Computers and Chemical Engineering 35 (2011) 2366–2377

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Graph theory augmented math programming approach to identify minimal reaction sets in metabolic networks

Sudhakar Jonnalagadda [a], Balaji Balagurunathan [a], Rajagopalan Srinivasan [a,b,*]

[a] Institute of Chemical and Engineering Sciences, Agency for Science, Technology and Research (A*STAR), 1, Pesek Road, Jurong Island, Singapore 627833, Singapore
[b] Department of Chemical and Biomolecular Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore

ARTICLE INFO

Article history:
Received 18 September 2009
Received in revised form 5 May 2011
Accepted 9 May 2011
Available online 14 May 2011

Keywords:
Metabolic engineering
Systems biology
Strain improvement
Minimal cell
Flux Balance Analysis
Optimization

ABSTRACT

Bioprocesses are of growing importance as an avenue to produce chemicals. Microorganisms containing only desired catalytic and replication capabilities in their metabolic pathways are expected to offer efficient processes for chemical production. Realizing such minimal cells is the holy grail of metabolic engineering. In this paper, we propose a new method that combines graph-theoretic approaches with mixed-integer liner programming (MILP) to design metabolic networks with minimal reactions. Existing MILP based computational approaches are computationally complex especially for large networks. The proposed graph-theoretic approach offers an efficient divide-and-conquer strategy using the MILP formulation on sub-networks rather than considering the whole network monolithically. In addition to the resulting improvement in computational complexity, the proposed method also aids in identifying the key reactions to be knocked-out in order to achieve the minimal cell. The efficacy of the proposed approach is demonstrated using three case studies from two organisms, *Escherichia coli* and *Saccharomyces cerevisiae*.

1. Introduction

Bioprocesses are becoming increasingly common for production of chemicals, fuels and food ingredients due to the depletion of fossil resources and concerns over global warming (Demain, 2000). In general, industrially used microorganisms do not produce significant quantities of desired products since microorganisms are typically evolved for maximizing other cellular objectives (e.g. growth). In order to make bioprocesses economically viable, it is essential to engineer and improve microbial strains with enhanced yield of desired product (Stephanopoulos, 2002). Improved strains can be developed by targeted modification of cellular metabolism using genetic engineering techniques (Lee, Lee, & Kim, 2005). Traditional approaches to gene target selection for strain improvement largely rely on known literature of organisms. The large number of inter-connected cellular compounds and redundancies in metabolic networks make the gene target selection a difficult problem. Often the effects of even simple genetic modifications are beyond intuition; hence the success rate is low. The recent availability of genome sequences of organisms and other experimental information now enable new approaches based on genome-scale models of metabolism (Covert et al., 2001).

Constraint-based analysis, generally known as Flux Balance Analysis (FBA), of metabolic models estimates internal flux distribution and identifies metabolic bottlenecks for production of desired products. FBA also aids in redesigning the metabolism to achieve optimal phenotype that improves yield of the desired products (Price, Reed, & Palsson, 2004). Systems engineering approaches can also be employed to identify specific gene targets for strain improvement. One such optimization framework using mixed-integer linear programming (MILP) has been proposed by Burgard, Pharkya, and Maranas (2003) to identify genes which when knocked-out from the organism yields more desired product. However, identifying a gene knock-out strategy is combinatorial, consequently the computational time increases exponentially with the size of the problem (number of reactions in the model). As a simplification, Burgard et al. (2003) considered only the reactions in the central metabolism to achieve results within a reasonable time. A more practical approach using Genetic Algorithms for large scale systems that does not required such simplification has been proposed by Patil, Rocha, Förster, and Nielsen (2005) for the same purpose. Genetic Algorithm is a stochastic optimization approach; it cannot offer a guarantee for the global optimal solution. Besides identification of gene knock-out strategies, researchers are also interested in identifying genes/reactions not presented in the organism for insertion into its genome to obtain desired microbial properties (Pharkya, Burgard, & Maranas, 2004).

Recently, attempts are being made to develop cells with minimal functionality, containing only desired catalytic properties for chemical conversion and replication capabilities (Forster &

* Corresponding author at: Department of Chemical and Biomolecular Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore. Tel.: +65 65168041; fax: +65 67791936.
E-mail address: chergs@nus.edu.sg (R. Srinivasan).

0098-1354/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2011.05.006

Church, 2006). Such minimal cells are expected to be the most efficient machinery for production of specific chemicals (Pohorille & Deamer, 2002). Several approaches based on comparative genomics, genetic and biochemical tools have been proposed for creating such minimal cells (Glass et al., 2006; Hutchison et al., 1999; Mushegian & Koonin, 1996). Since functional information of the components of organism is not fully known, creating minimal cell from scratch is still in infancy. However, it is possible to design cells with minimal metabolic reactions through identification of minimal reaction set for desired chemical conversion and knocking out unnecessary reactions (Trinh, Unrean, & Srienc, 2008). Computational procedures have been proposed for identifying minimal reaction sets through analysis of metabolic models. Burgard, Vaidyaraman, and Maranas (2001) proposed a math programming based approach to find the minimal reaction sets under different uptake environments. Their study finds that minimal reaction sets are strongly dependent on medium constituents and cellular objectives. Besides the high computational complexity of their approach, it does not provide any clue on how a minimal cell can be achieved, *i.e.* the set of reactions that need to be knocked out. Trinh et al. (2008) used Elementary Flux Modes (EMFs) for analysis of different pathways in *Escherichia coli* and identified a set of seven reactions that when knocked-out leads to a minimal strain for efficient production of ethanol. Elementary Flux Modes essentially represent different independent pathways available for the cell to achieve cellular objectives. The approach used by Trinh et al. (2008) considers the reduction in available EFMs as a result of removal of reactions for selecting reactions to knock-out. Since their approach considers individual reactions and ignores interactions between different reactions, the solution is not guaranteed to be optimal.

Here, we propose a method to reduce the computational complexity of the MILP formulation for identifying minimal reaction set. Graph theory based approaches (Mah, 1990) have been used widely for solving chemical engineering problems such as process synthesis, Friedler, Tarjan, Huang, and Fan (1992), sensor location (Meyer, Le Lann, Koehret, & Enjalbert, 1994), plant layout (Jayakumar & Reklaitis, 1994), HAZOP analysis (Srinivasan & Venkatasubramanian, 1995), waste minimization and inherent safety analysis (Palaniappan, Srinivasan, & Halim, 2002), heat exchange network synthesis (Shivakumar & Narasimhan, 2002), process monitoring (Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003), water network design (Prakash & Shenoy, 2005), cause-and-effect analysis (Thambirajah, Benabbas, Bauer, & Thornhill, 2009), and process analysis (Preisig, 2009). The proposed method combines graph-theoretic approaches with the math programming framework. Instead of solving the whole problem monolithically, the proposed method solves the problem in a hierarchical manner starting from the primary uptake reactions as the top layer and proceeds iteratively, identifying essential reactions at each stage. A small MILP, with only a few binary variables, is solved for this purpose thus improving the optimization performance. Along with the reduced computational complexity, the proposed approach identifies the key reactions to be knocked-out from the cell in order to achieve the minimal cell for efficient production of desired chemicals. The efficacy of the proposed approach is demonstrated using three case studies from two different organisms, *E. coli* and *Saccharomyces cerevisiae*.

2. Math programming approach in metabolic engineering

2.1. Modeling metabolism

The metabolic profile of an organism is directly related to its phenotype. Hence modeling and analysis of metabolic networks play an important role in bioprocess development. The metabolic network of a given organism with N metabolites and M reactions can be represented as (Covert et al., 2001):

$$\frac{dX_i}{dt} = \sum_{j=1}^{M} S_{ij} v_j \quad i = 1, 2, \ldots, N \tag{1}$$

where X_i is the concentration of metabolite i, S_{ij} is the stoichiometric coefficient of the ith metabolite in the jth reaction, v_j is the flux (rate) of reaction j. The mass balance constraints arising from the steady-state assumption can be written as:

$$\sum_{j=1}^{M} S_{ij} v_j = 0 \quad i = 1, 2, \ldots, N \tag{2}$$

The stoichiometric modeling approach characterizes all feasible metabolic phenotypes (flux distributions) in the organism. Since the number of reactions is generally higher than the number of metabolites, multiple flux distributions can satisfy the system of equations in Eq. (2). Flux Balance Analysis (FBA) is the approach generally used to determine the metabolic phenotype. FBA uses linear programming to identify flux distribution while optimizing a flux (objective function) through a reaction (or linear combination of reactions). Growth rate is generally selected as the objective function in FBA. Mathematically, this can be represented as:

$$\begin{aligned} &\text{maximize } z = V_{biomass} \\ &s.t. \quad \sum_{j=1}^{M} S_{ij} v_j = 0 \quad i = 1, 2, \ldots, N \\ &v_j^{\min} \le v_j \le v_j^{\max} \quad j = 1, 2, \ldots, M \\ &v_j \in \Re \end{aligned} \tag{3}$$

here $v_j^{\min}$ and $v_j^{\max}$ are the minimum and maximum bounds on v_j, and $V_{biomass}$ is the flux through the biomass reaction (Edwards & Palsson, 2000). The reactions j also include transport reactions for uptake and secretion of metabolites by the cell.

2.2. MILP approach for minimal reaction set identification

Solving Eq. (3) results in a particular flux distribution that leads to maximum biomass. However, when engineering the cell, the objective of strain improvement is to improve the yield of a desired chemical. This can be achieved by modifying the metabolic reactions so as to divert most of the flux towards the chemical production while maintaining an adequate biomass. Two different approaches currently in practice are to knock-out some of reactions and insert heterogeneous reactions into the cells. Identification of genetic targets for knock-out or insertion is generally formulated as a MILP problem. In this approach, associated to each gene/reaction is a binary variable, with 1 indicating the presence/activation of reaction and 0 its absence/inactivation.

$$y_i = \begin{cases} 1 & \text{if reaction flux of } v_j \text{ is active} \\ 0 & \text{if reaction flux of } v_j \text{ is inactive} \end{cases} \tag{4}$$

where y_j is a binary variable corresponding to v_j. In order to identify gene knock-out targets, the optimizer finds a flux distribution that maximizes chemical production along with biomass by eliminating some reactions from the network (Burgard et al., 2003). Similar approach is used for identification of minimal reaction sets sufficient to make predefined quantity of biomass (Burgard et al.,

2001). The mathematical representation of the minimal reaction set identification problem is given as:

$$\begin{aligned} &\text{minimize } z = \sum_{j=1}^{M} y_j \\ &s.t. \quad \sum_{j=1}^{M} S_{ij} v_j = 0 \quad i = 1, 2, \ldots, N \\ &v_j^{\min} \cdot y_j \le v_j \le v_j^{\max} \cdot y_j \quad j = 1, 2, \ldots, M \\ &y_j \in \{0, 1\} \quad j = 1, 2, \ldots, M \\ &v_{biomass} = v_{\max} \\ &v_j \in \Re \end{aligned} \tag{5}$$

where $v_{\max}$ is the maximum biomass possible for a given environment (glucose and other nutrients).

Though the MILP approach has been reported to be successful in some cases, its solution time increases tremendously with increasing number of reactions. It is essential to improve the efficiency of optimizing procedure in order to employ this approach for the genome-scale models. In this paper, we propose the use of graph theory to reduce the search space and thus reduce the computational complexity.

3. Hybrid approach for identifying minimal reaction sets

The proposed approach explicitly accounts for the network structure of metabolic networks and exploits this to solve the minimal reaction set identification problem in two stages (see Fig. 1). As an overview, first, we classify the reactions into two classes: Essential and Extraneous. *Essential reactions* (ER) are those that are essential to the organism *i.e.* removal of these reactions makes the cell incapable of supporting a pre-specified rate of growth ($v_{\max}$) or chemical production. These reactions are part of the minimal set. *Extraneous reactions* (XR) comprise the reactions which are not necessary for the cell. So the minimal reaction set identification problem becomes one of classifying every reaction into one of these two categories. The MILP scheme described above approaches this monolithically considering all the reactions in the network simultaneously. We use graph-theoretic insights in Phase 1 to derive small sets of reactions that can be classified in isolation. Only the reactions in each set are classified using a small MILP. While this local analysis is adequate for a majority of reactions, some reactions cannot be classified in isolation. We call such reactions as indeterminate. *Indeterminate reactions* (IR) occur in the path of biomass or chemical production but may or may not be the in the minimal set, and hence may or may not be essential. Indeterminate reactions can be classified only by considering the network holistically. In our approach, these reactions are classified in Phase 2, again using the same MILP approach. But since the number of indeterminate reactions is typically small, the resulting MILP can be solved more efficiently. These steps are described in detail below.

Table 1
Set of reactions representing a simple metabolic network.

	Reaction formula
r1	A → B
r2	B → E
r3	C → D
r4	D → E
r5	E → F
r6	E → F
r7	D → G
r8	F + G → H + I
r9	A_ext → A
r10	C_ext → C
r11	H → Biomass
r12	I → I_ext

3.1. Phase 1

First, the reactions in the network are represented by a graph. As an illustration, consider the set of reactions shown in Table 1 and Fig. 2a. This simple network takes two substrates A and C (through reactions r9 and r10, respectively) and produces metabolite H that goes to biomass. A number of reactions are involved in

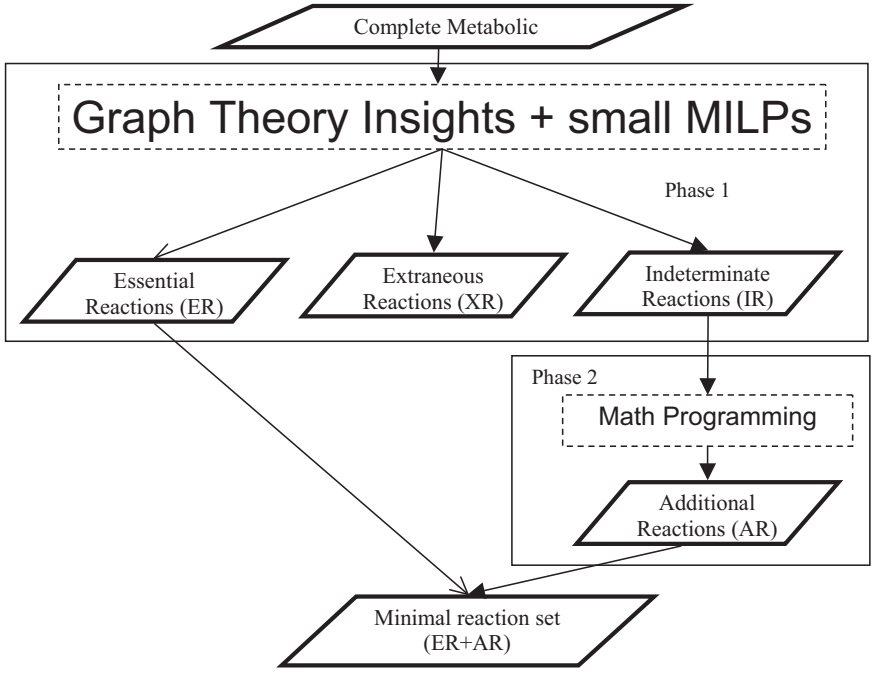

Fig. 1. Schematic representation of the proposed hybrid approach that combines graph theory insights with math programming to identify minimal reaction sets.

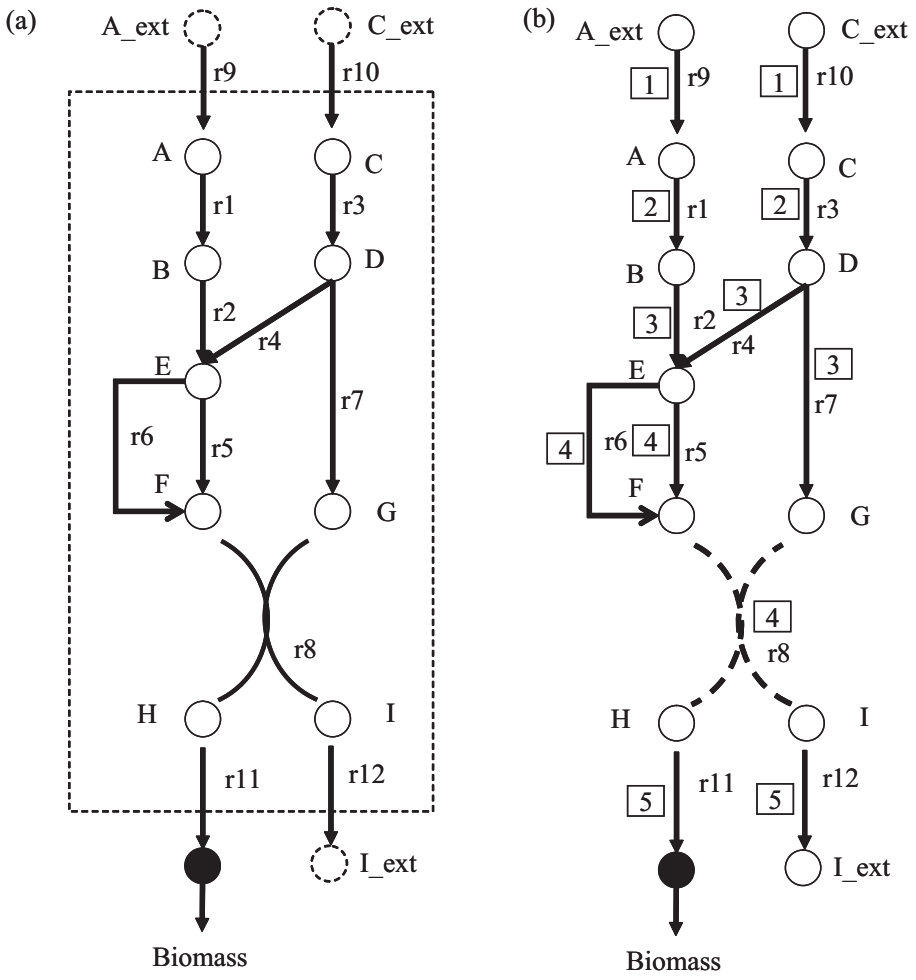

Fig. 2. (a) Sample metabolic network, and (b) its graph-theoretic representation with AND arcs represented by dash line.

this transformation. The dependencies between the reactions are first represented explicitly using a graph notation.

The metabolic network can be considered as an AND–OR graph, a well-known knowledge representation scheme (Braunschweig & Gani, 2002; Nilsson, 1980) that is used to represent relationships among the components of any system. In AND–OR graphs, each node depicts a specific component in the system. Further details of AND–OR graph are described in the Supplementary material. In the context here, each metabolite is represented as a node. Edges represent dependencies among nodes. Usually, the edges are directed and establish a parent–child relationship between the upstream and the downstream nodes. In metabolic networks, each arc represents a reaction, specifically one that converts the upstream metabolite(s) to the downstream metabolite(s). When a reaction has multiple reactants, there is a logical 'AND' relationship between the reactant metabolites, *i.e.* all the reactants have to be present for the reaction to occur. This logic is captured in AND–OR graphs using AND arcs. The same is the case with reactions producing multiple products. Consider reaction r8 in Table 1, which requires metabolites F and G to produce H and I. In the proposed scheme, this reaction would be depicted by an AND arc. Next, consider reactions r2 and r4 in Table 1. Both produce metabolite E, the former from B and the latter from D. The two reactions are depicted by two separate arcs both of which converge on the node representing metabolite E. The relationship between these two arcs can be considered to have a logical 'OR' nature since E can be produced through either of the reactions. Thus, OR edges depict parallel reactions sharing the same product. They can also depict parallel reactions sharing the same reactant (e.g. r5 and r6). Following this scheme, all the reactions in a metabolic network can be represented as a graph. Fig. 2b shows such the AND–OR graph representation for the set of 12 reactions in Table 1 and Fig. 2a. The interdependencies between the reactions in the metabolic network are thus effectively reflected in the graph structure.

The graph representation of the metabolic network allows the reactions to be classified into three classes – essential, extraneous, and indeterminate – mentioned above. In the example above, assume that the objective is to identify the minimal reaction set for producing metabolite H. Reactions r10, r3, r4, r7, r8, and r11 are essential since the network cannot produce biomass without any of these. Reaction r12 is also essential to ensure that there is no buildup of metabolite I. Reactions r9, r1 and r2 are extraneous since the network can produce metabolite H even without these (Metabolite E can be produced through the pathway r10, r3, and r4). Ambiguity arises while classifying reactions r5 and r6. Only one of these two reactions is required in the minimal reaction set. Classification of these two reactions cannot be performed in isolation; hence, these are labeled as indeterminate reactions in Phase 1.

Next, we describe an automated method for classifying the reactions using the graph representation. A *depth* can be associated with each node (metabolite) and arc (reaction) in the network. The depth d_j of a reaction r_j is derived from the depth of its reactants. Extracellular metabolites and primary uptake reactions (such as reactions r9 and r10 in Fig. 2) are considered to have depth 1. The depth of every other metabolite and reaction is assigned as an increment

over its predecessor's. If a metabolite is produced through multiple pathways, its depth is calculated based on that of its least deep predecessor. If reaction r_l requires as reactant metabolites that are products of reactions r_j and r_k, then

$$d_l = \min\{d_j, d_k\} + 1 \quad (6)$$

As illustration, the depth of all reactions in the network in Fig. 2b is indicated next to them within boxes. This procedure also breaks cycles, which are common in metabolic networks. For example, consider the cycle shown in Fig. 3 which is an extract from the well-known TCA cycle. The cycle starts with the conversion of two reactants, Acetyl-CoA and Oxaloacetate into products Citrate and CoA. Assume that the depth of r_A, the reaction producing Acetyl-CoA is d_j (since this reaction is upstream of the cycle, its depth would have been assigned earlier). Consider the depth d_k of reaction r_H producing Oxaloacetate. Since r_H is downstream of r_A, so from (6), depth of $r_B = d_j + 1$. Following the above procedure, all other depths would be assigned as shown in the figure. Thus the cycle has been converted into an open path by the depth assignment scheme using the directionality of the reactions.

This assignment of a depth to the reactions enables Phase 1 to exploit the dependency among reactions to reduce the complexity of the minimal reaction set determination problem. Instead of solving the whole network at once, the proposed method uses the above assigned depth to search the network in a breadth-first manner, starting from the primary uptake reactions (search-level $x = 1$) and progressing iteratively through reactions of increasing depth, until the bottom layer *i.e.* the secretion reactions. In each iteration (stage), all the reactions at a given depth are considered to be *independent* reactions (I). These independent reactions are analyzed in isolation and classified into essential, extraneous and indeterminate reactions using small MILPs as described below. In subsequent stage, reactions that are dependent on these reactions *i.e.* reactions whose reactants are products of the independent reactions in the previous stage are considered as *independent*. This is equivalent to incrementing the search-level $x \leftarrow x + 1$ and repeating. This procedure is continued until all the levels in the graph and therefore all reactions present in the network have been evaluated.

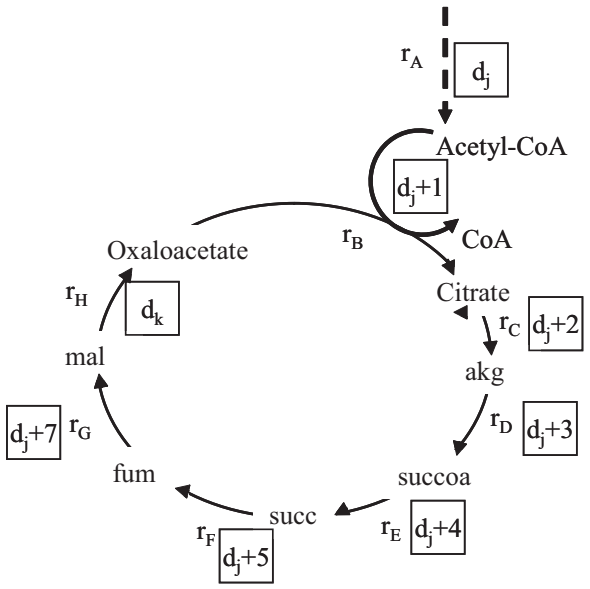

Fig. 3. Illustration of depth assignment for loops in the metabolic network.

The algorithm for classification of independent reactions into essential, extraneous or indeterminate reactions is shown in Fig. 4. The classification procedure uses the basic MILP formulation described in Section 2.2 but restricting its scope to the independent reactions at that search-level. Hence, binary valuables are associated with the independent reactions. The variables representing

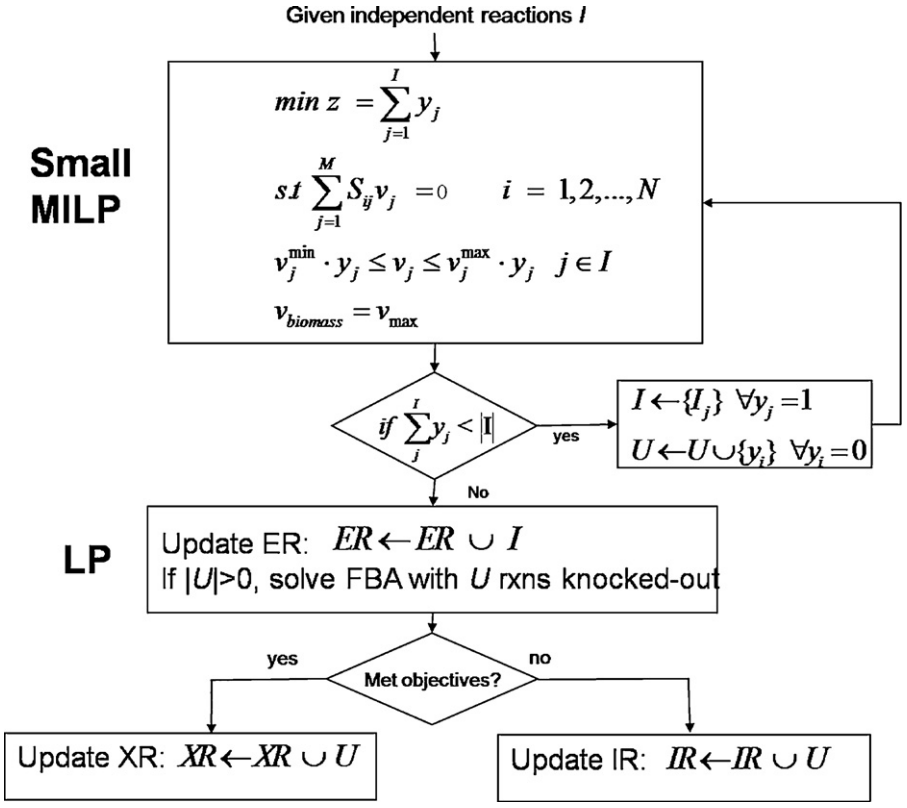

Fig. 4. Algorithm for classification of independent reactions at each search-level.

flux through the dependent reactions are allowed to be continuous. The objective function for optimization and other constraints are the same as those used for identifying minimal reaction set. Since only independent reactions are modeled as binary variables, the complexity of the optimization problem is low.

The classification of variables is based on the solution of the small MILP. In the optimal solution, if all the binary variables take the value 1 it indicates that *all the independent reactions are essential* for the organism to support the cellular objectives. These reactions are then appended to the essential reactions set. If the optimal objective value is less than the number of independent reactions, one or more reactions are extraneous or indeterminate. To identify only the essential reactions, independent reactions whose associated binary variable takes the value of 0 in the solution are removed from independent reactions set and included in a temporary set, U, for later analysis. Reactions whose associated binary variables take the value of 1 are potentially essential ones, but not guaranteed to be so. Hence, the MILP is solved again in another pass after restricting its scope further to consider only the potentially essential reactions in the independent reaction set. Thus, in every pass, reactions whose associated binary variables take 0 after optimization are removed from the independent reaction set and appended to U and the MILP is re-run considering the remaining independent reactions. This loop terminates when all binary variables takes value of 1 after optimization and the independent reaction set is guaranteed to contain solely essential reactions.

After identifying the essential reactions from the independent reactions, I, the remaining reactions in the temporary set U are further analyzed to determine whether they are extraneous. We use principles from Dynamic Programming (Williams, 1970) for this purpose. Dynamic Programming solves complex problems by breaking them into simpler stages and taking the best decision at each stage in a sequential manner. Let f_m be the optimal solution (for e.g. maximum) of an optimization problem that has m stages, mathematically f_m can be written as

$$f_m = \max\{s_m + s_{m-1} + \cdots s_1\} \tag{7}$$

where s_m is the solution for stage m. In general, the optimal solution f_m should be identified taken considering all the stages together. However, if stage m is independent of other stages, a recursive relation can be obtained for f_m as below:

$$f_m = \max\{s_m + \max(s_{m-1} + s_{m-2} + \cdots s_1)\} \tag{8}$$

which can also be rewritten as:

$$f_m = \max\{s_m + f_{m-1}\} \tag{9}$$

Eq. (9) shows that optimal decision can be taken for a stage independently of other stages. We use this insight to determine whether reactions U are extraneous and can be removed from further analysis. The procedure to check whether a given stage is independent of other stages involves knocking-out all the reactions in U from the FBA model (by setting their $v_j^{\min}$ and $v_j^{\max}$ to 0) and evaluating the capability of the organism to meet its predefined objectives. If the organism is found to be capable of meeting all its objectives, it indicates that the particular stage is independent of others and hence all the reactions in U are extraneous. This is because of the fact that reactions U are essentially different paths connecting the metabolites in that stage with the metabolites downstream. If the organism requires none of these paths to meet its objectives, it indicates that the stage is de-linked and the optimal solution can be selected for that stage independent of other stages. If the organism fails to meet its objectives, it means that some of the reactions in U are potentially essential; but their essentiality can only be resolved through an integrated analysis of the entire network. Hence, at the end of Phase 1 all these reactions are denoted as indeterminate reactions. Thus, at the completion of Phase 1, reactions have been classified as essential, extraneous, or indeterminate.

The above procedure for classification of independent reactions is based on the precept that the organism should require essential reactions irrespective of presence or absence of other reactions. This is captured in the iterative optimization scheme which evaluates the essentiality of the independent reactions (that are modeled as binary variables) while allowing other reactions to be used as necessary by considering the flux through them as continuous variables. At the end of each iteration, if the cell can support the predefined growth or chemical production even without any of the reactions in set U, then those indeterminate reactions are unnecessary for the cell and hence extraneous. Otherwise, if FBA reveals that the organism is untenable, it requires at least one reaction from U to meet its cellular objectives. Different subsets of U may serve the purpose. The selection of one such subset cannot be done during the iterative evaluation stage since at that juncture, results for downstream reactions are unknown and any classification of indeterminate reactions could be sub-optimal. Hence classification of indeterminate reactions is performed in Phase 2.

3.2. Phase 2

In Phase 2, extraneous reactions are removed from the model and only indeterminate reactions are evaluated further. A single MILP is formed with the indeterminate reactions represented as binary variables while the flux through the essential reactions is allowed to be continuous. The solution of this MILP identifies a subset of the indeterminate reactions (possibly null) that are required along with essential reactions identified in Phase 1 to form the minimal reaction set. This subset is denoted in Fig. 1 as the set of *additional reactions*. Once all reactions have been classified as essential or extraneous, the minimal set of reactions (genes) KO that have to be knocked in order to prevent all extraneous reactions can be determined by the graph traversal algorithm shown in Fig. 5.

Next, we illustrate the proposed procedure. Consider the simple network used as an example above. The depth of each reaction in Fig. 2b has been annotated next to it within boxes. Phase 1 requires 5 iterations since the maximum depth of this network is 5. In Iteration 1, with search-level initialized at 1, the two reactions at depth of 1 – r9 and r10 – are considered as independent reactions. A MILP with these two reactions modeled through binary variables and all other reactions as continuous variables is solved. The solution of this MILP has the objective function value of 1 revealing that at least one of these reactions is not essential in order to produce H. Since the binary variable corresponding to r9 is assigned a value of 0, r9 is not considered an independent variable in the next pass and is stored in set U. A new MILP is formed with only r10 as an independent reaction (binary variable) and solved. The optimal objective has a value of one revealing that r10 is essential. Then the flux through reaction r9 is constrained to 0 and the FBA identifies that the network can produce biomass. Hence, reaction r9 is declared as extraneous. Similarly, for iterations 2 and 3, reactions at depth of 2 – r1 and r3 – and depth of 3 – r2, r4 and r7 – are analyzed, respectively. Reactions r3 and {r4, r7} are identified as essential while r1 and r2 are classified as extraneous in iterations 2 and 3, respectively. For Iteration 4, reactions r5, r6 and r8 are independent reactions since their depth is 4. A MILP, with these three reactions as binary variables and all other reactions as continuous variables, is formulated and solved. The solution of this MILP has the optimal objective value 2 indicating that at least one reaction is not essential. The binary variables corresponding to reactions r6 and r8 take a value of 1 while that corresponding to r5 is assigned a value 0. Hence for the second pass, a new MILP is formed with only r6 and r8 as independent reactions while reaction

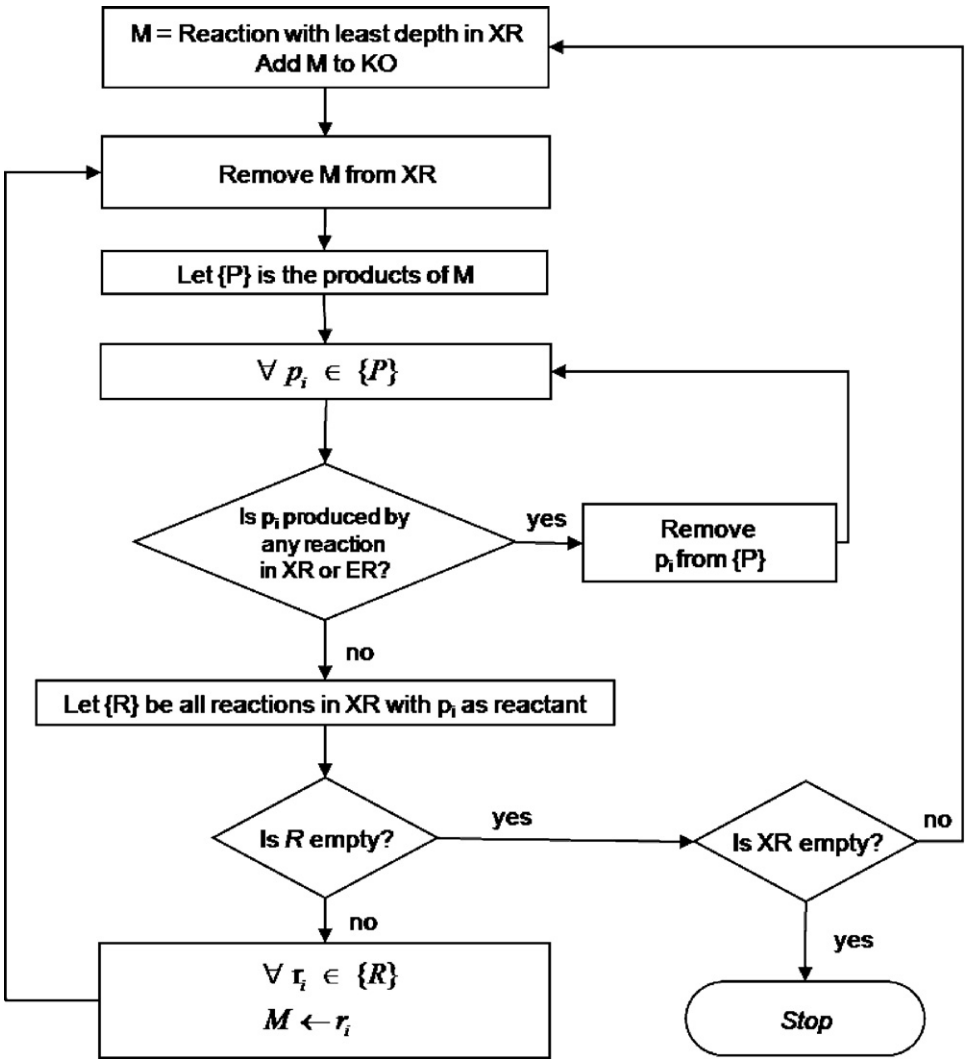

Fig. 5. Procedure for identifying reactions to be knocked-out from the cell to achieve minimal cell.

r5 is stored in set U. The solution of this MILP has the optimal objective value 1 and the binary variables corresponding to r6 and r8 are assigned a value of 0 and 1, respectively. In the third pass, a new MILP with only one binary variable, corresponding to r8, is formed and solved while r6 is added to U along with r5. The solution for this MILP has the objective value of 1 indicating r8 is essential. To classify r5 and r6 as extraneous or indeterminate, the FBA is conducted with the constraint of 0 flux through these reactions. FBA finds that the network is not capable of producing biomass. The network requires one of these in order to produce the biomass. Hence, reactions {r5, r6} are classified as indeterminate. The algorithm terminates at search-level of 5 after reactions r11 and r12 have been analyzed. Both these reactions are identified as essential. Reaction r11 is essential since it is the reaction exporting component H while reaction r12 is essential for maintaining the steady-state constraint. Hence, the proposed approach solves the sub-problem in each iteration rather than solving the whole optimization problem at once.

The indeterminate reactions {r5, r6} are analyzed in Phase 2 and r5 is identified as an additional reaction that is required by the cell. This reaction together with other essential reactions identified in Phase 1 forms the minimal set {r10, r3, r4, r5, r7, r8, r11 and r12} for this network. Reaction r6 could be selected instead of r5 since there are no associated downstream reactions for both of them. Such situations arise when there exist degenerate solutions in the optimization problem as is the case here.

Next, the procedure shown in Fig. 5 is then employed to identify the minimal set (KO) of reactions to be knock-out from the network in order to achieve the minimal cell. After Phase 1 and Phase 2, the extraneous reaction set contains reactions r9, r1, r2 and r6. The procedure is started by considering r9 as M since its depth is the smallest of all the reactions in XR. Reaction r9 is removed from XR and added to KO. Its product metabolite A is added to {P}. Since there is no other reaction producing component A, the reaction having A as reactant *i.e.* r1 is assigned to M and removed from XR. Similarly, r2 is assigned to M and removed from XR in the next iteration. The traversal of the graph has to stop here since the product of reaction r2, component E, is produced by another reaction in ER. Since XR is not empty, the algorithm continues with r6 as M as it has the lowest depth among all the remaining reactions in XR. Reaction r6 is then removed from XR and added to KO. The traversal of the network stops here since component F which is the product of reaction r6 is produced by a reaction in ER. The extraneous reaction set, XR, is now empty and the algorithm terminates. This procedure identified two reactions r9 and r6 as the minimal number of reactions to be knocked-out from the network in order to achieve the minimal cell. Though the result is intuitive for this sample network,

the procedure is very helpful for identifying minimal KO set in real large-scale networks.

4. Case studies

In this section, we illustrate the proposed method to identify the minimal reaction sets for *E. coli* and *S. cerevisiae*. The minimal reaction sets are identified such that the organisms are capable of supporting predefined growth rate under given environmental conditions. The results from the proposed method are compared with those obtained from the classical monolithic MILP approach. The above described method has been implemented in MATLAB™ and uses TOMLAB CPLEX solver.

4.1. Minimal reaction set for E. coli: aerobic growth on glucose

The first case study involves identification of minimal reaction set for *E. coli* growing on glucose as substrate. The metabolic model for *E. coli* used in this paper was developed by Edwards and Palsson (2000). It contains 63 metabolites participating in 76 reactions from the central metabolic pathways of *E. coli*.

For growth on glucose as the sole carbon source, the glucose uptake rate is constrained to 10 mmol/gDW h while uptake rates for oxygen and inorganic phosphate are unconstrained. Secretion routes are provided for all components capable of exiting the cell. Under these conditions, the Flux Balance Analysis predicts that the metabolic network is capable of achieving a maximal cell growth of 0.8614 g biomass/gDW h. The proposed approach is then employed to identify minimal reaction set for this organism while producing same quantity of biomass. Since the objective is to produce the biomass, the reaction leading to biomass production in the model is initially marked as Essential. Also, 9 other reactions involving only cofactors such as ADP, ATP, and NADP as both reactants and products are initialized as indeterminate since these cofactors are assumed to be always present in the system. Next, the essentiality of the reactions were determined using the proposed method as detailed below.

As described in Section 3, the proposed approach proceeds in a hierarchical manner starting from the primary uptake reactions as the top layer and continuing progressively until the secretion reactions in the bottom layer. In Iteration 1, the three exchange reactions for glucose, oxygen and phosphate, at depth 1 are considered as independent reactions. These three reactions are modeled as binary variables while flux through all other reactions is represented as continuous variables. On solution, CPLEX identifies that all three reactions are essential for the organism in order to meet the biomass constraint. In the next iteration, the reaction at depth 2, for transporting external glucose to the cytoplasm, is set to be independent and modeled using a binary variable. On solution, CPLEX determines it to be essential. Of the seven independent reactions at depth 3, three glucose conversion reactions are identified as essential and the Pyruvate conversion side reactions as extraneous. The algorithm thus continues till the last iteration as summarized in Table 2. In total, the proposed approach identified 38 reactions as essential and 29 reactions as extraneous. All reactions could be classified and no reaction was deemed as indeterminate during Phase 1. In Phase 2, only the cofactor reactions that were initialized as indeterminate need to be resolved. For this, the extraneous reactions identified in Phase 1 are removed from the model and the MILP is used considering only the nine indeterminate reactions as binary variables. Six out of the 9 reactions were then classified as essential, leading to a set of 44 reactions in the minimal set under the given conditions (see Table 3). For comparison, the minimal reaction set was also identified using the classical monolithic MILP procedure. The minimal reaction set identified by both approaches were exactly the same which confirms that the proposed method correctly identifies the minimal reaction set. In both approaches, the solution time was a few seconds.

Table 2
Results for Phase 1 of the proposed approach identifying minimal reaction set in *E. coli* supporting predefined quantity of biomass using glucose as sole carbon source.

Iteration #	# Ind Rxns	# ER	# IR	# XR
Initialization	0	1 (biomass)	9	0
1	3	3	0	0
2	1	1	0	0
3	7	3	0	4
4	17	11	0	6
5	21	12	0	9
6	12	5	0	7
7	4	2	0	2
8	1	0	0	1
Total	66	38	9	29

4.2. Minimal reaction set for E. coli: ethanol production from pentose sugar

We now identify the minimal reaction set for the same organism using Xylose as carbon source and producing ethanol as product while satisfying the specified growth rate condition. Production of biofuels such as ethanol from biomass is gaining in importance considering the depletion of fossil fuels and global warming. Xylose along with other pentose sugar Arabinose form a major part of hemi-cellulose and its conversion to bioethanol is essential for economical production of biofuels (Jeffries & Jin, 2004). Here, we identify the minimal reaction set as well as the set of reactions to be knocked-out for *E. coli* conversion of Xylose into ethanol while satisfying specified growth rate constraint. Since the proposed approach considers the dependency of reactions, it is useful for designing such minimal reaction organism in contrast to the regular MILP approach which identifies only the minimal reaction set but does not provide any clue on how to achieve such a minimal reaction organism through metabolic engineering.

The *E. coli* metabolic model used in this study is shown in Fig. 6. We have added 4 new reactions, two exchange reactions for exporting external Xylose to cytoplasm, one for converting Xylose to Xylulose and the fourth reaction for converting Xylulose to Xylulose-5-phosphate, to the network to enable *E. coli* to use Xylose as a substrate. The Xylose uptake is constrained to 10 mmol/gDW h and the growth rate is constrained to be at least 0.45 g biomass/gDW h. Under these conditions, according to FBA, the organism is capable of producing 7.233 mmol/gDW h of ethanol. The proposed approach is employed to identify the minimal reaction set while producing the same quantity of ethanol along with the growth rate constraint. Of the total of 80 reactions in the model, 49 reactions were identified to be in the minimal reaction set. The reactions which are not part of the set are shown with 'X' mark in Fig. 6. Reactions in the pathways leading to production of acetate, lactate, and formate should be silenced. Other reactions needed to be removed from the organism are those converting Oxaloacetate (Oaa) to Phosphoenolpyruvate (pep) and converting L-Malate (mal) to Pyruvate (pyr). Several exchange reactions including exchange of Oxoglutarate and Succinate should also be removed as shown in Fig. 6 to make the minimal cell.

The same result is obtained using the classical monolithic MILP approach. However, the proposed approach provides additional insights that are useful for designing the minimal reaction organism. Following the procedure shown in Fig. 5, the proposed approach identified only seven reactions – PFL, LDH_D, PTAr, ME, PPS, ICL and PPCK – which have to be knocked-out from the cell in order to achieve the minimal cell that produces ethanol from

Table 3
Minimal reaction set identified by proposed method in *E. coli*.

Rxn #	Rxn name	Formula	Depth	Result
1	ACKr	ac[c] + atp[c] ⇔ actp[c] + adp[c]	5	Extraneous
2	ACONT	cit[c] ⇔ icit[c]	5	Essential
3	ACt2r	ac[e] + h[e] ⇔ ac[c] + h[c]	6	Extraneous
4	ADHEr	accoa[c] + 2 h[c] + 2 nadh[c] ⇔ coa[c] + etoh[c] + 2 nad[c]	4	Extraneous
5	ADK1	amp[c] + atp[c] ⇔ 2 adp[c]	Initialization	Extraneous
6	AKGDH	akg[c] + coa[c] + nad[c] → co2[c] + nadh[c] + succoa[c]	4	Essential
7	AKGt2r	akg[e] + h[e] ⇔ akg[c] + h[c]	7	Extraneous
8	ATPM	atp[c] + h2o[c] → adp[c] + h[c] + pi[c]	Initialization	Essential
9	ATPS4r	adp[c] + 4 h[e] + pi[c] ⇔ atp[c] + h2o[c] + 3 h[c]	Initialization	Essential
10	Biomass	Biomass	Initialization	Essential
11	CO2t	co2[e] ⇔ co2[c]	Initialization	Essential
12	CS	accoa[c] + h2o[c] + oaa[c] → cit[c] + coa[c] + h[c]	4	Essential
13	CYTBD	2 h[c] + 0.5 o2[c] + q8h2[c] → h2o[c] + 2 h[e] + q8[c]	7	Essential
14	D_LACt2	h[e] + lac-D[e] ⇔ h[c] + lac-D[c]	4	Extraneous
15	ENO	2pg[c] ⇔ h2o[c] + pep[c]	4	Essential
16	ETOHt2r	etoh[e] + h[e] ⇔ etoh[c] + h[c]	5	Extraneous
17	EX_ac(e)	ac[e] ⇔	7	Extraneous
18	EX_akg(e)	akg[e] ⇔	8	Extraneous
19	EX_co2(e)	co2[e] ⇔	5	Essential
20	EX_etoh(e)	etoh[e] ⇔	6	Extraneous
21	EX_for(e)	for[e] ⇔	5	Extraneous
22	EX_glc(e)	glc-D[e] ⇔	1	Essential
23	EX_h2o(e)	h2o[e] ⇔	6	Essential
24	EX_h(e)	h[e] ⇔	4	Essential
25	EX_lac_D(e)	lac-D[e] ⇔	5	Extraneous
26	EX_o2(e)	o2[e] ⇔	1	Essential
27	EX_pi(e)	pi[e] ⇔	1	Essential
28	EX_pyr(e)	pyr[e] ⇔	4	Extraneous
29	EX_succ(e)	succ[e] ⇔	6	Extraneous
30	FBA	fdp[c] ⇔ dhap[c] + g3p[c]	5	Essential
31	FBP	fdp[c] + h2o[c] → f6p[c] + pi[c]	5	Extraneous
32	FORt	for[e] ⇔ for[c]	4	Extraneous
33	FRD	fadh2[c] + fum[c] → fad[c] + succ[c]	6	Extraneous
34	FUM	fum[c] + h2o[c] ⇔ mal-L[c]	5	Essential
35	FUMt2_2	fum[e] + 2 h[e] → fum[c] + 2 h[c]	6	Extraneous
36	G6PDH2r	g6p[c] + nadp[c] ⇔ 6pgl[c] + h[c] + nadph[c]	3	Essential
37	GAPD	g3p[c] + nad[c] + pi[c] ⇔ 13dpg[c] + h[c] + nadh[c]	5	Essential
38	GLCpts	glc-D[e] + pep[c] → g6p[c] + pyr[c]	2	Essential
39	GND	6pgc[c] + nadp[c] → co2[c] + nadph[c] + ru5p-D[c]	5	Essential
40	H2Ot	h2o[e] ⇔ h2o[c]	Initialization	Essential
41	ICDHyr	icit[c] + nadp[c] ⇔ akg[c] + co2[c] + nadph[c]	6	Essential
42	ICL	icit[c] → glx[c] + succ[c]	6	Extraneous
43	LDH_D	lac-D[c] + nad[c] ⇔ h[c] + nadh[c] + pyr[c]	3	Extraneous
44	MALS	accoa[c] + glx[c] + h2o[c] → coa[c] + h[c] + mal-L[c]	4	Extraneous
45	MDH	mal-L[c] + nad[c] ⇔ h[c] + nadh[c] + oaa[c]	5	Essential
46	ME1	mal-L[c] + nad[c] → co2[c] + nadh[c] + pyr[c]	5	Extraneous
47	ME2	mal-L[c] + nadp[c] → co2[c] + nadph[c] + pyr[c]	5	Extraneous
48	NADH11	3 h[c] + nadh[c] + q8[c] → 2 h[e] + nad[c] + q8h2[c]	7	Essential
49	NADTRHD	nad[c] + nadph[c] → nadh[c] + nadp[c]	Initialization	Extraneous
50	O2t	o2[e] ⇔ o2[c]	Initialization	Essential
51	PDH	coa[c] + nad[c] + pyr[c] → accoa[c] + co2[c] + nadh[c]	3	Essential
52	PFK	atp[c] + f6p[c] → adp[c] + fdp[c] + h[c]	4	Essential
53	PFL	coa[c] + pyr[c] → accoa[c] + for[c]	3	Extraneous
54	PGI	g6p[c] ⇔ f6p[c]	3	Essential
55	PGK	3pg[c] + atp[c] ⇔ 13dpg[c] + adp[c]	6	Essential
56	PGL	6pgl[c] + h2o[c] → 6pgc[c] + h[c]	4	Essential
57	PGM	2pg[c] ⇔ 3pg[c]	5	Essential
58	PIt	pi[c] ⇔ pi[e]	Initialization	Essential
59	PPC	co2[c] + h2o[c] + pep[c] → h[c] + oaa[c] + pi[c]	4	Essential
60	PPCK	atp[c] + oaa[c] → adp[c] + co2[c] + pep[c]	5	Extraneous
61	PPS	atp[c] + h2o[c] + pyr[c] → amp[c] + 2 h[c] + pep[c] + pi[c]	3	Extraneous
62	PTAr	accoa[c] + pi[c] ⇔ actp[c] + coa[c]	4	Extraneous
63	PYK	adp[c] + h[c] + pep[c] → atp[c] + pyr[c]	4	Essential
64	PYRt2r	h[e] + pyr[e] ⇔ h[c] + pyr[c]	3	Extraneous
65	RPE	ru5p-D[c] ⇔ xu5p-D[c]	5	Essential
66	RPI	r5p[c] ⇔ ru5p-D[c]	6	Essential
67	SUCCt2_2	2 h[e] + succ[e] → 2 h[c] + succ[c]	6	Extraneous
68	SUCCt2b	h[c] + succ[c] → h[e] + succ[e]	5	Extraneous
69	SUCD1i	fad[c] + succ[c] → fadh2[c] + fum[c]	5	Essential
70	SUCD4	fadh2[c] + q8[c] ⇔ fad[c] + q8h2[c]	6	Essential
71	SUCOAS	atp[c] + coa[c] + succ[c] ⇔ adp[c] + pi[c] + succoa[c]	4	Essential
72	TALA	g3p[c] + s7p[c] ⇔ e4p[c] + f6p[c]	4	Essential
73	THD2	2 h[e] + nadh[c] + nadp[c] → 2 h[c] + nad[c] + nadph[c]	Initialization	Extraneous
74	TKT1	r5p[c] + xu5p-D[c] ⇔ g3p[c] + s7p[c]	5	Essential
75	TKT2	e4p[c] + xu5p-D[c] ⇔ f6p[c] + g3p[c]	4	Essential
76	TPI	dhap[c] ⇔ g3p[c]	5	Essential

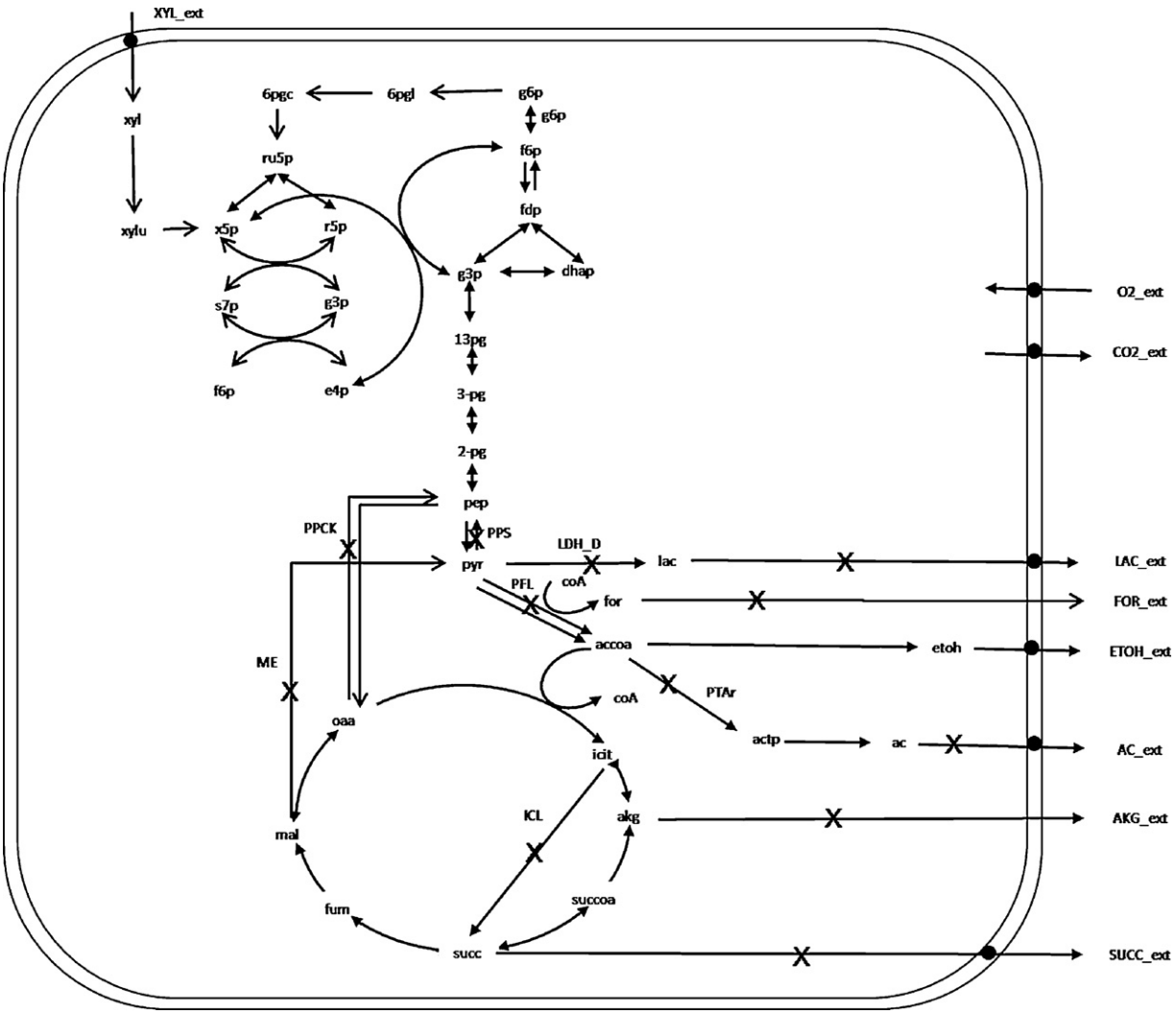

Fig. 6. The central metabolic network of *E. coli*. Reactions marked 'X' are the reactions identified by proposed method. Knocking-out of genes corresponding to these reactions will make the minimal cell for efficient conversion of Xylose to ethanol.

Xylose in most efficient way. Blocking of these seven reactions will automatically silence other extraneous reactions and the minimal cell is achieved. Three of these – PTAr, LDH and ME – are among the set of seven reactions that were knocked-out by Trinh et al. (2008) in an *E. coli* strain for producing growth associated ethanol from Xylose at theoretical yield.

4.3. Minimal reaction set for S. cerevisiae: aerobic growth on glucose

In this section, we illustrate the computational efficiency of the proposed approach by identifying the minimal reaction set for a genome-scale model of *S. cerevisiae* containing 1061 metabolites and 1266 reactions (Forster, Famili, Fu, Palsson, & Nielsen, 2003). The minimal reaction set is identified such that the organism achieves maximum growth rate on glucose as sole carbon source. The number of reactions in this model is reduced to 629 in the preprocessing step which removes reactions (or pathways) that have no connection with the glycolysis pathway. These reactions not connected to the glycolysis pathway do not carry flux under the specified conditions; thus they can be removed without affecting the final result. In this case study, the glucose uptake rate is constrained to 1 mmol/gDW h with unconstrained supply of oxygen, inorganic phosphate, sulfate and ammonia. Under these conditions, the maximum possible growth rate is found to be 0.0973 g biomass/gDW h using FBA. The proposed approach is then employed to identify the minimal reaction set that is capable of achieving the maximum possible growth rate.

The proposed approach identified 211 reactions as essential and 37 reactions as extraneous at the end of Phase 1. The remaining 381 reactions are designated as indeterminate. These indeterminate reactions are evaluated in Phase 2 and 67 of them found to be essential. In total, the minimal reaction set thus contains 278 reactions. The proposed approach requires 20,739 CPLEX iterations for Phase 1 that can be executed in approximately 32 s on a work station with 3 GB RAM and 3 GHz Xeon® processor. Phase 2 requires 23,761,130 CPLEX iterations and takes approximately 54 min. For comparison, we also identified the minimal reaction set using the classical monolithic MILP approach. The number of reactions in the minimal set is the same in both cases. However, the actual reactions are slightly different with 271 of the 278 reactions being common between the two sets. Both the sets satisfy the growth rate constraints and of equal size, hence we hypothesize that they are degenerate solutions to the optimization problem. The classical monolithic MILP approach requires 50,131,735 CPLEX iterations and approximately 120 min to reach the solution. The proposed approach thus requires less than 50% of the iterations and solution time compared to the classical monolithic MILP approach.

5. Discussion

In this paper, we proposed a hybrid approach combining graph-theoretic insights with a math program to identify minimal reaction sets. In contrast to the classical MILP approach which solves the whole optimization problem monolithically, the proposed approach solves the problem in a hierarchical manner considering only a few reactions at each iteration and using small MILPs. Thus the computational time is significantly reduced while achieving the same solution quality. The proposed approach also identifies key reactions from among the extraneous reactions that have to be knocked out from the organism in order to make the minimal cell. The proposed approach has been illustrated using three case studies. In the first case study, where the minimal reaction set is identified for *E. coli* under aerobic conditions with glucose as the sole substrate, the actual reactions identified by both the proposed and the classical methods match perfectly. Since it is a small model comprising only 76 reactions, both these methods solves the problem in a few seconds. This case study shows that the proposed approach correctly identifies the minimal reaction sets.

In the second case study, the proposed algorithm is used to design the minimal *E. coli* cell that produces ethanol at theoretical maximum yield using Xylose as the sole carbon source. This case study is intended to highlight the usefulness of the proposed approach to identify the key reactions that need to be knocked-out to achieve the minimal cell. Both the approaches identify the same 49 reactions for the minimal reaction set. In addition to this, the proposed method identified seven key reactions that can be knocked out for achieving the minimal cell. Knocking-out of these seven reactions precludes the flux through their downstream reactions thus making the cell minimal for the given conditions. Three of these seven reactions are a subset of the seven reaction knock-out strain reported by Trinh et al. (2008). The difference in the other reactions has to be expected since the *E. coli* model used by Trinh et al. (2008) and their growth requirements are different from the ones used in this study. Some of the reactions knocked-out by Trinh et al. (2008) are not included the model used here. Also, the minimal reaction set is known to be dependent on the growth constraint and nutrient conditions (Burgard et al., 2001). Given these, the observed extent of commonality in the reactions to be knocked-out should be viewed as confirmation.

In case study three, the proposed approach is used to identify the minimal reactions set for the larger genome-scale model of *S. cerevisiae*. In this study, both methods identified minimal reaction sets with equal number of reactions – an indication of degenerate solutions to the optimization problem. The number of iterations and time required for the proposed approach is far lesser than those required for the classical MILP approach. This clearly illustrates the benefit of the proposed approach. A closer look at the time taken by the two phases of the proposed approach brings out the fact that the time required (~32 s) for Phase 1 is negligible compared to that required (54 min) for Phase 2 (which uses the monolithic MILP for classification of *indeterminate* reactions). This indicates that bringing graph theory into MILP formulation is highly beneficial.

We also compared the minimal reaction set identified by the proposed approach with that from the monolithic MILP approach. For the *E. coli* case study, both the sets were the same. However, there were some differences in the minimal reaction sets for the *S. cerevisiae* case study indicating redundant pathways in the metabolic network (Lee, Fan, et al., 2005; Mahadevan & Schilling, 2003). Both sets had the same number of reactions in the minimal set (278), indicating them to be alternate optima. Of these, 271 reactions were in common. All the 211 reactions classified as essential (see Section 4.3) were present in both sets, as expected. In addition, 60 reactions were common between these two sets. During Phase 1 analysis, these 60 were classified as indeterminate as were the unmatched 14 reactions from the two solutions. The complete set of 381 indeterminate reactions can lead to many alternate solutions with a theoretical limit of $\binom{381}{67}$ (order of 10^{75}). This is a much smaller number compared to the original challenge of identifying all possible solutions without the graph-theoretic analysis which would have up to $\binom{629}{278}$ (order of 10^{186}) possibilities. Nevertheless, our current research is directed at identifying efficient approaches to identify all possible sets of additional reactions.

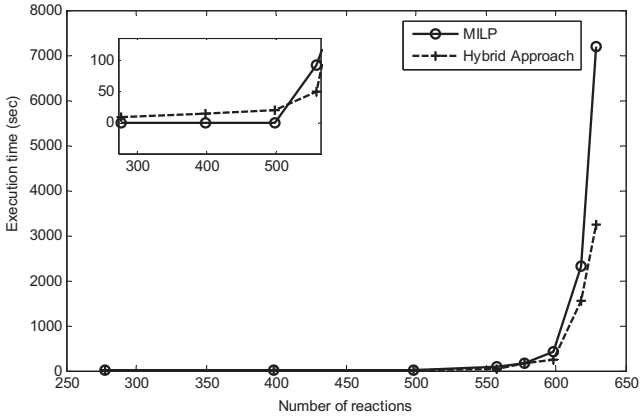

Fig. 7. Comparison of computational time required for both proposed and monolithic MILP approaches to identify minimal reaction set from networks with varying number of reactions. The computational time required for proposed hybrid approach is far lesser than the MILP approach as the size of network increases.

In order to show the efficacy of the proposed approach, we compare the computational time required with the monolithic MILP approach to identify minimal reaction set from networks of varying number of reactions. For this, we start with the minimal reaction network of *S. cerevisiae* identified in case study 3 and sequentially increase the number of reactions by adding new reactions randomly from the total reactions in this network. The computational results from this incremental increase in the problem complexity are shown in Fig. 7. Initially, the execution time required for the proposed hybrid approach to identify minimal reaction set is marginally higher than that of monolithic MILP (See inset in Fig. 7). This is due to the pre-processing steps such as depth assignment and classification of reactions in the proposed hybrid approach. As the number of reactions increases, the computational time required for the proposed hybrid approach is far lesser compared to the monolithic MILP approach although both approaches identify the same results. Further analysis revealed that the increasing computational time of the proposed approach is entirely from Phase 2 which involves the solution of the same MILP formulation as the monolithic one, but with fewer variables. The time required for Phase 1 is only 0.77% of the total execution time which clearly shows the benefits of incorporating graph theoretic ideas. We are now exploring the use of graph-theoretic insights for classification of indeterminate reactions in Phase 2 to further reduce the total computational time. The graph theoretic insights used in paper can also be extended to other computational approaches used in metabolic engineering such as methods for identifying gene insertion candidates. Such hybrid approaches can enable the analysis of large scale genome-scale models efficiently.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2011.05.006.

References

Braunschweig, B., & Gani, R. (2002). Software architectures and tools for computer aided process engineering. *Computer-Aided Chemical Engineering*, *11*

Burgard, A. P., Vaidyaraman, S., & Maranas, C. D. (2001). Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnology Progress*, *17*, 791–797.

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, *84*, 647–657.

Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E., et al. (2001). Metabolic modeling of microbial strains *in silico*. *Trends in Biochemical Sciences*, *26*, 179–186.

Demain, A. L. (2000). Small bugs, big business: The economic power of the microbe. *Biotechnology Advances*, *18*, 499–514.

Edwards, J. S., & Palsson, B. O. (2000). The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of National Academy of Sciences*, *97*, 5528–5533.

Forster, J., Famili, I., Fu, P., Palsson, B. O., & Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, *13*, 244–253.

Forster, A. C., & Church, G. M. (2006). Towards synthesis of a minimal cell. *Molecular Systems Biology*, *2*, 45.

Friedler, F., Tarjan, K., Huang, Y. W., & Fan, L. T. (1992). Graph-theoretic approach to process synthesis axioms and theorems. *Chemical Engineering Science*, *47*, 1973–1988.

Glass, J. I., Assad-Garcia, N., Alperovich, A., Yooseph, A., Lewis, M. R., Maruf, M., et al. (2006). Essential genes of a minimal bacterium. *Proceedings of National Academy of Sciences*, *103*, 425–430.

Hutchison, C. A., III, Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. *Science*, *286*, 2165–2169.

Jayakumar, S., & Reklaitis, G. V. (1994). Chemical plant layout via graph partitioning-1. Single level. *Computers and Chemical Engineering*, *18*, 441–458.

Jeffries, T. W., & Jin, Y. S. (2004). Metabolic engineering for improved fermentation of pentoses by yeasts. *Applied Microbiology and Biotechnology*, *63*, 495–509.

Lee, S. Y., Lee, D. Y., & Kim, T. Y. (2005). Systems biotechnology for strain improvement. *Trends in Biotechnology*, *23*, 349–358.

Lee, D. Y., Fan, L. T., Park, S., Lee, S. Y., Shafie, S., Bertok, B., et al. (2005). Complementary identification of multiple flux distributions and multiple metabolic pathways. *Metabolic Engineering*, *7*, 182–200.

Mah, R. S. (1990). *Chemical process structures & information flows*. Butterworth-Heinemann.

Mahadevan, R., & Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, *5*, 264–276.

Meyer, M., Le Lann, J. M., Koehret, B., & Enjalbert, M. (1994). Optimal location of sensor location on a complex plant using a previous termgraphnext term oriented approach. *Computers and Chemical Engineering*, *18*(Suppl.), S535–S540.

Mushegian, A. R., & Koonin, K. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of National Academy of Sciences*, *93*, 10268–10273.

Nilsson, N. J. (1980). *Principles of artificial intelligence*. CA: Palo Alto.

Palaniappan, V., Srinivasan, R., & Halim, I. (2002). A material-centric methodology for developing inherently safer environmentally benign processes. *Computers and Chemical Engineering*, *26*, 757–774.

Patil, K. R., Rocha, I., Förster, J., & Nielsen, J. (2005). Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics*, *6*, 308.

Prakash, R., & Shenoy, U. V. (2005). Design and evolution of water networks by source shifts. *Chemical Engineering Science*, *60*, 2089–2093.

Preisig, H. A. (2009). A graph-theory-based approach to the analysis of large-scale plants. *Computers and Chemical Engineering*, *33*, 598–604.

Pharkya, P., Burgard, A. P., & Maranas, C. D. (2004). OptStrain: A computational framework for redesign of microbial production systems. *Genome Research*, *14*, 2367–2376.

Pohorille, A., & Deamer, D. (2002). Artificial cells: Prospects for biotechnology. *Trends in Biotechnology*, *20*, 123–128.

Price, N. D., Reed, J. L., & Palsson, B. O. (2004). Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews*, *2*, 886–897.

Shivakumar, K., & Narasimhan, S. (2002). A robust and efficient NLP formulation using graph theoretic principles for synthesis of heat exchanger networks. *Computers and Chemical Engineering*, *26*, 1517–1532.

Srinivasan, R., & Venkatasubramanian, V. (1995). Automating HAZOP analysis of batch chemical plants: Part I. The knowledge representation framework. *Computers and Chemical Engineering*, *22*, 1345–1355.

Stephanopoulos, G. (2002). Metabolic engineering: Perspective of a chemical engineer. *AIChE Journal*, *48*, 920–926.

Thambirajah, J., Benabbas, L., Bauer, M., & Thornhill, N. F. (2009). Cause-and-effect analysis in chemical processes utilizing XML, plant connectivity and quantitative process history. *Computers and Chemical Engineering*, *33*, 503–512.

Trinh, C. T., Unrean, P., & Srienc, F. (2008). Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and Environmental Microbiology*, *74*, 3634–3643.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers and Chemical Engineering*, *27*, 293–311.

Williams, K. (1970). *Dynamic programming sequential decision making*. London: Longman.