NUTRIENT-RELATED ANALYSIS OF PATHWAY/GENOME DATABASES

P.R. ROMERO AND P. KARP
Artificial Intelligence Center, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025, USA
{promero,pkarp}@ai.sri.com

We present an algorithm that solves two related problems in the analysis of metabolic networks stored within a pathway/genome database. (1) The Forward Propagation Problem: given a set of nutrients that are inputs to the metabolic network, what compounds will be produced by the metabolic network? (2) The Backtracking Problem: given the results of a forward propagation, and given a set of essential compounds that are not produced as a result of the forward propagation, what precursors must be supplied to produce those essential compounds? A program based on this algorithm is applied to the EcoCyc database, which is a pathway/genome database for *E. coli* that consists of annotated genomes and the metabolic reactions and pathways associated with the known gene products. The inputs to the program are a description of the metabolic network of an organism (EcoCyc), a set of nutrients corresponding to a known minimal growth medium, and a list of essential compounds to be produced. The program "fires" the microorganism's metabolism contained in the database and predicts all synthesized and nonsynthesized essential compounds, along with the missing precursors required to produce the latter. When applied to the EcoCyc database, the program identifies a number of missing precursors that indicate incomplete regions of the database. Thus the program results can be used to evaluate existing pathway databases like EcoCyc.

1 Introduction

Recent years have seen an increase in the number of small genomes that have been completely sequenced. Still, genome annotation efforts have yet to reach completion, so for many sequenced genomes 40-80% of the corresponding proteome may have unknown functions[1].

A number of groups have developed integrated pathway/genome databases through prediction of the metabolic pathways of an organism from its genome. These initiatives include the KEGG project[2], the WIT project[3] and the project that has produced EcoCyc and other pathway/genome databases[4]. These latter databases consist of microorganism-specific pathway/genome datasets and software tools for query, visualization and analysis of the data.

The EcoCyc (*E. coli Encyclopedia*) knowledge base (*KB*)[5] contains the complete *E. coli* (strain K-12) genome and its known gene products along with the set of all metabolic and transport reactions that are known to occur in the microorganism. All this information is kept in a highly structured database schema, which allows for computational analysis of the data[6]. Metabolic information is represented by four object types: metabolic compounds (1), pathways (2) and reactions (3), and the enzymes that catalyze these reactions (4). Genome information is characterized by

three object types: chromosomes and plasmids (5), genes (6) and the corresponding gene products (7). The primary link between metabolic and genomic information relates gene products that encode enzymes to the reactions these enzymes catalyze. EcoCyc also includes transport reactions and signaling pathways.

1.1 Metabolic Analysis and Nutrition-Related Studies

Metabolic analysis is usually carried out through quantitative calculation of the fluxes of chemical species through metabolic networks. This approach has been applied to pathway analysis and prediction[7,8], metabolic simulation[9] and metabolic engineering[10,11]. Here we employ a qualitative metabolic analysis that does not depend on mathematical models, but rather on connectivity aspects of the studied metabolic network. With this kind of analysis, we can use the metabolic information contained in a pathway/genome database like EcoCyc to estimate the nutrition requirements for a microorganism such as *E. coli*. The approach can be also applied to other unicellular organisms and /or cell types for which metabolic information is known.

1.1.1 Nutrition of Microorganisms

A microorganism can be made to grow in a *culture medium*, from which it can obtain its *nutrients*, that is, the chemical compounds necessary for it to grow and reproduce. When the exact chemical composition of the culture medium is known, it is called a *defined medium*. A defined medium can contain just a *minimal set* of nutrients such that the microorganism can grow only if each and every one of those nutrients is present. It is then called a *minimal growth medium*. The culture media known for many microorganisms are not defined media, but are complex media that contain undefined components such as beef serum.

Nutrients are those compounds that the organism can utilize as *sources* for the chemical elements necessary to sustain life, especially carbon, sulfur, phosphorus, nitrogen, potassium, iron and magnesium, among others. Some microorganisms can synthesize all needed organic compounds from simple carbon sources (e.g., usually glucose, lactose or other simple sugars). Sometimes, more complex carbon sources are needed to supply important organic compounds that the organism cannot synthesize, called *growth factors*. So, a nutrient set for a given microorganism will consist of sources of needed inorganic elements, plus a carbon source and/or growth factors (a growth factor usually also serves as a carbon source) such that the organism can grow using that nutrient set.

Some elements and compounds are also necessary for the organism's enzymes to carry out their catalytic function. These enzyme-required elements/compounds can

be of two types, called *cofactors* and *prosthetic groups*. For the sake of simplicity, we use the term *cofactor* to encompass both cofactors and prosthetic groups, although their relationships to the enzymes is different.

1.1.2 Metabolic Information

The metabolic information used in this study consists of the set of biochemical (metabolic and transport) reactions and pathways — as well as the enzymes and respective cofactors needed for these reactions to take place — known to occur in *E. coli* and contained in EcoCyc. We use the term *transport reactions* because all transport processes are modeled as reactions in EcoCyc. For our nutrition studies, we are interested in how the organism generates the chemical building blocks used to synthesize the macromolecules essential for life: proteins, nucleic acids (DNA and RNA), phospholipids (components of cell membranes), lipopolysaccharides and oligosaccharides (components of cell walls). For the purposes of this research, these building blocks of macromolecules (i.e., amino acids, purines and pyrimidines, lipids and saccharides) are considered *essential compounds* that the microorganism needs to produce in order to grow. From this perspective, we can divide reaction and pathway information into two classes: the *small molecule metabolism* (SMM) where the above-mentioned building blocks are generated, and the *macromolecule metabolism* where these building blocks are assembled into final cell components.

1.2 Problem Definitions

This article addresses two computation al problems:

1. The forward propagation problem: given a set of input nutrients, what compounds will be produced by the SMM when it metabolizes those nutrients? We seek a qualitative prediction of *what* compounds will be produced, but not the quantities of those compounds.
2. The backtracking problem: given the results of a forward propagation, and given a set of essential compounds that are not produced as a result of the forward propagation, generate alternative sets of precursors that, if supplied, will produce those essential compounds. Notice that a precursor may or may not be a nutrient, depending on its transportability into the cell. It is just a compound needed as a reactant in order to fire a given reaction.

Solving these two problems can provide useful information on the characteristics of a pathway database. For example, we can verify the completeness and consistency of a pathway/genome database by supplying the known minimal

nutrient set of an organism as the input for a forward propagation, and asking whether all known essential compounds for that organism are produced. This study shows the application of this approach to the analysis of the EcoCyc database.

The following sections explore our study in more depth. Section 2 details both the algorithm and data used. The selected inputs and results obtained are shown in Section 3 and discussed in Section 4.

2. Algorithm

The algorithm used here is divided into two phases that solve problems 1 and 2.

Inputs: (1) metabolic information for the studied organism, (2) a set of essential compounds, (3) a known minimal set of nutrients for the microorganism, (4) a set of bootstrapping compounds (see below).

Outputs: (5) the final set of metabolites (initial metabolites plus the compounds produced from the initial metabolites), (6) a list of which essential compounds are synthesized and which are not, (7) a list of additional precursors that, if added to (3), would produce the essential compounds that could not be synthesized.

Algorithm:

Preamble:

1. Find which nutrients can be transported into the cell.
2. Assemble an initial metabolite set by combining the bootstrapping compounds and the transported nutrients.

Phase I (forward propagation):

3. Starting from the initial metabolite set, fire the metabolism (see below) and gather all compounds thus produced (the initial metabolite set plus all synthesized metabolites).

Phase II (backtracking):

4. Check non-produced essential compounds and find their missing precursors, i.e., those that were either not provided or not synthesized, and whose absence prevented some essential compounds from being produced.

2.1 Essential Compounds

The set of essential compounds can be determined by literature searches on the structure and composition of the microorganism, which will provide the compounds that must be assembled into the essential cell components. Essential compounds

include building blocks of the macromolecules required by the cell (such as amino acids and nucleoside triphosphates). When information is scarce, the essential compounds can be estimated from general information about similar organisms (i.e., gram positive or gram-negative bacteria). Any estimated essential compound thus found must be checked against the pathway database: only those appearing in the database are to be included in the final essential compounds set.

2.2 Minimal Nutrient Set

Our program, applying the algorithm explained above, analyzes a pathway database for completeness by asking whether a known minimal nutrient set for the studied microorganism can produce all of the essential compounds for that organism. If this is the case, the program ends after forward propagation has been performed (Phase I); otherwise, backtracking (Phase II) is executed.

2.3 Bootstrapping Compounds

General metabolic information tells us that most reactions require energy sources like ATP, redox acceptors/donors like NAD, and acyl carriers like CoA. Although the organism may be able to synthesize these compounds, they are required to bootstrap the entire metabolism — including sometimes their own synthesis — and allow for most reactions to start firing on the provided nutrients so that the whole metabolic process can be started in step 3. For example, although glycolysis produces ATP, it requires an early input of ATP before any ATP is produced. Thus, a set of such "bootstrap" compounds is added to the initial set of metabolites. Biosynthetic pathways that produce the bootstrap compounds are analyzed after forward propagation has been completed, to check whether they were fired, which implies that bootstrap compounds can be synthesized from the supplied nutrients.

2.4 Firing the Metabolism

Any reaction for which all reactants and cofactors are present in the current list of metabolites is considered "fired", and its products are added to the current set of metabolites, which grows monotonically. This process repeats, looping over all non-fired reactions until no more reactions can be fired with the current set of metabolites. The directionality of a reaction determines which of its substrates are reactants and which are products, so the program must know the directionality of a reaction before attempting to fire it. This can be inferred from the directionality of a pathway that includes the reaction (different pathways may use the reaction in

different directions). If the reaction does not appear in any pathway, the database can specify its directionality. A reversible reaction can be fired on either direction.

2.5 Finding Precursors: Backtracking

Step 4 is accomplished by backtracking from non-produced essential metabolites through reactions and pathways that can synthesize them, and recursively gathering combinations of those precursors not present in the final set of metabolites produced by the forward propagation process. The backtracking process stops when it reaches the *inputs* of the SMM *reaction network*, that is, compounds that are not the products of any reaction in the SMM. Because a compound can be the product of more than one reaction or pathway, backtracking results can be a combination of different sets of precursors, any of which can produce the desired compounds. The program will output all possible combinations of precursors (see Figure 1).

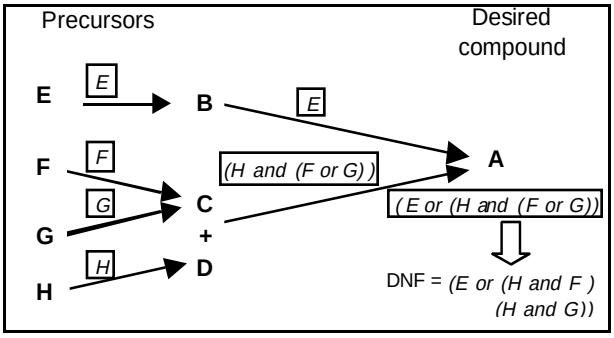

Figure 1. Backtracking from compound A results in an expression involving reaction-network inputs needed to synthesize A. Expressions resulting from backtracking through reactions (arrows) are shown in boxed, italic text. The final expression can be put into disjunctive normal form (DNF). For example, synthesizing A through the reaction involving condensation of C and D yields an expression that combines H (precursor of D) with either F or G (precurors of C).

Care should be taken not to fall into loops when backtracking. For efficiency, the program caches the precursors computed for a given compound so that they need not be recomputed.

2.6 Precursor Analysis

The resulting expressions for all essential compounds not produced are manually analyzed in order to find precursors that are needed because of database incompleteness. Directly examining the pathway database will help explain why these precursors are not present after forward propagation, providing useful insights on the database's completeness.

2.7 Implementation of the Algorithm

We developed a Common Lisp program that can implement the preceding algorithm on EcoCyc or any of the other pathway/genome databases. Working on a database like EcoCyc and fed with an approximate nutrient set, appropriate bootstrap compounds, and a set of essential compounds, the program can to provide the user with many different user-selected outputs, including

- List of transported and nontransported nutrients.
- List of all fired and nonfired reactions.
- List of all fired, nonfired and partially fired pathways.
- List of used and unused nutrients.
- List of synthesized and non-synthesized essential compounds.
- For each non-synthesized essential compound, an expression showing possible network inputs' combinations that must be supplied in order to synthesize it, as well as a backtracking log that shows all intermediate precursors, reactions and pathways covered to reach the final expressions. This log is used as complementary analysis information.

3. Results of Applying the Program to EcoCyc

Using the metabolic information stored in EcoCyc, the program was run using a known minimal medium for *E. coli*. The corresponding data is shown in section 3.1. Results are shown and analyzed in sections 3.2 and 3.3, respectively.

3.1. Metabolic and Nutrition Data

Table 1. shows statistics on the small-molecule metabolism information we used from EcoCyc. The other inputs to the program are shown in Tables 2 and 3.

Table 1. Pathway data studied.

Objects	*Quantity in EcoCyc*
Reactions (total)	1740
Small molecule metabolism	804
Transport	164
Pathways (total)	198
Small molecule metabolism	141
Compounds	1891

Table 2. M63 minimal growth medium composition.

Growth medium component	*Corresponding compound(s) included in initial nutrient set*
Glucose	Glucose
KH_2PO_4	K^+, H^+, PO_4^{-3}
$(NH_4)_2SO_4$	NH_4^+, SO_4^{-2}
$MgSO_4.7H_2O$	Mg^{+2}, SO_4^{-2}, H_2O
$FeSO_4.7H_2O$	Fe^{+2}, SO_4^{-2}, H_2O
KOH	K^+, OH^-
Trace elements found in water	Mn^{+2}, Co^{+2}, Cu^{+2}, Mo^{+2}, Ca^{+2}, Zn^{+2}, Cd^{+2}, Ni^{+2}

Table 3. *E. coli* essential compounds.

Cellular component	*Building blocks*
Proteins	All 20 amino acids
DNA & RNA[12]	Pyrophosphate nucleotides: DNA: dATP, dTTP, dGTP, dCTP RNA: ATP, UTP, GTP, CTP
Cytoplasmic membrane[13]	Phospholipids: Phosphatidylethanolamine Phosphatidylglycerol Cardiolipin
Outer membrane (peptidoglycan)[12,14,15]	Muropeptide monomer(C_6): undecaprenyl-pyrophosphoryl-MurNAc-(pentapeptide)-N-acetylglucosamine
Cell wall[12-14] • Lipid A layer • Core oligosaccharide and repeating oligosaccharide chain (O antigen)	• Lipid-A disaccharide • Heptose (L-glycero-D-manno-heptose) KDO (2-keto-3-deoxyoctonate) Ethanolamine UDP-glucose UDP-galactose dTDP-rhamnose GDP-mannose N-acetylglucosamine

Table 2 shows the minimal growth medium M63[16] and the corresponding compounds used to generate the initial nutrient set.

The essential compounds shown in Table 3 were compiled from literature sources referenced within the table. It should be noted that the O antigen part of the cell wall is not essential for growth, but it was included anyway as a component of a normal cell.

The bootstrap compounds used were those mentioned in the previous section: ATP, NAD, Coenzyme-A, and oxygen (included because M63 is a medium used for aerobic growth).

3.2 Results

Table 4. Precursors needed for production of essential compounds in EcoCyc.

Precursors	**Comments**
Reactants:	
3-hydroxymyristoyl-ACP apo-ACP Thioredoxin	All are proteins (ACP = acyl carrier protein), which are not small molecules, and are therefore outside the scope of our analysis
THF	THF derivatives are needed to produce GTP, but THF needs GTP as precursor (folic acid biosynthesis).
HCO_3	Obtained from CO_2 in aqueous solution (non enzymatic reaction to be added to EcoCyc).
delta(3)-isopentenyl-pyrophosphate	No precursors in EcoCyc
Heptose	No precursors in EcoCyc
Cofactors:	
Pyridoxal phosphate (Vitamin B_6)	Needed as a cofactor in its own biosynthesis.
Cob(I)alamin (Vitamin B_{12}) OR 5-methyltetrahydropteroyltri-L-glutamate	Used as cofactors for the same reaction under different conditions: B_{12} must be gathered from the environment, while 5-methyltetrahydropteroyltri-L-glutamate has no precursor in EcoCyc
Siroheme	Needed to produce cysteine, but siroheme needs cysteine as a precursor.
THZ (5-(2-hydroxyethyl)-4-methylthyazole)	No precursors in EcoCyc

The program was run with the inputs shown in Section 3.1, and the missing essential compounds and their precursors were manually analyzed. Precursors that appeared repeatedly when backtracking from different essential compounds were carefully studied. Our strategy was to select each precursor in turn, add it to the list of bootstrap compounds, and rerun the program in an incremental fashion. This

approach simplified the precursor-selection process, as each successive run produced simpler and simpler expressions for the remaining non-synthesized essential compounds.

Table 4 shows the end result of this precursor selection process: it lists the final set of compounds found to be needed as extra nutrients, so that all essential compounds are produced by the SMM defined in EcoCyc after forward propagation from the M63 minimal growth medium.

3.3 Analysis of Results

One essential compound, heptose, has no precursors in EcoCyc. This is because little is known about its biosynthesis in *E. coli*[13,17].

An interesting case occurs with cob(I)alamin (vitamin B_{12}) and 5-methyltetrahydropteroyltri-L-glutamate. These serve as cofactors for the same reaction, but cob(I)alamin is used only during anaerobic growth, whereas the other cofactor is used when growth is aerobic. It is known that *E. coli* cannot synthesize B_{12}[13], and so the microorganism must take this nutrient from the environment. But growth medium M63 is used, as we explained previously, under aerobic conditions The program selects cob(I)alamin as a precursor, though, because such information on the regulatory aspects of reactions is not currently available in EcoCyc. On the other hand, the aerobic (not cob(I)alamin dependent) version of the reaction could not be fired either because it needed 5-methyltetrahydropteroyltri-L-glutamate, which has no precursor in EcoCyc. The database is incomplete in this respect because the mechanism by which this compound is synthesized is unknown[13].

There are several cases of loops in the precursor lists: Either a compound needs a precursor that in turn needs the compound as a precursor (THF and siroheme), or the compound itself is needed as cofactor in its own biosynthesis (pyridoxal-phosphate). These compounds should be treated as bootstrap compounds, because their synthetic pathways are known, but they are needed to start the metabolic process — and even their own synthesis — as with ATP or NAD. Here we must remember that growth within a cell always starts from a whole cell, and not from an "empty cell" as in our simulations. A whole cell will already have trace amounts of these compounds present.

3-hydroxymyristoyl-ACP, apo-ACP and Thioredoxin are proteins produced by the macromolecule metabolism, so they can also be treated as bootstrap compounds, as we are dealing with the SMM only. Having the unmodified acyl carrier protein (apo-ACP) as an input to the program guarantees the production of many, but not all moieties of ACP, as is the case with 3-hydroxymyristoyl-ACP, whose synthesis mechanism in *E. coli* is not known.

Bicarbonate (HCO_3^-) is present in the cell as a byproduct of CO_2, but there are no explicit metabolic pathways/ reactions for the synthesis of this ion in EcoCyc. This analysis have prompted the addition of this reaction to EcoCyc.

Finally, we are left with other precursors that are known to be produced by *E. coli*, but for which no complete synthesis information is yet available. In the case of delta(3)-isopentenyl-pyrophosphate there is a proposed biosynthetic pathway that has not been confirmed yet[13]. In the case of thiamine's thiazole subunit (THZ), the complete biosynthetic pathway has not been elucidated[13].

In summary, analysis of the program's output against the EcoCyc data and literature searches found metabolic information missing in EcoCyc, most of it because of gaps in our current knowledge. Once the absent information is accounted for, the analysis confirmed the facts that *E. coli* can grow on the M63 medium without the addition of any growth factor, and that the organism uses all chemical components of the minimal medium, as expected.

4. Discussion

This study has shown how we can apply qualitative analysis to pathway/genome databases in order to address metabolic problems at a complete organism level. The algorithm we present is suitable for assessing the completeness of a pathway/genome database, and for performing forward propagation of a set of compounds through a metabolic network.

One limitation of this study is the absence of regulatory information in EcoCyc, the effects of which can be appreciated in the cob(I)alamin example above. Incorporating this kind of data to the database will allow for the inclusion of growth conditions (temperature, pH and the like) as inputs for the program, providing more accurate results.

The application of the backtracking process used here to the problem of finding minimal nutrition requirements is the next step of this ongoing research. In this case, backtracking will be performed from every essential compound to all transportable precursors. There is no forward propagation phase, as we are trying to find alternative nutrient sets. The lack of a forward-propagation phase means that many more essential compounds are considered at one time, which produces a much larger search space than in the present studies, where backtracking is carried out only through reactions that were not fired during the forward propagation stage.

Several other applications for this kind of analysis can be envisioned, ranging from finding metabolic routes between given compounds to checking for the existence of isolated (disconnected) pathways/reactions. The latter analysis could

help diagnose predicted metabolic networks, like those produced by our Pathologic program[18].

The data resulting from these qualitative analyses is very useful, despite not taking into account complex aspects of metabolism, like regulation and metabolic fluxes, an analysis of which requires determining, estimating or, in many cases, assuming the values of many parameters. The idea behind the analysis presented here is to study what a metabolic machinery is capable of, and the information obtained, valuable in its own right, can also help guide more complex analysis endeavors.

References

1. C.H. Schilling, J.S. Edwards, B.O. Palsson and R. Heinrich, *Biotechnol. Prog*. 15:3 296 (1999).
2. M. Kanehisa and S. Goto, *Nucleic. Acids Res*. 28:1, 27 (2000).
3. R. Overbeek, N. Larsen, G.D. Pusch, M. D'Souza, E. Selkov Jr, N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov, *Nucleic. Acids Res*. 28:1, 123 (2000).
4. P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, S.M. Paley and A. Pellegrini-Toole, *Nucleic. Acids Res.* 28:1, 56 (2000).
5. P.D Karp, M. Riley, S.M. Paley, A. Pellegrini-Toole and M. Krummenacker, *Nucleic. Acids Res.* 26:1, 50 (1998).
6. C.A. Ouzounis and P.D. Karp, *Genome Res*. 10:4, 568 (2000).
7. C.H. Schilling and J.S. Edwards and B.O. Palsson, *Biotechnol. Prog*. 15:3, 288 (1999).
8. S. Schuster, D.A. Fell and T. Dandekar, *Nat. Biotechnol*. 18:3, 326 (2000).
9. Goryanin, T.C. Hodgman and E. Selkov, *Bioinformatics* 15:9, 749 (1999).
10. B. Christensen and J.Nielsen, *Adv. Biochem. Eng. Biotechnol*. 66, 209 (2000).
11. J.S. Edwards and B.O. Palsson, *Biotechnol. Bioeng*. 58:2-3 162 (1998).
12. J. Mandelstam, K. McQuillen and I. Dawes, *Biochemistry of Bacterial Growth* 3rd. Ed. (John Wiley & Sons, New York, 1982).
13. *Escherichia coli and Salmonella Cellular and Molecular Biology* Neidhardt, F. et al. eds. 2nd. ed. (ASM Press, Washington, 1996).
14. D. Metzler, *Biochemistry* (Academic Press, New York, 1977).
15. G. Gottschalk, *Bacterial Metabolism*. 2nd. ed. (Springer, New York, 1986).
16. J. Sambrook, E. Fritsch and T. Maniatis, *Molecular Cloning: a Laboratory Manual* 2nd. Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).
17. J. Kadrmas and C. Raetz, *Biol. Chem*. 273:5, 2799 (1998).
18. P.D. Karp, M. Krummenacker, S. Paley, J. Wagg, *Trends Biotechnol*. 17:7, 275 (1999).