systematic procedure for the reduction of complex biological reaction pathways and the generation of macroscopic equivalents, Chem. Eng. Sci. 60 (2005) 459–465.
[4] S. Schuster, T. Dandekar, D.A. Fell, Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, Trends Biotechnol. 17 (1999) 53–60.
[5] C.H. Schilling, D. Letscher, B.O. Palsson, Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective, J. Theor. Biol. 203 (2000) 229–248.
[6] C. Wagner, R. Urbanczik, The geometry of the flux cone of a metabolic network, Biophys. J. 89 (2005) 3837–3845.
[7] S. Klamt, J. Stelling, Two approaches for metabolic pathway analysis? Trends Biotechnol. 21 (2003) 64–69.
[8] J.A. Papin, J. Stelling, N.D. Price, S. Klamt, S. Schuster, B.O. Palsson, Comparison of network-based pathway analysis methods, Trends Biotechnol. 22 (2004) 400–405.
[9] F. Llaneras, J. Pico, Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators, J. Biomed. Biotechnol. 2010 (2010) 753904.
[10] H.S. Song, D. Ramkrishna, Reduction of a set of elementary modes using yield analysis, Biotechnol. Bioeng. 102 (2009) 554–568.
[11] C.T. Trinh, A. Wlaschin, F. Srienc, Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism, Appl. Microbiol. Biotechnol. 81 (2009) 813–826.
[12] F. Llaneras, J. Pico, Stoichiometric modelling of cell metabolism, J. Biosci. Bioeng. 105 (2008) 1–11.
[13] A. Provost, G. Bastin, Dynamic metabolic modelling under the balanced growth condition, J. Process Control 14 (2004) 717–728.
[14] Z.I.T.A. Soons, E.C. Ferreira, I. Rocha, Selection of elementary modes for bioprocess control, Comput. Appl. Biotechnol. (Leuven) (2010).
[15] T. Pfeiffer, I. Sanchez-Valdenebro, J.C. Nuno, F. Montero, S. Schuster, METATOOL: for studying metabolic networks, Bioinformatics 15 (1999) 251–257.
[16] M.G. Poolman, K.V. Venkatesh, M.K. Pidcock, D.A. Fell, A method for the determination of flux in elementary modes, and its application to *Lactobacillus rhamnosus*, Biotechnol. Bioeng. 88 (2004) 601–612.
[17] J.M. Schwartz, M. Kanehisa, A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes, Bioinformatics 21 (2005) 204–205.
[18] S.J. Wiback, R. Mahadevan, B.O. Palsson, Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum, J. Theor. Biol. 224 (2003) 313–324.
[19] F. Llaneras, J. Pico, An interval approach for dealing with flux distributions and elementary modes activity patterns, J. Theor. Biol. 246 (2007) 290–308.
[20] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
[21] P.F. Suthers, A.P. Burgard, M.S. Dasika, F. Nowroozi, S. Van Dien, J.D. Keasling, C.D. Maranas, Metabolic flux elucidation for large-scale models using C-13 labeled isotopes, Metab. Eng. 9 (2007) 387–405.
[22] E. Fischer, N. Zamboni, U. Sauer, High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived C-13 constraints, Anal. Biochem. 325 (2004) 308–316.

[23] R. Carlson, F. Srienc, Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: Creation of overall flux states, Biotechnol. Bioeng. 86 (2004) 149–162.
[24] S. Klamt, S. Schuster, E.D. Gilles, Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria, Biotechnol. Bioeng. 77 (2002) 734–751.
[25] C. Chassagnole, N. Noisommit-Rizzi, J.W. Schmid, K. Mauch, M. Reuss, Dynamic modeling of the central carbon metabolism of *Escherichia coli*, Biotechnol. Bioeng. 79 (2002) 53–73.
[26] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, M. Tomita, Multiple high-throughput analyses monitor the response of *E. coli* to perturbations, Science 316 (2007) 593–597.
[27] E. Fischer, U. Sauer, A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*, J. Biol. Chem. 278 (2003) 46446–46451.
[28] R. Carlson, Systems biology – metabolic systems cost-benefit analysis for interpreting network structure and regulation, Bioinformatics 23 (2007) 1258–1264.
[29] W.L. Price, Controlled random search procedure for global optimization, Comput. J. 20 (1977) 367–370.
[30] W.L. Price, Global optimization by controlled random search, J. Optimiz. Theory Appl. 40 (1983) 333–348.
[31] E.J. Crampin, P.E. McSharry, S. Schnell, Extracting biochemical reaction kinetics from time series data, in: Proceedings of the Knowledge-based Intelligent Information and Engineering Systems, Pt. 2, vol. 3214, 2004, pp. 329–336.
[32] K. Judd, A. Mees, On selecting models for nonlinear time-series, Physica D 82 (1995) 426–444.
[33] A.P. Teixeira, C. Alves, P.M. Alves, M.J.T. Carrondo, R. Oliveira, Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control, BMC Bioinform. 8 (2007).
[34] J.C. Liao, S.Y. Hou, Y.P. Chao, Pathway analysis, engineering, and physiological considerations for redirecting central metabolism, Biotechnol. Bioeng. 52 (1996) 129–140.
[35] N. Vijayasankaran, R. Carlson, F. Srienc, Metabolic pathway structures for recombinant protein synthesis in *Escherichia coli*, Appl. Microbiol. Biotechnol. 68 (2005) 737–746.
[36] T. Dandekar, F. Moldenhauer, S. Bulik, H. Bertram, S. Schuster, A method for classifying metabolites in topological pathway analyses based on minimization of pathway number, Biosystems 70 (2003) 255–270.
[37] Q. Hua, C. Yang, T. Baba, H. Mori, K. Shimizu, Responses of the central metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts, J. Bacteriol. 185 (2003) 7053–7067.
[38] M. Emmerling, M. Dauner, A. Ponti, J. Fiaux, M. Hochuli, T. Szyperski, K. Wuthrich, J.E. Bailey, U. Sauer, Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*, J. Bacteriol. 184 (2002) 152–164.
[39] M.A. Hoque, H. Ushiyama, M. Tomita, K. Shimizu, Dynamic responses of the intracellular metabolite concentrations of the wild type and pykA mutant *Escherichia coli* against pulse addition of glucose or NH_3 under those limiting continuous cultures, Biochem. Eng. J. 26 (2005) 38–49.
[40] U. Sauer, D.R. Lasko, J. Fiaux, M. Hochuli, R. Glaser, T. Szyperski, K. Wuthrich, J.E. Bailey, Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism, J. Bacteriol. 181 (1999) 6679–6688.
[41] D. Riesenberg, V. Schulz, W.A. Knorre, H.D. Pohl, D. Korz, E.A. Sanders, A. Ross, W.D. Deckwer, High cell density cultivation of *Escherichia coli* at controlled specific growth rate, J. Biotechnol. 20 (1991) 17–28.
[42] M. Rahman, M.R. Hasan, T. Oba, K. Shimizu, Effect of rpoS gene knockout on the metabolism of *Escherichia coli* during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations, Biotechnol. Bioeng. 94 (2006) 585–595.
[43] G.L. Kleman, W.R. Strohl, Acetate metabolism by *Escherichia coli* in high-cell-density fermentation, Appl. Environ. Microbiol. 60 (1994) 3952–3958.
[44] I. Rocha, Model-based strategies for computer-aided operation of a recombinant *E. coli* fermentation, Thesis, 2003.