Metabolic Pathway Analysis via Integer Linear Programming

A thesis submitted for the degree of
Doctor of Philosophy

by

Francisco J. Planes

Department of Mathematical Sciences,
Brunel University,
Uxbridge, Middlesex, UB8 3PH, United Kingdom.

May 2008

Abstract

The understanding of cellular metabolism has been an intriguing challenge in classical cellular biology for decades. Essentially, cellular metabolism can be viewed as a complex system of enzyme-catalysed biochemical reactions that produces the energy and material necessary for the maintenance of life. In modern biochemistry, it is well-known that these reactions group into metabolic pathways so as to accomplish a particular function in the cell. The identification of these metabolic pathways is a key step to fully understanding the metabolic capabilities of a given organism. Typically, metabolic pathways have been elucidated via experimentation on different organisms. However, experimental findings are generally limited and fail to provide a complete description of all pathways. For this reason it is important to have mathematical models that allow us to identify and analyze metabolic pathways in a computational fashion. This is precisely the main theme of this thesis.

We firstly describe, review and discuss existent mathematical/computational approaches to metabolic pathways, namely stoichiometric and path finding approaches. Then, we present our initial mathematical model named the Beasley-Planes (BP) model, which significantly improves on previous stoichiometric approaches. We also illustrate a successful application of the BP model to optimally disrupt metabolic pathways. The main drawback of the BP model is that it needs as input extra pathway knowledge. This is especially inappropriate if we wish to detect unknown metabolic pathways. As opposed to the BP model and stoichoimetric approaches, this issue is not found in path finding approaches. For this reason a novel path finding approach is built and examined in detail. This analysis serves us as inspiration to build the Improved Beasley-Planes (IBP) model. The IBP model incorporates elements of both stoichometric and path finding approaches. Though somewhat less accurate than the BP model, the IBP model solves the issue of extra pathway knowledge. Our research clearly demonstrates that there is a significant chance of developing a mathematical optimisation model that underlies many/all metabolic pathways.

Contents

List of Tables

List of Figures

Main Notation

B The set of pairs of reactions (α,β) such α and β are the reverse of each other.

b_c Binary variable indicating whether compound c is balanced in the pathway.

C Number of compounds in the metabolic network.

Δ Input parameter for applying balancing constraints in the BP model.

D_1 Set of high presence compounds.

D_2 Set of inorganic compounds.

D_3 Set of cofactors.

D_4 Set of main compounds.

δ_c The percentage presence of compound c.

d_{rc} 1, if compound c is an output compound for reaction r; 0, otherwise.

E Number of enzymes in the metabolic network.

e_c Binary variable indicating whether compound c is produced to excess in the pathway.

f_c Binary variable indicating whether compound c is consumed (freely available) in the pathway.

G_r Standard Gibbs free energy involved in one tick of reaction r.

H_r Real Gibbs free energy involved in one tick of reaction r.

K Maximum number of metabolic paths.

L Length of the pathway.

m_{cr} 1, if compound c is an input compound for reaction r; 0, otherwise.

n_{cr} Number of molecules of compound c needed as input for one tick of reaction r.

N_{cr} Relative presence for compound c as input to reaction r.

p_{cr} Number of molecules of compound c produced as output for one tick of reaction r.

P_{cr} Relative presence for compound c as output to reaction r.

Q_S Number of molecules of the source compound S involved in the pathway.

Q_T Number of molecules of the target compound T involved in the pathway.

R Number of reactions in the metabolic network.

S Source compound.

T Target compound.

t_r Number of ticks of reaction r in the pathway.

u_{cr} Binary variable indicating whether the arc from compound node c to reaction node r is in metabolic path.

u_{crk} Binary variable indicating whether the arc from compound node c to reaction node r is in metabolic path k.

v_{rc} Binary variable indicating whether the arc from reaction node r to compound node c is in metabolic path.

v_{rck} Binary variable indicating whether the arc from reaction node r to compound node c is in metabolic path k.

W Number of unbalanced main compounds.

W_c Connectivity of compound c in the metabolic network.

x_e Binary variable indicating whether enzyme e is involved in the pathway.

z_r Binary variable indicating whether reaction r is active in the pathway.

Ψ Specificity of the pathway.

$\Omega_{\alpha\beta}$ Frequency of a given pair of compounds (α,β).

$\Omega^{\alpha\beta}$ Relative frequency for the compound α with respect to the pair (α,β).

Acknowledgements

Firstly, I would like to thank University of Navarra for funding my research. I really appreciate the trust you have placed in me. In particular, I would like to mention to Carlos Bastero, Joseba Campos, Angel Rubio and Javi Santos for their constant support.

Special thanks to John Beasley, my supervisor, for his magnificent guidance, advice and support during this research. I will never forget our discussions in the room M505 about what a metabolic pathway should or should not be. He has brilliantly formed my idea as to what a researcher must be.

Many thanks to my friends at Mathematics Department of Brunel: Rodrigo, Norbert, Sovan and Harry. To my friends at Netherhall House: Peter Brown, Father Joe, Alvaro Tintore, Uncle George, Maiki, Cheles, Sergi, Javi Munoz, Hinoyo, Eoin, Jopi, Villacorta, Alvaro Maestro, Borja Arostegui, Jose Gabriel Erdozain and many others. In particular, I will never forget Manton and PGB. You all managed to convert my stay in England in an unforgettable experience. Of course to all my friends in San Sebastian, Javi Tejero, Martin, Iñaki Garmendia, Dani Pardo, Dani Fernández, Jose Manuel Torres, etc.

Of course thanks to my parents and brothers/sisters, because you are the best of the world. To Paquita, my wife, you only know the effort to get here. This thesis had not been possible without your constant company and support. It is to you whom I dedicate this thesis.

Publications arising from this dissertation

Beasley, J.E. and Planes, F.J. (2007) Recovering metabolic pathways via optimization. *Bioinformatics,* 23(1), 92-98.

Planes, F.J. and Beasley, J.E. (2007) Path finding approaches and metabolic pathways. To appear in *Discrete Applied Maths*.

Planes, F.J and Beasley, J.E. (2008) Critical examination of path finding and path-finding approaches to metabolic pathways. To appear in *Briefing in Bioinformatics*.

Chapter 1

Introduction

1.1 Basic Biology

The basic unit of life is the cell. All living organisms are made of cells which are small membrane-bounded units filled with a concentrated aqueous solution of chemicals, called *cytoplasm*. Each cell is an independent entity, capable of creating copies of itself by growing and dividing into two identical daughter cells. The complete characteristics of an organism are carried by each of its cells. This information is stored within the DNA (*deoxyribonucleic acide*) molecule.

The DNA molecule is a nucleic acid consisting of a double-stranded helix-twisted polymer composed of four basic molecular units called nucleotides. Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases. The four different bases found in DNA are adenine, guanine, cytosine and thymine.

The DNA molecule encodes the information for building the different parts of the cell. This information is located in the genes. A gene consists of a specific segment of DNA that specifies how to generate a (number of) protein(s) in a process called expression. Proteins are fundamental components of all living cells that are necessary for the proper functioning of an organism. In particular many proteins act as enzymes and catalyse biochemical reactions. From the biochemical point of view, proteins are composed of linear chains of amino acids which are linked together by peptide bonds. There are 20 different types of amino acids, each containing an amino group and a carboxyl group.

The expression of the genes eventually leading to the production of proteins, occurs in several stages, and it is often referred as the central dogma of molecular biology, as can be observed in Figure 1.1. We briefly describe below the two stages that constitute the central dogma.

During the first phase, the DNA in the gene is transcribed into *messenger ribonucleic acid* (mRNA). One strand of the DNA double helix is used as a template by the RNA polymerase enzyme to synthesise mRNA: a single-stranded complementary copy of the base sequence in the DNA molecule. The sequence of mRNA is dictated by the order of the nucleotides in the transcribed part of the gene. The base *uracil* replaces *thymine*.

In the translation phase, the mRNA migrates to the cytoplasm. During this step, mRNA goes through different types of maturation processes including one called splicing, which eliminates the non-coding sequences (introns). Then, the mRNA carries coded information to the ribosomes. The function of the ribosome is to take individual amino acids and link them in a chain in the right order, based on the sequence of the mRNA. Once the amino acids are linked into the chain, they are released from the ribosome and fold into a new protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the genetic code, which relates nucleotide triplets to amino acids. Note here that sequences within mRNA may not be translated.

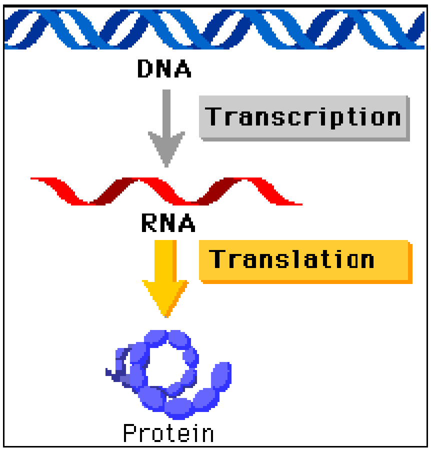

Figure 1.1: Central dogma in molecular biology

In summary, proteins, which are coded by genes, are involved in almost all biological activities. As will become apparent below, proteins are crucial in cellular metabolism, which constitutes the main topic of this thesis.

1.2 Cellular metabolism

Cellular metabolism can be viewed as a complex chemical engine by which the cell produces the energy and material necessary for the maintenance of life. This process is achieved via a vast number of enzyme catalysed biochemical reactions. As introduced above, enzymes are proteins that speed up (catalyse) biochemical reactions so that they occur at significant rates.

In order to better comprehend the intricate functioning of this process, cellular metabolism has been commonly organised into metabolic pathways. Traditionally, biochemistry has defined metabolic pathways as a sequence of enzyme catalysed reactions by which a living organism transforms an initial source compound into a final target compound (Nelson and Cox, 2005). These pathways have been elucidated via experimentation on different organisms. Different authors refer to these pathways using different terms, e.g. consensus pathway (Arita, 2000), annotated pathway (Croes *et al.*, 2005, 2006), experimentally elucidated pathway (Keseler *et al.*, 2005). Henceforth, we refer to them as experimentally determined pathways.

Figure 1.2 shows the glycolate degradation pathway by which two molecules of *glycolate* (glyclt) are converted into one molecule of *3-phospho-D-glycerate* (3pg) by four different enzyme catalysed reactions. The source and target compound, *glycolate* and *3-phospho-D-glycerate* respectively, are coloured yellow. The numbers associated with each arc are the number of molecules of each compound. For example, the enzyme catalysed reaction R45 takes two molecules of *ubiquinone-8* (q8) and two molecules of *glycolate* (glyclt) and transforms them into two molecules of *ubiquinol-8* (q8h2) and two molecules of *glyoxylate* (glx). The numbers in brackets after each reaction label are the number of ticks. The number of ticks defines the activity of a biochemical reaction in a particular metabolic pathway. R460, for example, whose stoichiometry is 2 glx + h → 2h3oppan + co2, “ticks” once in our example pathway, converting two molecules of *glyoxylate* (glx) and one molecule of *hydrogen ion* (h) into one molecule of *2-Hydroxy-3-oxopropanoate* (2h3oppan) and one molecule of *carbon dioxide* (co2). R45, whose stoichiometry is glyclt + q8 → glx + q8h2, "ticks" twice in the glycolate degradation pathway, each time converting one molecule of *ubiquinone-8* (q8) and one molecule of *glycolate* (glyclt) into one molecule of *ubiquinol-8* (q8h2) and one molecule of *glyoxylate* (glx). Clearly, the number of ticks of a biochemical reaction that is not involved in the pathway is zero.

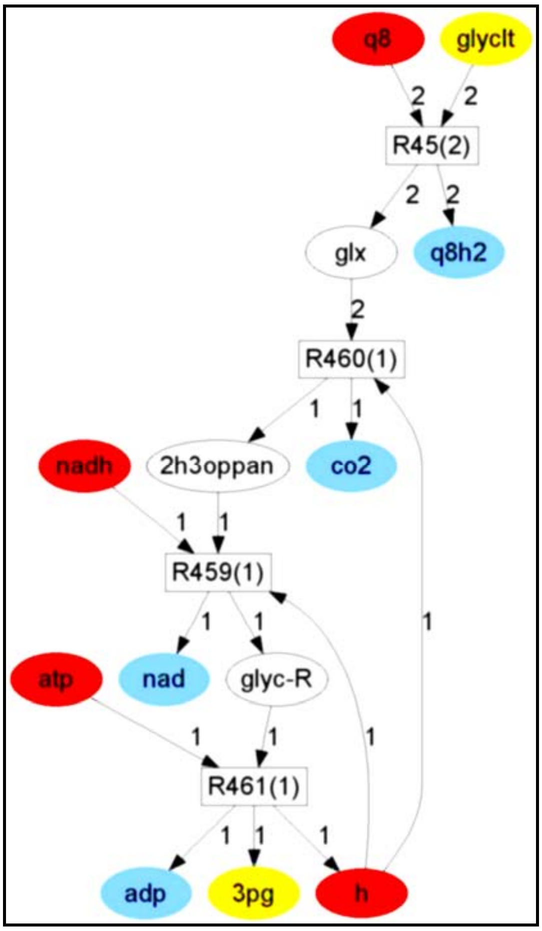

Figure 1.2: Glycolate degradation pathway

Compounds coloured blue in Figure 1.2 are produced to excess (the total number of molecules consumed by the reactions involved in the pathway is less than the total number of molecules produced by the reactions involved in the pathway). For example, *carbon dioxide* (co2) is coloured blue because R460 (2 glx + h→ 2h3oppan + co2), which ticks once, produces one molecule of co2 and no reaction consumes co2. Thus, co2 is, in aggregate (net) terms, produced in the pathway. Similarly, compounds coloured red are freely available (the total number of molecules consumed by the reactions involved in the pathway is greater than the total number of molecules produced by the reactions involved in the pathway). For example, *hydrogen ion* (h) is coloured red because R460 (2 glx + h→ 2h3oppan + co2), which ticks one, consumes one molecule of h, R459 (2heoppan + h + nadh → glyc-R + nadh), which ticks one, consumes one molecule of h and R461 (glyc-R + atp → 3pg + adp + h), which ticks once, produces one molecule of h. Thus, h is, in aggregate (net) terms, consumed in the pathway. Finally, compounds coloured white are balanced (the total number of molecules consumed by the reactions involved in the pathway is equal to the total number of molecules produced by the reactions involved in the pathway). For example, *glyoxylate* (glx) is coloured white because R45 (glyclt + q8 → glx + q8h2), which ticks twice, produces two molecules of glx and R460 (2 glx + h→ 2h3oppan + co2), which ticks once, consumes two molecules of glx. Thus, glx is, in aggregate (net) terms, balanced in the pathway.

We should note in passing here that the representation of a metabolic pathway as in Figure 1.2 is not usual in the literature or databases that are available. This representation, which we developed, enables one to more clearly see a metabolic pathway than the standard representations given in the literature/databases.

Experimental findings have provided insight into metabolic pathways and how organisms commonly undertake their mass-energetic requirements (Nelson and Cox, 2005). However, such findings have several shortcomings. First and foremost, there are few organisms in which the experimentally determined pathway structure is complete. To the best of our knowledge, only Escherichia Coli has a full description available, Karp *et al.*, 2002b. Secondly, experimental findings do not provide an answer as to why the cell makes use of a particular pathway (via evolution) and not another different pathway. For example, referring to Figure 1.2, one might find different possibilities to convert two molecules of *glycolate* into one molecule of *3-phospho-D-glycerate*. In addition, the absence of a general logic (or mathematical model) that provides detailed insight into specific metabolic pathways makes it difficult to answer other practical questions related to metabolism, such as the discovery of novel alternatives pathways for biomedical and pharmacological issues or the regulation of different metabolic pathways.

In order to direct the analysis of metabolic pathways towards these fundamental questions, computational/mathematical methods emerged in the late 80's (Seressiotis and Bailey, 1986, 1988; Mavrovouniotis, 1992a, 1992b, 1993). Given a set of biochemical reactions belonging to a particular organism or cell, different algorithms based on artificial intelligence were used to compute meaningful metabolic pathways from a source compound to a target compound. However, the set of biochemical reactions was generally too incomplete to carry out a valid analysis.

Early in this century, the availability of the genome sequences, along with the effort for storing in metabolic databases (Kanehisa and Goto, 2000; Schomburg *et al.*, 2002; Joshi-Tope *et al.*, 2003) biochemical reaction data from literature sources, made it possible to draft the complete set of biochemical reactions for a particular organism (Schilling *et al.*, 2003; Förster *et al.*, 2003; Becker and Palsson, 2005; Feist *et al.*, 2007). Essentially, each metabolic gene of a particular organism is assigned a biochemical reaction. As explained above, the link between genomes and metabolism is that some (metabolic) genes codes for enzymes, which are proteins that catalyse biochemical reactions. This assignment can be found directly through experimental methods (e.g. enzyme isolation). However, experimental methods are typically time-consuming. For this reason, much effort has been expended in finding analytical/computational methods that automatically carry out this task (Karp *et al.*, 1999). For example, DNA sequence homology is typically used as a strong evidence for the presence of a reaction in an organism. These methods assume that two genes (from different organisms) that are homologous (have a similar sequence of nucleotides) have the same function. Therefore, should we know the function of one of these two genes, we know the function of the other gene. These computational procedures are not completely reliable. Indeed, the existence of the inferred reaction is hypothetical until an experimental validation is done. However, they provide a reference for future improvements (Palsson, 2006).

This process of inferring the set of biochemical reactions for a particular organism is usually referred as to metabolic reconstruction. The interaction of the entire set of biochemical reactions is usually characterised as a directed graph and referred to as a genome-scale metabolic network. Figure 1.3 shows an example network, represented as a bipartite graph, comprising 8 reactions (labelled R1 to R8 respectively) and 8 compounds (labelled C1 to C8 respectively). Reaction R3, for example, converts one molecule of C5 into two molecules of C4, one molecule of C6 and one molecule of C7. Each reaction has a specified direction so a reversible reaction contributes two different reactions. For example, R6 and R7 are the reverse of each other. We should like to state here that such a network is an input data to our research.

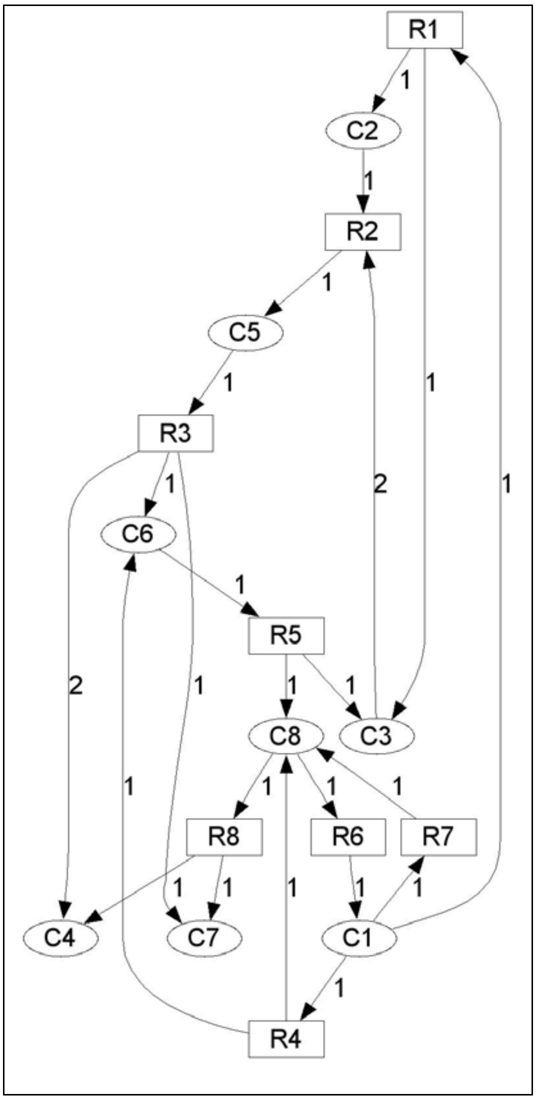

Figure 1.3: An example metabolic network

The availability of the complete set of biochemical reactions for an organism propelled the appearance of computational/mathematical approaches so as to carry out an analysis of metabolic pathways in genome-scale metabolic networks. In Chapter 2 a survey of the mathematical/computational approaches reported so far in the literature is provided. For the sake of clarity, approaches have been divided into two types: stoichiometric approaches and path finding approaches. Essentially, stoichiometric approaches, in contrast to path finding approaches, make use of the stoichiometry of the reactions when trying to find meaningful metabolic pathways in the metabolic network. The stoichiometry of R2 in Figure 1.3, for example, is that one molecule of C2 and two molecules of C3 are converted into one molecule of C5. Typically, both types of approaches have focused on determining a particular set of pathways in which it is expected to find experimentally determined pathways. Although somewhat more emphasised in path finding approaches, little validation has been carried out in this respect. As will become apparent in Chapter 2, path finding approaches, which are based on graph theory, turn out to be too simplistic to model metabolic pathways. In addition, stoichiometric approaches are currently inefficient and impractical at the genome-scale, since they are concerned with obtaining a special set of pathways and this set explodes in size in a combinatorial fashion as the number of reactions in the metabolic network increases.

1.3 Contribution

In this thesis a novel mathematical framework based on combinatorial optimisation (integer linear programming) to analyse metabolic pathways is presented. As will be shown in the thesis, we think that optimisation is a suitable concept to analyze metabolic pathways for two reasons:

- it presents more flexibility to add meaningful biological constraints than typical path finding and stoichiometric approaches;
- it focuses on computing single biologically meaningful metabolic pathways instead of particular sets of pathways. This simplifies the analysis and computational effort at the genome-scale.

In order to validate whether (or not) our mathematical optimisation models produce biologically meaningful metabolic pathways, the research has been mainly focused upon the problem of recovering experimentally determined metabolic pathways from the metabolic network. This problem consists of defining a mathematical optimisation model that, when solved, provides a solution identical to a known experimentally determined pathway, given the following input data:

- the metabolic network of a particular organism;
- the source and target compound of the experimentally determined pathway;
- and (perhaps) some pathway knowledge.

Note here that, as we are concerned with investigating whether (or not) there exists a mathematical optimisation model that underlies experimentally determined (known) metabolic pathways, then to do this, we need to verify whether (or not) the optimal solution to our model is equal to the actual pathway. If we use a heuristic algorithm, then we can never be sure as to the outcome. For example, if the heuristic solution is equal to the actual pathway, we can never be sure if our model is good or it just indicates that our model has (by chance) found a non-optimal solution, which is equal to the actual pathway. For this reason the solution to our model must guarantee optimality.

The initial mathematical model for recovering experimentally determined metabolic pathways, referred as to the Beasley-Planes (BP) model, is presented in Chapter 3. The BP model is fundamentally an stoichiometric approach. It is applied to forty experimentally determined pathways of *Escherichia Coli*, which is a well-known organism in the biological world. Although the BP model shows quite good results, the BP model needs to know the low presence unbalanced (whether produced to excess or freely available) compounds in the experimentally determined pathways. A low presence compound is one that appears in very few of the reactions in the metabolic network. The precise numeric definition of a low presence compound is given in Chapter 3. This limitation constrains the scope of the approach, as many pathways include low presence unbalanced compounds.

Chapter 4 presents a practical application of the BP model to best disrupt metabolic pathways. We distinguish two cases: the disruption of a single metabolic pathway; and the disruption of two (related) metabolic pathways. The study is carried out with ATP producer pathways, which have special interest in cancer research.

In Chapter 5, we develop a path finding approach based on integer linear programming so as to examine in detail the accuracy of such approaches. This approach served as inspiration to derive the Improved Beasley-Planes (IBP) model, which is described in Chapter 6. Although the IBP model presents slightly worse results with respect to the BP model, the IBP model does not need any further knowledge of the experimentally determined pathway aside from the source and target compound.

Finally, a discussion is carried out in Chapter 7 so as to summarize the conclusions and future lines of research. Appendices providing detailed results and presenting full details of the forty metabolic pathways we have considered in this thesis are also given.

Chapter 2

Literature Survey

Advances in the field of genomics have made it possible to draft the complete set of biochemical reactions involved in different organisms. Accordingly, a computational analysis of metabolic pathways at the genome-scale has become possible. Singular attention has been devoted in the literature to stoichiometric approaches and path finding approaches to metabolic pathways. Essentially, stoichiometric approaches make use of the stoichiometry of the reactions when trying to determine meaningful metabolic pathways. In contrast, path finding approaches propose an alternative view based on graph theory in which the stoichiometry of the reactions is not considered.

In this chapter we give an overview of the theory, applications and challenges of stoichiometric approaches and path finding approches to metabolic pathways. One point to note here is that, given the volume of papers relating to metabolic pathways that have been published, it is impossible to review each and every such paper. Rather we have chosen to focus on selected papers that seem to us, either from our own reading or from knowledge of citations given by others, of especial relevance.

2.1 Introduction

Cellular metabolism is a highly complex biological process by which the cell produces the energy and material necessary for the maintenance of life. In order to better comprehend the intricate functioning of this process, cellular metabolism has been commonly organised into metabolic pathways. Traditionally, biochemistry has defined metabolic pathways as a sequence of enzyme catalysed reactions by which a living organism transforms an initial source compound into a final target compound (Nelson and Cox, 2005). These pathways have been elucidated via experimentation on different organisms. A comprehensive collection of experimentally determined metabolic pathways are available in different electronic databases (Selkov *et al*., 1996, 1998; Karp *et al*., 2002a, 2002b).

For those organisms in which experimentally determined metabolic pathway knowledge is sparse, the complete availability of genome sequence data has enabled metabolic pathways to be inferred computationally. A number of qualitative tools have been developed to achieve this by assigning identified enzymes to experimentally determined (reference) pathways (Karp *et al.*, 1999; Overbeek *et al.*, 2000; Kanehisa *et al.*, 2000). However, the scope of these tools is restricted by prior knowledge of experimentally determined metabolic pathways and they find difficulties, for example, in the discovery of novel alternative metabolic pathways which may be of interest for biotechnological or biomedical reasons.

Advances in genomics have made it possible to draft the entire set of biochemical reactions involved in a particular organism or cell, along with their underlying data: compounds, stoichiometry, reversibility, enzymes and genes (Schilling *et al.*, 2002; Reed *et al.*, 2003). The interaction of the entire set of biochemical reactions and compounds is usually characterised as a directed graph and referred to as a genome-scale metabolic network. Referring back to Chapter 1, Figure 2.1 shows an example network, represented as a bipartite graph, comprising 8 reactions (labelled R1 to R8 respectively) and 8 compounds (labelled C1 to C8 respectively). Reaction R3, for example, converts one molecule of C5 into two molecules of C4, one molecule of C6 and one molecule of C7. Each reaction has a specified direction so a reversible reaction contributes two different reactions. For example, R6 and R7 are the reverse of each other. Such graph-theoretic representations provide a framework for a complete computational search for functional metabolic pathways within the genome-scale metabolic network.

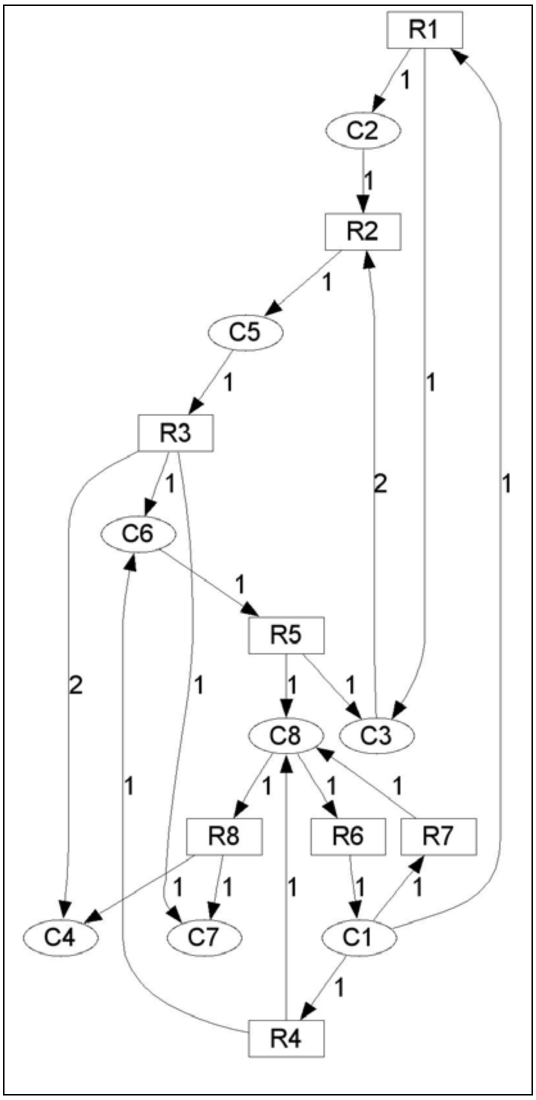

Figure 2.1: Previous Figure 1.2 in Chapter 1

Although not explicitly shown as such Figure 2.1 is a bipartite graph, since there are no reaction to reaction arcs, nor any compound to compound arcs. Although other graphical representations are possible (Deville *et al.*, 2003) we have found the representation shown to be the most useful.

To illustrate approaches to the computational analysis of metabolic pathways suppose that we are concerned with finding pathways which convert C1 into C7 in the network shown in Figure 2.1. A pathway, as illustrated in Figure 2.2a, will be a subgraph of this entire network satisfying the condition that for each reaction node included in the subgraph all compound nodes associated with that reaction node (either as an input compound or as an output compound) also appear in the subgraph. In Figure 2.2a the subgraph comprises reactions nodes R1, R2 and R3, plus all of their associated compound nodes. As one might expect, within the metabolic network there may be more than one pathway. Figure 2.2b shows another possible pathway that convert C1 into C7.

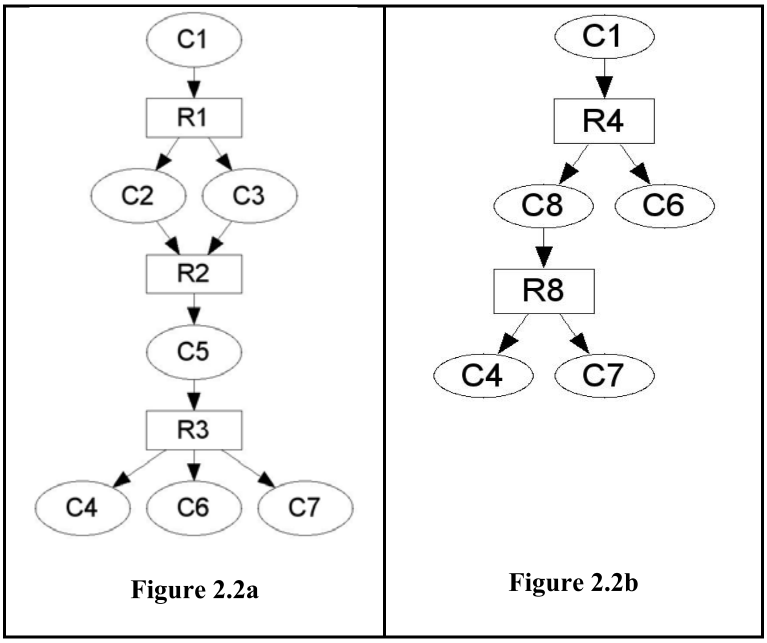

Figure 2.2a

Figure 2.2b

Figure 2.2: Two possible pathways for converting C1 into C7

However, not each possible pathway will be biologically meaningful, i.e. will have a valid biochemical/physiological interpretation in the cell/organism. Thus, approaches to metabolic pathways aim to find biologically meaningful metabolic pathways within the metabolic network. Clearly, the ability of the approach to find experimentally determined pathways will show the reliability of any approach to detect (novel) biologically meaningful metabolic pathways.

Computational/mathematical approaches to metabolic pathways can be conveniently divided into two types: stoichiometric approaches and path finding approaches. As will become apparent in the following sections, there are two key distinctions between these approaches:

- stoichiometric approaches force pathways to satisfy biologically meaningful constraints related to compounds stoichiometry, path finding approaches do not.

- path finding approaches search for linear sequences of biochemical reactions (directed paths) in the metabolic network from a given source compound to a target compound, stoichiometric approaches do not.

As stoichiometric approaches appeared earlier in the literature, we begin by describing these approaches.

2.2 Stoichiometric approaches

Stoichiometric approaches aim to find pathways in which the compound nodes satisfy a variety of different, but biochemically meaningful, stoichiometric constraints. Following the example described above, suppose now we are concerned with the problem of finding pathways converting a source compound, C1, into a target compound, C7, such that a subset of compounds, for example, C2, C5 and C8, are balanced. Here by balanced we are referring to a stoichiometric related balance, namely a compound is balanced if the total number of molecules consumed by the reactions involved in the pathway is equal to the total number of molecules produced by the reactions involved in the pathway.

Unbalanced compounds are either, in aggregate (net) terms, produced or consumed. Clearly the source compound, C1, must be, in aggregate, consumed (total number of molecules of C1 consumed by the reactions involved in the pathway is greater than the total number of molecules of C1 produced by the reactions involved in the pathway). Similarly the target compound, C7, must be, in aggregate, produced (total number of molecules of C7 consumed by the reactions involved in the pathway is less than the total number of molecules of C7 produced by the reactions involved in the pathway). Note here that, for the purposes of illustration, we assume that the source and target compound are different compounds.

The reason to balance a subset of compounds, such as C2, C5 and C8, is that the some compounds can be regarded as present in the organism (e.g. cell) as represented by the metabolic network (such as in Figure 2.1) purely as intermediate compounds necessary for producing other, more crucial, compounds. Differently, C3, C4 and C6 are stoichiometrically unconstrained, i.e. they could be produced, balanced or consumed. They represent cofactors as well as other compounds such as water or hydrogen (ions). A cofactor is generally defined as a biochemical compound that fulfils the same specific and secondary function in a considerable number of reactions. Cofactors are typically coupled in the biochemical reactions: one as input compound and the other as output compound. e.g. atp-adp. To illustrate this, observe, for instance, the following biochemical reactions: *glu-C* + *atp* → *g6p* + *adp*. Clearly, the principal biotransformation is that converting *glu-C* (D-Glucose) into *g6p* (D-Glucose 6-phosphate). The *atp* (Adenosine Triphosphate) *–adp* (Adenosine Diphosphate) pair undergoes the specific function of giving a phosphate group (-P) to *glu-C* so that this compound converts into *g6p*. Thus, *atp-adp* pair behaves here as a cofactor. One point to note here is that the list of cofactors for a given metabolic network has not been defined unambiguously. This is due to the fact that some biochemical compounds might (or not) behave as cofactors. Indeed, one might find biochemical reactions in which *atp* conducts the principal biotransformation and does not behave as a cofactor. In this reaction, for example, *atp* + *h* + *nmn* → *nad* + *ppi*, it is not clear *atp* operates as a cofactor.

Other approaches (Schuster *et al*., 2000; Schilling *et al*., 2000) include exchange reactions in the metabolic network so as to balance (if necessary) compounds. An exchange (transfer) reaction usually involves a physical transportation from the internal organism to the external environment. For instance in Figure 2.1 both C4 and C7 have no reactions to which they provide input, and hence must both have exchange reactions, namely C4→C4(external) and C7→C7(external), otherwise they would build up as the organism operates. However, the criteria as to how exchange reactions are included has not been unambiguously defined (Klamt and Stelling, 2003). Moreover explicitly adding to Figure 2.1 compounds such as C4(external) and C7(external) merely (in graph-theoretic terms) transfers the problem of unbalanced compounds from the internal compounds {C4,C7} to the external compounds {C4(external),C7(external)}, since these external compounds can only themselves be balanced by considering other organisms. For the sake of simplicity therefore, we will neglect exchange reactions here, assuming that unbalanced compounds may be balanced (if necessary) by such reactions.

In order to consider stoichiometric approaches mathematically let R be the total number of reactions in the metabolic network and let t_r be the number of ticks of reaction r in the pathway under consideration. This variable must take integer values, having the value zero if the reaction is not used. The tick variable for a reaction relates to its associated stoichiometry. In Figure 2.1, for example, one tick of reaction R3 converts one molecule of C5 into two molecules of C4, one molecule of C6 and one molecule of C7. Two ticks of reaction R3 converts two molecules of C5 into four molecules of C4, two molecules of C6 and two molecules of C7; three ticks of reaction R3 converts three molecules of C5 into six molecules of C4, three molecules of C6 and three molecules of C7; etc.

Clearly a particular pathway can be fully represented by its associated tick vector, $[t_1, t_2, ..., t_R]$. In other works (Schuster *et al.*, 2000; Schilling *et al.*, 2000) the ticks are usually referred to as reaction fluxes. Differently, reaction fluxes are allowed to take continuous values (whereas ticks take integer values). It is important to note that both reaction fluxes and ticks are proportional to the absolute fluxes operating inside the cell. An absolute flux is a physiological variable that measures the rate at which the concentration of substrates are degraded (and products synthesised) per unit of time in a particular reaction. The absolute fluxes depend on many different factors (other pathways, environmental conditions, kinetics, etc). Because we are concerned with the structural non-dynamic problem of finding pathways, it seems to appropriate to define the tick vector as a discrete (integer valued) vector.

Let s_{cr} be the stoichiometric coefficient of compound c in reaction r (defined as, for one tick of reaction r, the number of molecules of compound c produced as output minus the number of molecules of compound c consumed as input). Then compound c constrained to being, in aggregate (net) terms, consumed implies that $\sum_{r=1}^{R} s_{cr}t_r < 0$ must be satisfied.

Here, as mentioned above, we are interested in pathways from C1 to C7. In order to ensure that the source compound C1 is consumed, pathways in the metabolic network shown in Figure 2.1 must satisfy:

$$t_6 - t_1 - t_4 - t_7 < 0 \tag{2.1}$$

A compound c constrained to being, in aggregate (net) terms, produced must satisfy $\sum_{r=1}^{R} s_{cr}t_r > 0$. Hence, in order to ensure that the target compound C7 is produced, pathways in the metabolic network shown in Figure 2.1 must satisfy:

$$t_3 + t_8 > 0 \tag{2.2}$$

Finally, a compound c constrained to being, in aggregate (net) terms, balanced must satisfy $\sum_{r=1}^{R} s_{cr}t_r = 0$. Hence, in order to ensure that {C2,C5,C8} are balanced, pathways in the metabolic network shown in Figure 2.1 must satisfy:

$$t_1 - t_2 = t_2 - t_3 = t_4 + t_5 + t_7 - t_6 - t_8 = 0 \tag{2.3}$$

Figure 2.3a shows an example pathway that satisfies constraints (2.1)-(2.3). Here {C2,C5,C8} are balanced, two of these compounds {C2,C5} being involved in the pathway, the other {C8} not being involved. In Figure 2.3a the source and target compounds (C1 and C7 respectively) are coloured yellow. Compounds coloured red are (in net/aggregate terms) consumed in the pathway and those coloured white are balanced. The numbers in brackets after each reaction label are the number of ticks. For example reaction R1 ticks one, thereby converting one molecule of C1 into one molecule of C2 and C3. The tick vector for this pathway has the value [1, 1, 1, 0, 0, 0, 0, 0].

The criteria by which the stoichiometric constraints are systematically imposed is clearly crucial, as the constraints define the set of possible pathways converting C1 into C7. Indeed, different stoichiometric constraints produce a different solution space. Figure 2.3b, for example, shows a pathway converting C1 into C7 that satisfies different stoichiometric constraints, {C3, C5, C8} balanced. Compounds coloured blue are (in net/aggregate terms) produced in the pathway. Since C2 is not balanced in Figure 2.3b, this pathway would not satisfy the stoichiometric constraints defined in equation (2.3).

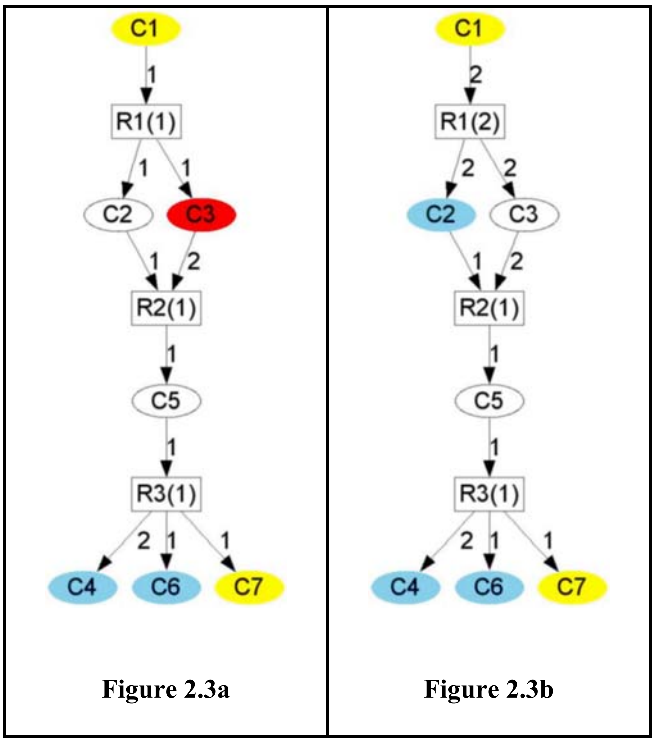

Figure 2.3a

Figure 2.3b

Figure 2.3: Two possible pathways for converting C1 into C7

Stoichiometric approaches have focused on the identification of all the pathways that satisfy given stoichiometric constraints. The key assumption here is that experimentally determined pathways are expected to satisfy the stoichiometric constraints, thus the experimentally determined pathway should be found amongst the complete set of pathways satisfying the mathematical stoichiometric constraints.

However, typically the large number of pathways that satisfy the stoichiometric constraints means that complete enumeration of these pathways would be computationally impracticable. This issue is usually referred to as "combinatorial explosion", and is mentioned further below. Subsequently, as described in the next section, stoichiometric approaches have developed to focus on finding a set of pathways capable of spanning the complete solution space of pathways defined by the stoichiometric constraints.

The next section describes the evolution of this set of pathways from genetically independent pathways as proposed by Seressiotis and Bailey, 1986, 1988, to extreme pathways as proposed by Palsson and co-workers (Schilling *et al.*, 2000).

2.2.1 Literature review

The work of Seressiotis and Bailey, 1986, 1988, constitutes the first stoichiometric approach for the computational analysis of metabolic pathways. They developed an algorithm based on artificial intelligence concepts so as to find a set of genetically independent pathways by the successive addition of reaction steps transforming a source compound into a target compound. For a pathway P to be genetically independent it means that there exists no other pathway (from source to target) utilising just a subset of the enzymes used in P. In their approach compounds involved in the pathway are constrained to being balanced, except for the source/target compounds and a cofactor set (although the constituents of this set are not clearly defined). They dealt with only a relatively small metabolic network (70 reactions, 100 compounds). However, the computational effort required of the algorithm prohibited its application to larger networks.

This work is of importance due to the fact that the concept of genetically independent pathways was established. In order to clarify this concept, consider three different pathways from the network shown in Figure 2.1 that convert C1 into C7 and satisfy the constraint that C2, C5 and C8 are balanced. In terms of their tick vector representation these are: P_1 = [1, 1, 1, 0, 0, 0, 0, 0], P_2 = [0, 0, 0, 1, 0, 0, 0, 1] and P_3 = [1, 1, 1, 0, 1, 0, 0, 1]. In the approach that Seressiotis and Bailey, 1986, 1988, adopted it was wrongly assumed (albeit based on what was known at the time) that we have a one-to-one unique association between gene-enzyme and enzyme-reaction. In other words one gene is uniquely associated with one enzyme and that enzyme in turn is uniquely associated with one reaction. Under this assumption let the genes be g1,g2,…,g8 each (by implication) having a unique one-to-one association with R1,R2,…,R8 respectively. Thus each metabolic pathway (defined by its corresponding tick vector) can be assigned a specific set of genes, referred as to genotypes, namely

G_1={g1,g2,g3} to P_1, G_2={g4,g8} to P_2 and G_3={g1,g2,g3,g5,g8} to P_3. Both G_1 (associated with P_1) and G_2 (associated with P_2) are genetically independent, since in each case no subset can generate a pathway from C1 into C7 that satisfies the stoichiometric constraints. However G_3 (associated with P_3) is dependent as it contains a subset, namely G_1 (where $G_1 \subset G_3$), that is a pathway from C1 to C7 satisfying the stoichiometric constraints. Hence, only P_1 and P_2 are genetically independent. Notice here that the concept of genetic independence differs from the concept of linear independence. Here, P_1, P_2 and P_3 are linearly independent, in particular there do not exist scalars $\alpha_1,\alpha_2 \in (-\infty, +\infty)$ such that $P_3 = \alpha_1 P_1 + \alpha_2 P_2$.

Genetic independence is defined above in relation to whether, or not, a single pathway contains a valid subset that is also a pathway. However, since our focus in this section is on sets of pathways, it is convenient to define genetic independence here in a slightly different way. Let P^* be the set of all pathways which satisfy the stoichiometric constraints, ρ be a particular pathway belonging to P^*, and G_ρ be the genotype associated with ρ. Then a pathway ρ is genetically independent if and only if:

$$\text{there does not exist } \tau \in P^* \text{ such that } G_\tau \subset G_\rho$$

This condition says that a pathway ρ is genetically independent if there is no other pathway τ for which the genotype G_τ is a subset of G_ρ.

In order to clarify the concept of genetically independent pathways, we have calculated manually (as not available software to do it) the list of genetically independent metabolic pathways within the metabolic network shown in Figure 2.1 that convert C1 into C7, subject to C2, C5 and C8 are balanced. We obtained the three genetically independent pathways shown in Figure 2.4.

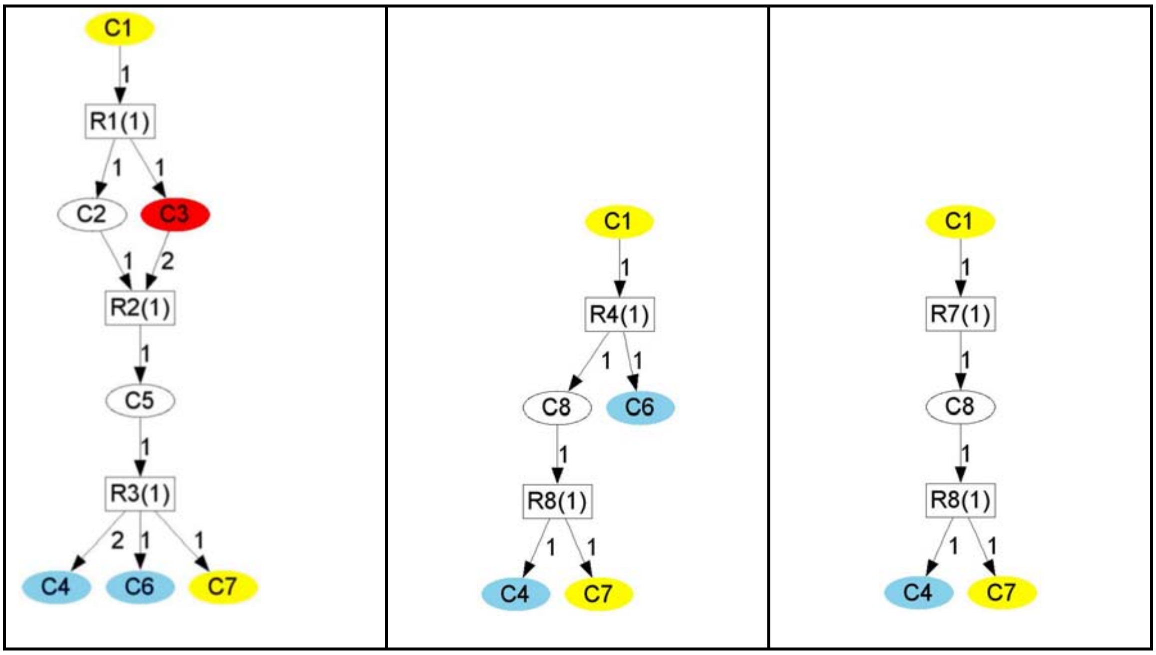

Figure 2.4: Three genetically independent pathways

Mavrovouniotis, 1992a, 1992b, 1993, presented an algorithm to generate a set of genetically independent pathways satisfying a given set of stoichiometric constraints. Differently to Seressiotis and Bailey, 1986, 1988, the algorithm successfully deals with pathways which comprise multiple sources and targets. However, it requires the specification of compound and reaction status (e.g. which compounds must be consumed in the pathway). The algorithm was successfully applied to metabolic networks of moderate size (250 reactions, 400 compounds).

After these early approaches, much effort was expended to provide a more theoretical foundation to the study of metabolic pathways. Different works on inorganic reaction networks (Milner, 1964; Clarke, 1980; Happel and Sellers, 1982, 1989) served as inspiration for subsequent stoichiometric approaches. In these approaches the cell is (essentially) considered as an open system, consuming nutrients and producing biomass, in which enzymatic reactions interact to produce an overall global flux distribution. Let f_r be the absolute flux of enzymatic reaction r, x_c be the concentration of compound c and b_c be the external flux for compound c, then applying the law of conservation of mass we have that $dx_c/dt = b_c + \sum_{r=1}^{R} s_{cr}f_r \ \forall c$.

The compound set is divided into two subsets: internal compounds, I, those compounds which cannot traverse the physical boundaries of the cell, and external compounds, E, those compounds able to traverse the physical boundaries of the cell. The external flux b_c is equal to zero for internal compounds, whilst for external compounds it can be different from zero.

The main assumption here is that the concentration of internal compounds remains constant over time. This is usually referred to as the pseudo steady state (henceforth PSS) condition. Thus, internal compounds satisfy $dx_c/dt = b_c + \sum_{r=1}^{R} s_{cr}f_r = 0 \ \forall c \in I$. As $b_c = 0 \ \forall c \in I$, this is $\sum_{r=1}^{R} s_{cr}f_r = 0 \ \forall c \in I$. We regard the system as being in a PSS condition since, when studying aspects of metabolism related to growth, the time constants associated with growth are much larger than those associated with individual reaction kinetics (Schilling *et al.*, 2000).

In the context of PSS systems, metabolic pathways are regarded as structural elementary units in the cell that preserve the steady state equilibrium for the whole cell. Interestingly, this view of metabolic pathways provides a link between the dynamic (absolute fluxes) and structural (ticks) view of metabolic pathways. Hence the concept of a metabolic pathway was redefined to be a set of enzyme catalysed biochemical reactions that satisfies two conditions: (i) the PSS condition; and (ii) a simplicity condition.

A pathway at PSS satisfies the condition that the internal compounds appearing in the pathway are stoichiometrically balanced, i.e. $\sum_{r=1}^{R} s_{cr}t_r = 0 \ \forall c \in I$. Note here that external compounds ($c \in E$) are stoichiometrically unconstrained, i.e. they could be produced, consumed or balanced. Indeed, they are not necessarily involved in the pathway. This is fundamentally different from previous approaches (Seressiotis and Bailey, 1986, 1988; Mavrovouniotis, 1992a, 1992b, 1993) in which pathways necessarily consumed a source compound(s) and produced a target compound(s).

In order to define the simplicity condition, the genetic independence condition presented by Seressiotis and Bailey, 1986, 1988, was reformulated to develop the non-decomposability condition. In the early work of Seressiotis and Bailey it was wrongly assumed, as mentioned above, that there is a one to one unique association between gene-enzyme and enzyme-reaction. In the post-genomic era is clear that a gene can express different enzymes and different enzymes can catalyse the same reaction. Let Q^* be the set of all pathways which satisfy the PSS condition, ρ be a particular pathway belonging to Q^*, and F_ρ be the set of reactions involved in pathway ρ. Then, a pathway ρ is non-decomposable if and only if:

$$\text{there does not exist } \tau \in Q^* \text{ such that } F_\tau \subset F_\rho$$

This condition says that a pathway ρ is non-decomposable if there is no other pathway τ for which the reaction set F_τ is a subset of F_ρ. Referring to the definition of genetically independent pathways above it is clear that this definition for non-decomposable pathways and that, are very similar – only one deals with genotypes and the other with reaction sets.

In order to illustrate the non-decomposability condition, we present below a non-decomposable pathway in Figure 2.5 and a decomposable pathway in Figure 2.6. These pathways were determined from the example metabolic network shown in Figure 2.1. We assumed that C2, C5 and C8 are internal compounds, while C1, C3, C4, C6 and C7 external compounds. The pathway in Figure 2.5 is non-decomposable since no reaction subset satisfies the stoichiometric constraints (equations (2.1)-(2.3)) related to internal compounds. Instead, pathway in Figure 2.6 can be decomposed into two different sub-pathways satisfying the stoichiometric constraints. One of them is precisely the pathway shown in Figure 2.5.

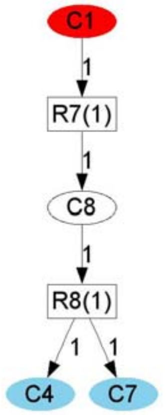

Figure 2.5: Non-decomposable pathway

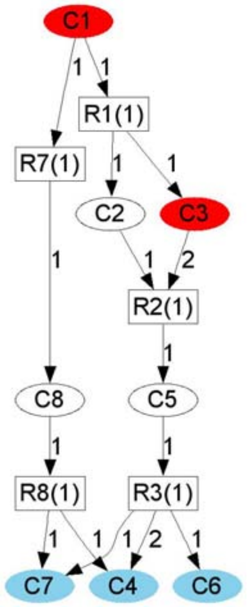

Figure 2.6: Decomposable pathway

2.2.1.1 Elementary flux modes

Non-decomposable pathways at PSS were termed elementary flux modes (henceforth EFMs) by Schuster and co-workers (Schuster and Hilgetag, 1994). The set of EFMs, $\{e_1,e_2,...,e_q\}$, is finite and unique for a given metabolic network and a given classification of compounds as internal or external. In order to clarify this concept we computed (using the YANA software package (Schwarz *et al.*, 2005)) the set of EFMs for the example metabolic network in Figure 2.1, assuming that {C2,C5,C8} are internal compounds, {C1,C3,C4,C6,C7} are external compounds. Note here that algorithms for computing EFMs effectively deal with reversibility by allowing reversible reactions to take negative values. For this reason it is not necessary to split a reversible reaction into two different reactions, as was done with R6 and R7 in Figure 2.1. Subsequently, our input network to YANA included the following reactions: {R1, R2, R3, R4, R5, R6, R8}, with R6 reversible. We obtained the six elementary flux modes $\{e_1,e_2,...,e_6\}$ shown in Figure 2.7. In particular, the fifth EFM has a negative value for R6, which (for convenience) is depicted here by R7. A SBML file for input to YANA containing the network shown in Figure 2.1 is available from http://people.brunel.ac.uk/~mastjjb/jeb/network.html.

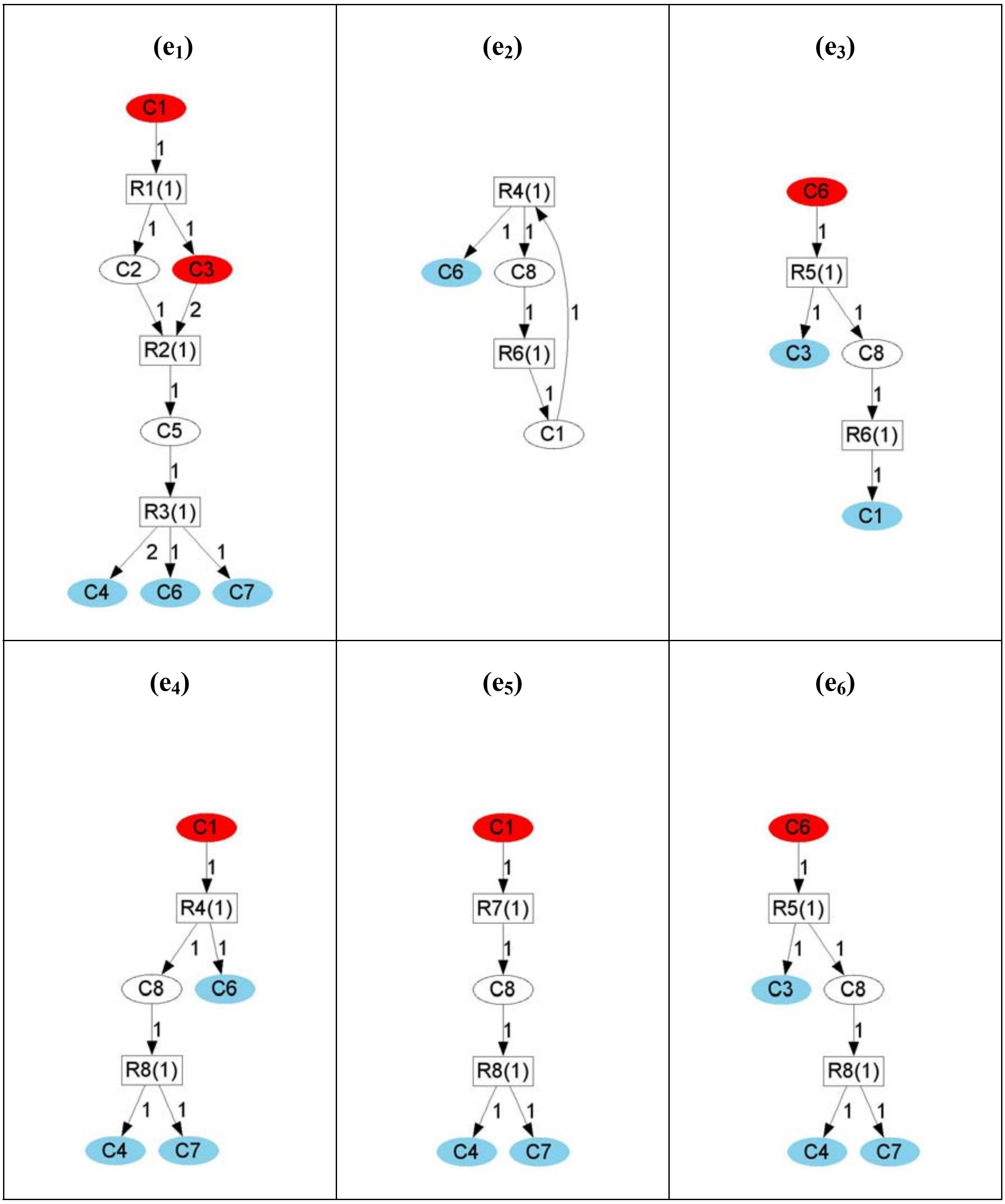

Figure 2.7: The six elementary flux modes

Different mathematical properties of the set of elementary flux modes were presented in a later work (Schuster *et al*., 2002a). One important characteristic is that all pathways at pseudo steady state, whether decomposable or not, can be expressed as a non-negative linear combination of the set of EFMs, i.e. as $\sum_{i=1}^{q} \alpha_i e_i$ where $\alpha_1, \alpha_2, ..., \alpha_q \geq 0$.

An algorithm to compute the complete set of EFMs was outlined in Schuster *et al.*, 2000. The algorithm was implemented in an open-source program named METATOOL (Pfeiffer *et al.*, 1999). Different improvements to the algorithm can be found in Gagneur and Klamt, 2004, and Urbanczik and Wagner, 2005. These algorithms have been shown to be effective for networks of small or moderate size. However, the calculation of the complete set of EFMs is computationally demanding in that increasing the size of the metabolic network causes a combinatorial explosion in the number of modes. For example, for a network with 13 reactions and 19 compounds, describing part of the glycolysis pathway and the non-oxidative pentose phosphate pathway, the number of EFMs is 14 (Pfeiffer *et al.*, 1999). However for a network with 112 reactions and 89 compounds, describing the central metabolism in *E.Coli*, the number of EFMs is 2450787 (Kamp and Schuster, 2006). An analysis of the combinatorial explosion in the number of EFMs can be found in Klamt and Stelling, 2002.

Different strategies have been suggested in order to overcome the combinatorial explosion in EFMs that appears in genome-scale metabolic networks. Schuster *et al.*, 2002b, proposed an algorithm to decompose the network into several sub-networks based on the connectivity of compounds, where the connectivity is defined by the number of reactions in which a compound participates either as an input compound or as an output compound. The logic is that if a sufficient number of compounds, initially constrained to be internal compounds, are considered external in addition to nutrients and biomass, the system disintegrates into subsystems, thereby each internal compound belongs to only one subsystem. The compounds defined to decompose the network are those whose connectivity is greater than four. The decomposition method was applied to the mycoplasma pneumoniae network, resulting in 19 sub-networks, in each of which the EFMs could be computed easily. In addition, Dandekar *et al.*, 2003, presented a stochastic optimisation program based on the Metropolis algorithm to find the classification of compounds (internal/external) that minimises the number of EFMs. The algorithm was applied to the network representing glutathione metabolism, 48 different combinations of compounds achieved the minimum number of EFMs. Among these 48 combinations, four combinations had the minimum number of external metabolites. One of these four combinations was randomly chosen. This method was specifically used to calculate EFMs of the genome-scale network of Lactobacillus Plantarum WCFS1 (Teusink *et al*., 2006), giving insight into the consumption of excess ATP under energy excess.

In the literature the calculation of the set of EFMs has been mostly carried out for the analysis of functional portions of metabolic networks with a limited number of reactions. We briefly summarise below a number of applications of EFMs in different fields, such as metabolic engineering, functional genomics or metabolic diseases. Liao *et al*., 1996, employed EFMs to construct an *E.Coli* strain that efficiently channelled the metabolic flux from carbohydrate to aromatic sugars. Dandekar *et al*., 1999, studied alternative pathways to the classic textbook glycolysis pathway for different organisms. Förster *et al*., 2003, linked EFMs analysis and metabolomics data in order to assign function to orphan genes in a simplified network of Saccharomyces Cerevisiae. Poolman *et al*., 2003, analysed viable pathways in the photosynthate metabolism network of the chloroplast stroma under different conditions. Carlson and Srienc, 2004 sorted EFMs according to their ATP production for a simplified metabolic network of *E.Coli*. Çakir *et al*., 2004, evaluated the importance of five deficient enzymes in the central network of human red blood cells.

Special attention has been given in the literature to the methods and algorithms used in Petri Net theory due to its inherent simplicity and ability to model metabolic networks. Petri Nets are bipartite graphs with two types of nodes: places and transitions, which for metabolic networks correspond to compounds and reactions respectively. The edges connect places (input compounds) with transitions and transitions with places (output compounds), but there are no edges between places or between transitions. The stoichiometric coefficients are used as edge weights. A further object, the token, describes the dynamics of a Petri Net. The number of tokens in a place stands for the number of molecules of that metabolite existing at a given moment. The tokens that exist in the system at a given time describe the state of system, usually called marking. The marking changes when a transition "fires", according to different firing rules. An extensive description of Petri Nets as applied to metabolic networks can be found in Zevendei-Oancea and Schuster, 2003, and Koch *et al.*, 2005. The analysis of T-invariants in Petri Nets has been of particular interest. A T-invariant is a particular set of transitions such that, after firing, the original marking is restored. T-invariants correspond to pathways at PSS in metabolic networks. Analogously, Petri Net theory contains methods to obtain a subset of T-invariants called minimal T-invariants, which correspond precisely to EFMs in metabolic networks. Colom and Silva, 1991, presented an algorithm to compute the set of minimal T-invariants. This algorithm has been used to help construct more powerful algorithms for the computation of EFMs.

2.2.1.2 Extreme pathways

Schilling *et al.*, 2000, proposed a refined view of the set of EFMs named extreme pathways (henceforth EPs). Apart from the PSS and non-decomposability conditions defined above, the set of EPs must satisfy the so-called systemic independence condition, i.e. no extreme pathway can be written as a non-trivial non-negative linear combination of other EPs. So, given a particular set of EFMs, $\{e_1, e_2, ..., e_q\}$, this set is systemically independent if and only if:

$$\text{there does not exist } e_j \subset \{e_1, e_2, ..., e_q\} \text{ for which } e_j = \sum_{i=1, i\neq j}^{q} \alpha_i e_i \text{ where } \alpha_1, \alpha_2, ..., \alpha_q \geq 0$$

In addition, the set of internal compounds is divided into two subsets: currency compounds and primary compounds. Currency compounds are typical cofactors which are involved in energy and redox levels. Primary compounds are the remaining internal compounds in the metabolic network. Whilst primary compounds must satisfy the PSS condition, currency compounds typically do not. Thus, the principal differences between EPs and EFMs are:

- EPs: satisfy the systemic independence condition; EFMs: do not
- EPs: not all internal compounds satisfy PSS; EFMs: all internal compounds satisfy PSS
- the treatment of reversible reactions

Interestingly, EPs have been classified into three classes: Types I, II and III. Type I EPs are those pathways which produce/consume (in net/aggregate terms) external compounds and currency compounds. Type II EPs, also called futile cycles, only produce/consume (in net/aggregate terms) currency compounds. Type III EPs, sometimes called internal cycles, do not produce/consume (in net/aggregate terms) any external or currency compounds. Classic pathways as found in the literature are generally Type I and Type II. Type III pathways lack biological interest, as there is not an overall transformation in the pathway.

In order to clarify the classification of EPs, following the example given in Figure 2.1, suppose C1 and C7 are external compounds; {C2,C5,C8} are primary compounds; and {C3,C4,C6} are currency compounds. We computed the set of EPs for this particular example using the software package Expa (Bell and Palsson, 2005). We obtained seven EPs. Six of these pathways {e_1, e_2, …, e_6} are identical to the six EFMs shown in Figure 2.7, the seventh EP e_7 is shown in Figure 2.8. Considering Figures 2.7 and 2.8 EPs e_1, e_3, e_4, e_5 and e_6 are Type I EPs, as primary and currency compounds are produced and consumed. EP e_2 illustrates a Type II EP in which only currency compounds {C6} are produced. Finally, EP e_7 shows a Type III EP, as no primary or currency compounds are produced or consumed.

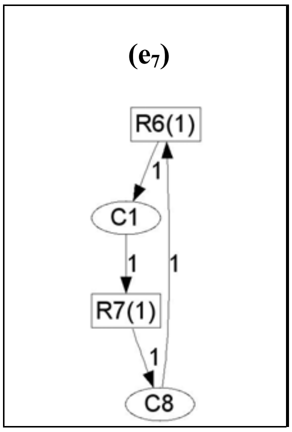

Figure 2.8: The seventh extreme pathway

Note here that EP e_7 appears due to the fact that the EP approach splits reversible reactions into two reactions, as we did with R6 and R7 in the example metabolic network in Figure

2.1. By contrast the EFM approach allows ticks/flux to be negative in reversible reactions. This is the reason as to why EP e_7 does not appear in the set of EFMs.

Aside from EP e_7, the set of EFMs and EPs turn out to be the same for our particular small example. Thus, in this example, the systemic independence condition has no effect. One additional point to note here is that, for the sake of simplicity, we have neglected the network configuration to compute the set of EPs, whose main difference is the addition of exchange fluxes. More as to EFMs and EPs can be found in Palsson *et al.*, 2003, and Papin *et al.*, 2004.

An algorithm to compute the complete set of EPs is outlined in Schilling *et al.* (2000). This algorithm was later implemented in the open-source software called Expa (Bell and Palsson, 2005). Wilback and Palsson, 2002, computed the set of EPs for the genome-scale network of human red blood cell metabolism, which comprises 32 reactions and 39 compounds.

As for EFMs, computing the set of EPs suffers a combinatorial explosion when applied to larger networks. For example Yeung *et al.*, 2007, estimate that there are 3 x 10^{18} EPs in a 904 reaction metabolic network for *E. coli*. For work attempting to limit the effect of this explosion see Schilling *et al.*, 2002, and Yeung *et al.*, 2007.

Whilst EFMs analysis has been applied for different biotechnological or biomedical issues as described above, EPs analysis has usually been focused on elucidating different properties of metabolic networks. Papin *et al.*, 2002, calculated the participation of each reaction in the set of EPs so as to find essential reactions in the metabolic network for the production of individual amino acids in Haemophilus influenzae and for individual amino acids and protein production in Helicobacter pylori. In addition, the length of EPs (as measured by the number of reactions involved in the pathway) was calculated to study, for instance, the minimum number of steps needed to synthesise a given product. Price *et al.*, 2002, evaluated the redundancy of the metabolic network of Helicobacter pylori to meet its biomass objectives under different conditions.

2.2.2 Discussion

Stoichiometric approaches have proposed a novel mathematical definition for metabolic pathways. A metabolic pathway is considered here as a set enzyme catalysed biochemical reactions that satisfy the PSS and non-decomposability conditions.

The most important constraints are those resulting from the PSS condition. However, the criteria to define whether (or not) a compound is internal has not been clearly stated. Indeed, the decision is often modified by authors in the literature depending on the case being studied. It is clear that this decision is of crucial importance due to the fact that the set of EFMs and EPs changes according to the set of internal compounds.

Another issue arising here is that of the PSS condition which requires internal compounds to be stoichiometrically balanced, i.e. they do not appear as by-products or co-substrates in the pathway. This does not fully fit with what is known with respect to experimentally determined pathways. We examined the PSS condition in ten experimentally determined pathways taken from EcoCyc (Karp *et al*, 2002b). We used the *E.Coli* model presented by Reed *et al.*, 2003, so as to define internal/external compounds. Table 2.1 summarises the results, showing that only half of the pathways satisfy the PSS condition. Thus, there will be a high number of pathways which, as they do not satisfy the PSS condition, will never be determined. Indeed, at a more theoretical level, it is not at all clear why individual functional pathways in which internal compounds appear as by-products/co-substrates, thus violating the PSS condition, cannot exist.

Pathway	*Number of balanced internal compounds*	*Number of unbalanced internal compounds*	*Satisfies PSS condition?*
Gluconeogenesis	8	0	Yes
Glycogen	2	0	Yes
Glycolysis	9	0	Yes
Proline biosynthesis	4	0	Yes
Ketogluconate metabolism	2	0	Yes
Pentose phosphate	7	1	No
Salvage pathway	3	1	No
Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	7	2	No
NAD biosynthesis	4	3	No
Arginine biosynthesis	7	1	No
Total	53	8	

Table 2.1: Pathways examined with respect to unbalanced internal compounds

One point to note though from Table 2.1 is that, reflecting the PSS condition for the entire organism, many (but not all) of the internal compounds are balanced in individual pathways, namely 87% (=100×53/(53+8)) of the compounds. It is clear that a significant step towards ensuring that the PSS condition is satisfied for the entire organism (regarded as the entire set of pathways acting together) can be made if individual pathways themselves satisfy the PSS condition with respect to the majority of compounds involved.

Obviously, the removal of the PSS condition might lead to determination of a high number of insignificant metabolic pathways. However, our view is that this condition should (somehow) be relaxed so as to include the whole set of meaningful metabolic pathways whilst preserving the biological logic of pathways. The addition of novel constraints, regulatory (Covert and Palsson, 2003), topological or energetic (Henry *et al*., 2007), might be helpful to avoid meaningless pathways. However it is clear that the algorithms presented to date to compute EFMs and EPs do not show much flexibility in terms of adding new constraints. The algorithm presented by Schilling *et al*., 2000, for example, does not directly remove Type III pathways and posterior analysis is necessary to eliminate them.

With respect to the non-decomposability condition, we believe that the idea of pathways satisfying a simplicity or independence condition is a fruitful one for refining the search for meaningful metabolic pathways. However, to the best of our knowledge, no experimental validation has been carried out so as to show independent operation of a non-decomposable set of enzymes. A further limitation lies in the fact that the non-decomposability condition is quite sensitive with respect to the PSS condition, as explained in Schuster *et al.*, 2002b.

Assuming that the PSS and non-decomposability conditions are appropriate, the main difficulty EFMs and EPs meet is the combinatorial explosion. Even though different attempts have been made to reduce the number of pathways found, the number is still too high for genome-scale networks to carry out a detailed analysis and interpret the pathways obtained, as noted Wiback and Palsson, 2002. This drawback leads us to question the utility of attempting to find the complete set of EFMs or EPs. Type III EPs, for example, illustrate the fact that every extreme pathway does not necessarily have biological significance. The high average values for the length of EPs presented in Papin *et al.*, 2002b, 84 reactions in Helicobacter pylori, 46 reactions in Haemophilus influenzae, which stand in sharp contrast to the typical length of pathways found in the biochemical literature (generally less than 10 reactions), also leads us to question the utility of EPs. Therefore, a better strategy might be to find a small number (perhaps five to ten) of EPs, or EFMs, under a given optimisation criteria. We would make two points here:

- moving from enumerating a large number of possibilities, to generating a small number of possibilities using an optimisation approach, is precisely what has happened with regard to path finding approaches to metabolic pathways (as will become apparent in the discussion as to these approaches given below).

- with regard to the optimisation criteria to be adopted this is an open research question but there are clear parallels in the literature as to what might be valuable:

- maximise (net) *ATP* production (Meléndez-Hevia *et al.*, 1996,1997; Heinrich *et al.*, 1997; Stephani *et al.*, 1998; Stephani and Heinrich, 1998; Ebenhöh and Heinrich, 2001)
- minimise the number of reactions (Meléndez-Hevia and Isidoro, 1985; Meléndez-Hevia and Torres, 1988; Meléndez-Hevia, 1990; Meléndez-Hevia *et al.*, 1994,1996; Mittenthal *et al.*, 1998; Ebenhöh and Heinrich, 2003)
- maximise growth (biomass), as is commonly done in flux balance analysis (Kauffman *et al.*, 2003; Price *et al.*, 2004; Lee *et al.*, 2006)

Note here that to avoid the combinatorial explosion new algorithms will have to be developed to find (without excessive enumeration) a small number of EPs, or EFMs, under a given optimisation criteria. Here there is a parallel with path finding approaches, since algorithms have been developed there that find optimal paths without explicitly enumerating all possible paths.

In summary, our view is that the mathematical and computational concept of a metabolic pathway as proposed by stoichiometric approaches should be re-examined, with emphasis placed on generating a small number of pathways via an optimisation approach.

2.3 Path finding approaches

Path finding approaches emerged as an alternative methodology to analysis metabolic pathways in genome-scale metabolic networks, given the shortcomings presented in the stoichiometric approaches. In contrast to stoichiometric approaches, path finding approaches do not make use of the reaction stoichiometry. Instead, they focus on the fact that there is a (directed) path (containing no cycles) from the source compound to the target compound in experimentally determined metabolic pathways. We refer to this directed path as the metabolic path for a particular experimentally determined metabolic pathway. Of course this path may not be unique, in particular when the pathway is branched. Suppose, for illustration, that the subgraph shown in Figure 2.2a (and reproduced in Figure 2.9 below) is the experimentally determined metabolic pathway that converts C1 into C7. Here, for example, we have the two paths shown in Figure 2.9.

Note, as in Figure 2.9, the difference between a (metabolic) pathway and a metabolic path. The pathway defines all the reactions/compounds involved. The metabolic path is a directed path from the source compound to the target compound in the pathway and may (as in both the metabolic paths seen in Figure 2.9) contain only a subset of the reactions/compounds involved in the pathway.

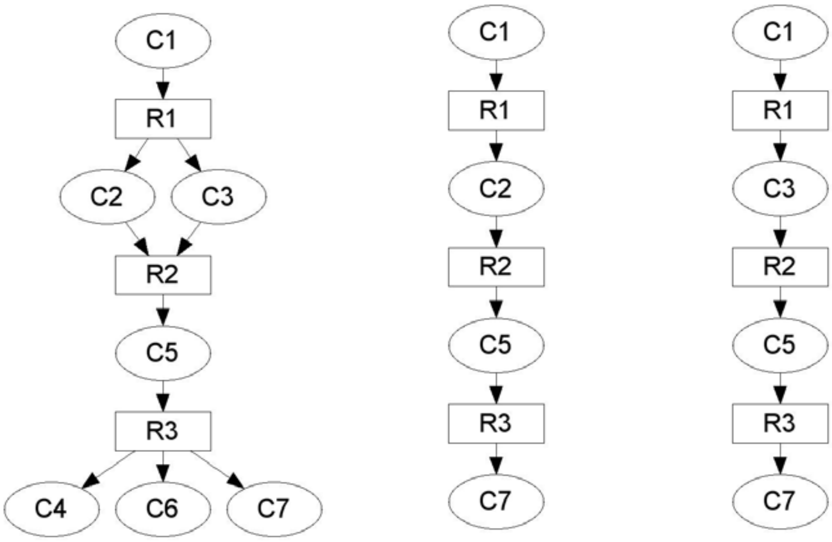

Figure 2.9: Two metabolic paths in the metabolic pathway shown in Figure 2.2

The key assumption behind path finding approaches is that finding directed paths between the source compound and the target compound in the entire metabolic network will give insight into the intermediate reactions/compounds used in the experimentally determined metabolic pathway between the source/target.

In terms of the experimentally determined metabolic pathway we need only focus on the reactions involved in the metabolic path (since for each reaction we know the set of compounds involved). Both of the metabolic paths shown in Figure 2.9 involve reactions R1, R2 and R3. Since these are the only reactions involved in the metabolic pathway then, for this example, knowledge of either of the metabolic paths shown in Figure 2.9 would give us complete insight into the underlying experimentally determined metabolic pathway also shown in Figure 2.9. Note here that with respect to the stoichiometry of the metabolic pathway shown then, once the set of reactions involved in the pathway are known, it is a relatively simple matter (albeit possibly involving some decisions as to which compounds should be balanced with respect to production and consumption) to deduce the stoichiometry of the pathway. With respect to deciding which compounds might be balanced with respect to production and consumption we would note that the intermediate compounds in the metabolic path are themselves often balanced compounds. This can be seen, for example, in Table 2.1 above where 87% of the internal compounds in ten example pathways are balanced.

The work of Küffner *et al.*, 2000, who showed that there were some 500,000 paths from glucose to pyruvate, illustrated that a complete enumeration of paths from the source compound to the target compound made very complicated the analysis of the paths, thus a more sophisticated approach was needed. Accordingly the focus of path finding work moved to defining a suitable distance metric on the directed graph representation of the metabolic network and finding the shortest path from the source node to the target node. Often approaches in the literature move beyond considering just the shortest path to consider the k-shortest paths (for small values of k). For readers unfamiliar with the concept of k-shortest paths k=1 corresponds to the shortest path; k=2 corresponds to the second shortest path; k=3 to the third shortest path; etc.

In the next section we present a literature review of relevant work dealing with path finding in metabolic networks.

2.3.1 Path finding literature review

Küffner *et al.*, 2000, described the database of reactions/compounds as a Petri Net (bipartite graph), where there are two types of nodes, places (compounds) and transitions (reactions). The edges connect places (input compounds) with transitions and transitions with places (output compounds). Paths are formed via a "firing rule". The solution approach adopted was a branch and bound algorithm. They found that there were over 500,000 paths from glucose to pyruvate. Their firing rule reduced the total number of paths to approximately 80,000. This was reduced to 170 paths by imposing further restrictions.

Because of the large number of possible paths identified by Küffner *et al.*, 2000, the subsequent work reported in the literature has focused on enumerating just a small number of paths. We describe these path finding approaches below.

Arita *et al.*, 2000, proposed the use of a k-shortest path algorithm to find paths in metabolic networks, where compounds are represented at the atomic level. They applied their approach to a number of example pathways, where "shortest" is interpreted as minimising the number of arcs (reactions) involved in the path. For Glycolysis (regarded as a path from glucose to pyruvate) they reported that they find some, but not all, of the compounds appearing in that pathway. They noted that one advantage of their approach is the enumeration of multiple (or in the limit, all) paths.

McShan *et al.*, 2003, considered metabolic pathways in terms of a biochemical state-space: compounds define the states and reactions define the state-transitions. The state-space, compounds, are defined as a vector $x = (x_1,x_2,\ldots,x_n)$ of 145 chemical descriptors. The state-transitions, reactions, are considered as transitions between states. Each reaction is simplified to only one input and output compound, avoiding side compounds. The cost of the transitions is defined as the Manhattan distance of the Δx vector, Δx being defined as the difference between the x vectors belonging to the input and output compounds. The problem of finding metabolic paths is viewed as searching for a path from an initial state to a destination state through a series of transitions. An algorithm (A^* search) to minimise the cost of the transitions was applied to find metabolic paths. They reported that they found A^* search to be more efficient than other search techniques they examined such as breadth-first or depth-first search. However, no biological validation was carried out to examine the accuracy of the approach.

Dooms *et al.*, 2005, proposed the use of constraint programming to find constrained paths in metabolic networks. They cite the PhD thesis of Croes (work later reported in Croes *et al.*, 2005, 2006) and, although their wording is imprecise, it does appears that all compound nodes in their approach were assigned a weight proportional to their degree of connectivity (number of reactions in which the compound participates), as in Croes *et al.*, 2005, 2006. One limitation is that in their work they need to know some of the reactions participating in the metabolic path that represents the metabolic pathway. In addition, they note that their approach cannot guarantee to find the optimal constrained shortest path.

In Rahman *et al.*, 2005, by comparison to other approaches, there are no reactions nodes in the metabolic network, only compound nodes. Edges between any two compounds are assigned according to their structural similarity. A breadth-first search algorithm was applied to compute the k-shortest paths between an initial source compound and a final target compound. In a related work, Rahman and Schomburg, 2006, used a k-shortest path approach to identify "load points" and "choke points". They defined a load value for each compound based on the ratio of the number of k-shortest paths passing through it and the number of links associated with the compound. They defined a choke value for each compound based on the number of k-shortest paths passing through it and the load value. Results from their approach were presented for two related bacteria.

Croes *et al.*, 2005, 2006, presented a path finding approach that utilises connectivity. They define connectivity for a compound to be the number of reactions (in the reaction database) in which the compound participates (either as an input compound or as an output compound). They define connectivity for a reaction node to be one. Node connectivities are then taken as the distance metric to be minimised when finding shortest paths. They use a depth-first backtracking (tree search) algorithm to find the k-shortest paths (k=1,2,3,4,5) not between a source compound and a target compound, but between a source reaction and a target reaction**.** Their view of a metabolic pathway as being between a source reaction and a target reaction is not usual in the literature. They systematically (and numerically) compare the k-shortest paths they find with a number of metabolic pathways. Their approach, which appears to be the most effective of all path finding approaches presented to date in the literature, is based on the observation that many of the intermediate compounds in a metabolic path appear to have low connectivity. Evidence presented in Croes *et al.*, 2005, 2006, indicates that biologically meaningful pathways can be found using k-shortest path approaches.

Although outside the scope of the thesis, note here that path finding approaches are broadly used in different biological contexts. In particular, a number of approaches have been proposed to detect biologically significant pathways in protein interaction networks (Hüffner *et al.*, 2007; Scott *et al.*, 2006; Shlomi *et al.*, 2006; Steffen *et al.*, 2002). Essentially these approaches carry out a search for shortest paths where the cardinality of the path (the number of nodes it contains) is specified/constrained. Further reading can be found in (Aittokallio and Schwikowski, 2006; Bebek and Yang, 2007; Kelley *et al.*, 2003).

2.3.2 Discussion

We believe path finding approaches constitute a considerable advance with respect to stoichiometric approaches for several reasons:

- the problem of finding k-shortest paths from a source compound to target compound is a well-known problem in graph theory and computationally tractable for genome-scale metabolic networks;

-the problem as to define compounds which are internal or external as required in EFMs and EPs is avoided and thus any possible pathway can be determined;

- computing k-shortest paths according to a suitable distance metric, instead of computing all the paths, appears to be a quite logical concept as not necessarily all the computed paths will have a biological significance. In addition, the analysis of the computed metabolic paths becomes simpler since we usually restrict k to very small number.

In order to determine biologically significant metabolic paths, the key decision is choice of an appropriate distance metric. The distance metric proposed in Croes *et al.*, 2005, 2006, appears to be the most effective to date presented in the literature.

One major drawback of path finding approaches however is that they are relatively inflexible in terms of adding additional, biologically meaningful, constraints. The biological significance of the paths found is implicitly completely determined by the distance metric adopted. Currently, for example, even deducing stoichiometric information for a metabolic path must be done as a separate stage (e.g. by balancing intermediate compounds in the path), once the path has been computed without using stoichiometric information**.** Being able to add biologically based constraints, e.g. stoichiometric, regulatory (Covert and Palsson, 2003), topological or energetic (Henry *et al.*, 2007), as an intrinsic part of the path-finding process, would significantly refine the search for biologically meaningful metabolic paths; provided this can be done without excessively complicating the algorithmic/computational expense of finding k-shortest paths.

2.4 Conclusions

In this chapter computational/mathematical approaches to metabolic pathways have been described, reviewed and discussed. For the sake of clarity, approaches were divided into two types: stoichiometric approaches and path finding approaches.

Despite the fact that stoichiometric approaches present a theoretical basis (although this is debatable from certain respects), EFMs and extreme pathways find severe difficulties when applied to genome-scale networks from both the computational and analytical point of view. With regard to approaches of this kind our judgement is that the key research challenge is to move from enumeration of all possibilities to generating a small number of possibilities using an optimisation approach, and we suggested a number of possible optimisation criteria based on parallels in the literature.

In contrast to stoichiometric approaches, path finding approaches do enable analysis of genome-scale metabolic networks to be performed. With regard to approaches of this kind we believe that the key research challenges are choice of an appropriate distance metric and addition of biologically based constraints.

In summary, we think that the mathematical concept of metabolic pathways must be re-examined. A first step forward is to understand the underlying mathematical logic of experimentally determined metabolic pathways, which will allow us to determine (unknown) biologically meaningful metabolic pathways. This is precisely the aim of this doctoral thesis, which is fully developed in the following chapters.

Chapter 3

The Beasley-Planes model

In order to determine biologically meaningful metabolic pathways inside the metabolic network, the underlying logic of experimentally determined metabolic pathways must be investigated. In this chapter we present a novel mathematical approach, referred as to the Beasley-Planes (BP) model, so as to computationally recover experimentally determined metabolic pathways inside the metabolic network. The effectiveness of the approach was tested in forty experimentally determined pathways of *E.Coli*.

3.1 Introduction

In Chapter 2 we concluded that, in order to search for meaningful metabolic pathways inside the metabolic network, the underlying logic of experimentally determined pathways must be (somehow) investigated. In this chapter we present our initial mathematical approach so as to recover experimentally determined metabolic pathways inside the metabolic network. As we described in Chapter 1, the problem of recovering metabolic pathways consists of defining a mathematical model that, when solved, provides a solution identical to a known experimentally determined pathway, given the following input data:

- the metabolic network of a particular organism;
- the source and target compound of the experimentally determined pathway;
- and (perhaps) some pathway knowledge, such as the number of molecules of the source or target compound.

In order to illustrate the problem, suppose we have the example metabolic network shown in Figure 3.1, which, as described in Chapter 1, comprises 8 reactions (labelled R1 to R8 respectively) and 8 compounds (labelled C1 to C8 respectively).

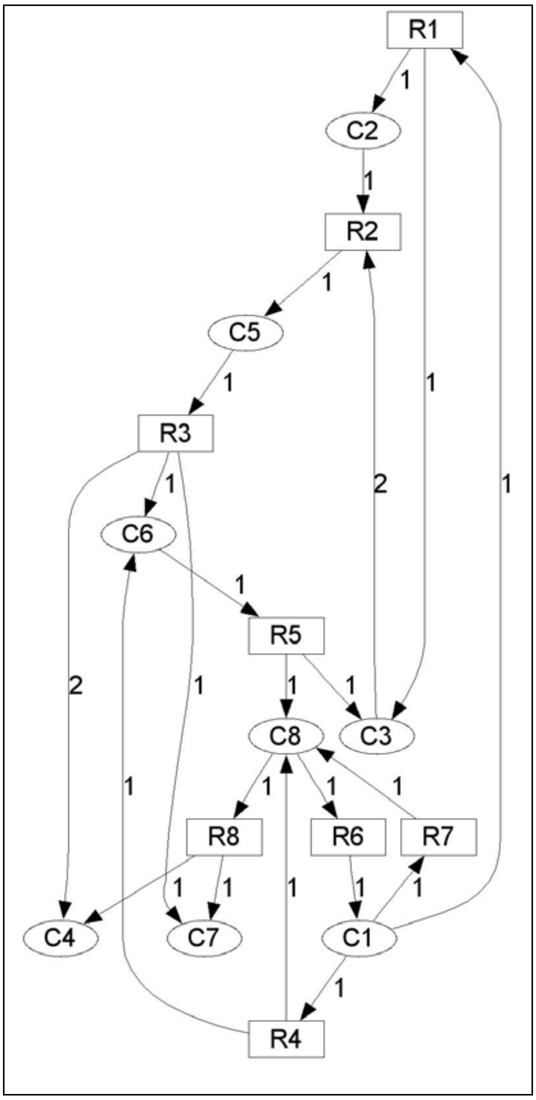

Figure 3.1: Previous Figure 1.2 in Chapter 1

Assume now that we are concerned with recovering the experimentally determined metabolic pathway that converts C1 into C7. Such pathway will be a subgraph of this entire metabolic network that has the property that for each reaction node included in the subgraph all compound nodes associated with that reaction node (either as an input compound or as an output compound) also appear in the subgraph. Clearly, there will be more than one such subgraph that converts C1 into C7 within the metabolic network. Figure 3.2a and Figure 3.2b, for example, show two possible subgraphs converting C1 into C7.

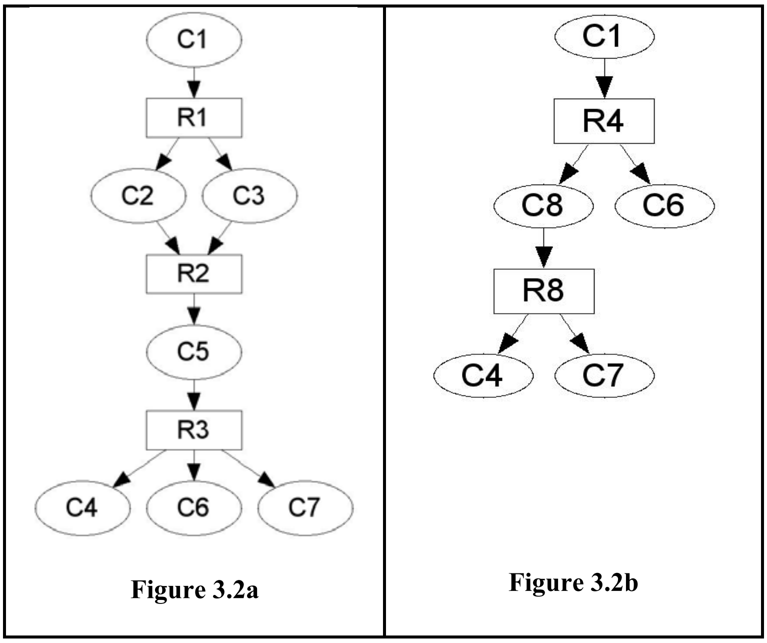

Figure 3.2a

Figure 3.2b

Figure 3.2: Two possible subgraphs for converting C1 into C7

Whilst each subgraph determined inside the metabolic network can essentially be regarded as a feasible pathway from a biochemical viewpoint, only one of them will correspond to the experimentally determined pathway. The problem of determining (recovering) precisely one such subgraph is clearly a highly complex combinatorial problem, as the number of possible subgraphs converting a source compound into a target compound is very high in genome-scale metabolic networks.

One further point to note is that, aside from recovering the precise set of reactions involved in the experimentally determined pathway, we directly address pathway stoichiometry. Indeed, experimentally determined pathways define a unique stoichiometry, i.e. the number of ticks for each reaction involved in the pathway. To illustrate this, suppose that the subgraph shown in Figure 3.2a is the experimentally determined metabolic pathway for converting C1 into C7. Figure 3.3a and Figure 3.3b show two metabolic pathways with the same set of reactions as shown in Figure 3.2a. However, they differ in the pathway stoichiometry, as R1 ticks once in Figure 3.3a and twice in Figure 3.3b. Suppose, for example, that the stoichiometry of experimentally determined pathway for converting C1 into C7 is that appearing in Figure 3.3a. One can easily observe that the pathway shown in Figure 3.3b produces/consumes/balances different compounds in aggregate (net) terms with respect to the pathway shown in Figure 3.3a and thus, the cellular function of the metabolic pathway becomes different. Clearly the fact of determining (recovering) the precise stoichiometry of experimentally determined pathways introduces an additional combinatorial complexity.

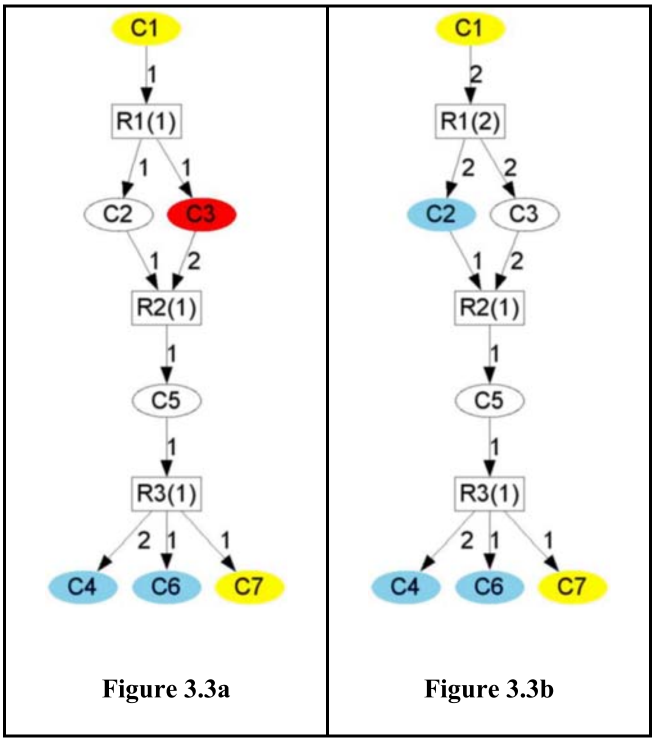

Figure 3.3a **Figure 3.3b**

Figure 3.3: Two possible stoichiometries for the example pathway in Figure 3.2a

We present below our mathematical optimisation model based on integer linear programming so as to recover the precise set of reactions involved in a experimentally determined metabolic pathway, along with its stoichiometry. Henceforth, this model is referred as to the Beasley-Planes (BP) model. Due to the combinatorial complexity of the problem, it may happen that some prior pathway knowledge must be introduced into the model. Clearly the aim is to recover experimentally determined pathways with the minimal use of prior pathway knowledge.

3.2 Mathematical model

3.2.1 Reaction variables and constraints

In the BP model we have a metabolic network of R reactions (where each reaction has a specified direction so a reversible reaction contributes two different reactions to the total number R) which collectively involve C different compounds. Suppose we are seeking a pathway that transforms Q_S molecules of source compound S into Q_T molecules of target compound T. A reaction may, or may not, be active in the pathway. So we have the binary (zero-one) variable:

z_r = 1 if reaction r is active in the pathway, 0 otherwise (r=1,…,R)

and the associated tick variable:

t_r the number of ticks of reaction r in the pathway (this must be an integer variable (≥0) with value 0 if the reaction not active)

We need a constraint relating the number of ticks of a reaction to the zero-one variable signifying whether the reaction is active or not, this is:

$$t_r \leq M_1 z_r \qquad r=1,\ldots,R \tag{3.1}$$

where M_1 is a large positive constant that represents the maximum number of ticks of any reaction (since z_r=1 implies $t_r \leq M_1$). If the reaction does not tick then it must be inactive, so we have the constraint:

$$z_r \leq t_r \qquad r=1,\ldots,R \tag{3.2}$$

3.2.2 Compound variables and constraints

The BP model involves variables relating to whether compounds are balanced (or not). A balanced compound is one where the number of molecules needed (consumed) is equal to the number produced. A compound which is balanced can either be active (number of molecules needed = number produced > 0) or inactive (number of molecules needed = number produced = 0) in the pathway. Considering Figure 3.3a, for example, the active balanced compounds are C2 and C5.

Let n_{cr} be the number of molecules of compound c needed as input for one tick of reaction r and p_{cr} be the number of molecules of compound c produced as output by one tick of reaction r. For each compound c (c=1,…,C) define:

b_c=1 if for compound c the number of molecules needed is equal to the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r = \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If b_c=1 compound c is balanced.

e_c=1 if for compound c the number of molecules needed is less than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r < \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If e_c=1 compound c is produced to excess, since we have “spare” molecules of the compound to be disposed of (in other pathways).

f_c=1 if for compound c the number of molecules needed is greater than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r > \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If f_c=1 compound c must be freely available, since we need “spare” molecules of the compound that have come from other pathways.

Considering Figure 3.3a, for example, compound C4 is produced to excess (denoted by the blue colouring) and compound C3 is freely available (denoted by the red colouring).

We have the constraint:

$$b_c + e_c + f_c = 1 \qquad c=1,...,C \qquad (3.3)$$

In order to link the variables e_c and f_c to the number of molecules of each compound produced we need the constraints.

$$e_c \geq (\sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r)/M_2 \qquad c=1,\ldots,C \qquad (3.4)$$

$$e_c \leq 1 + (\sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r -1)/ M_2 \quad c=1,\ldots,C \qquad (3.5)$$

$$f_c \geq (\sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r)/M_2 \qquad c=1,\ldots,C \tag{3.6}$$

$$f_c \leq 1 + (\sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r - 1)/M_2 \qquad c=1,\ldots,C \tag{3.7}$$

where M_2 is a large positive constant. Equation (3.4) forces the zero-one variable e_c to be one if $\sum_{r=1}^{R} n_{cr}t_r < \sum_{r=1}^{R} p_{cr}t_r$ whilst equation (3.5) forces e_c to be zero if $\sum_{r=1}^{R} n_{cr}t_r \geq \sum_{r=1}^{R} p_{cr}t_r$. Equations (3.6) and (3.7) are as equations (3.4) and (3.5) but with n_{cr} and p_{cr} interchanged.

3.2.3 Metabolic constraints

The above has defined the variables that we need and the constraints that logically (mathematically) must be satisfied given these variables. We now present the metabolic constraints that we included in the BP model.

We need constraints specifying that the required number of molecules of the source compound S (Q_S) and target compound T (Q_T) are involved – these are:

$$\sum_{r=1}^{R} n_{Sr}t_r = Q_S \quad \text{and} \quad \sum_{r=1}^{R} p_{Tr}t_r = Q_T \tag{3.8}$$

If the source compound and target compound are different then we produce none of the source compound and consume none of the target compound, i.e.

$$\sum_{r=1}^{R} p_{Sr}t_r = \sum_{r=1}^{R} n_{Tr}t_r = 0 \qquad \text{if } S \neq T \tag{3.9}$$

We have found it necessary in the BP model to distinguish between compounds that appear in a significant number of different reactions and compounds that appear in just a few reactions. We define the percentage presence (δ_c) of a compound c to be δ_c = 100(number of reactions in which c appears)/R = $100\sum_{r=1}^{R} \min(\max(p_{cr},n_{cr}),1)/R$. Note that δ_c is defined purely with respect to the set of reactions that are considered and hence will vary as that set of reactions changes. Compounds for which $\delta_c \leq \Delta$ (where Δ is an input parameter) we call low presence compounds. Compounds for which $\delta_c > \Delta$ we call high presence compounds. Other authors (Croes *et al.*, 2006; Horne *et al.*, 2004; Jeong *et al.*, 2000; Ma and Zeng, 2003; Wagner and Fell, 2001) have also found it necessary to distinguish compounds that commonly appear from those that appear less often when considering metabolic networks. In the computational results reported later we used Δ=4%. Although this might seem a small value, for our relatively large database (R=880 reactions, involving C=605 compounds) there were only 16 compounds (shown in Table 3.1) that had $\delta_c > \Delta$ and so were considered high presence compounds.

Compound	Percentage presence
Hydrogen ion	43.86
Water	28.98
Adenosine triphosphate	18.98
Adenosine diphosphate	14.89
Phosphate	14.32
Nicotinamide adenine dinucleotide	9.77
Nicotinamide adenine dinucleotide – reduced	9.32
Diphosphate	8.98
Nicotinamide adenine dinucleotide phosphate	7.16
Carbon dioxide	7.05
Nicotinamide adenine dinucleotide phosphate – reduced	6.93
L-Glutamate	5.91
Coenzyme A	5.23
Pyruvate	4.77
Ammonium	4.43
Adenosine monophosphate	4.43

Table 3.1: High presence compounds

The logic behind this distinction is that high presence compounds appear in so many reactions that we can reasonably assume that if the metabolic pathway we are seeking either needs to obtain molecules of a high presence compound (produced by other pathways); or produces molecules of a high presence compound that have to be disposed of (in other pathways); then this can be achieved. High presence compounds can therefore be regarded as being “freely available” or being “produced to excess” if necessary. Another way to view high presence compounds is that they represent the interaction/interface between the pathway we are considering (which is unknown, but is to be found) and all the other pathways that exist (which are unknown, and remain unknown in terms of our mathematical model).

Low presence compounds, by contrast, cannot be reasonably assumed to be so easily obtained from, or disposed of in, other pathways and so must be balanced, i.e. any molecules involved must be internally produced/disposed of in the pathway chosen from S to T. Hence we have the constraint:

$$b_c = 1 \qquad \text{if } \delta_c \leq \Delta;\ c \neq S,T;\ c=1,\ldots,C \tag{3.10}$$

which forces low presence compounds (excluding S and T) to be balanced. This constraint does not force compounds to be active in the pathway, merely to be balanced.

Equation (3.10) is precisely the pseudo steady state (PSS) condition described in Chapter 2. This links our approach to stoichiometric approaches (Schilling *et al.*, 2000; Schuster *et al.*, 2000. However, the BP model applies the PSS condition to a different set of biochemical compounds, namely low presence compounds. As opposed to the set of internal compounds in elementary flux modes approach, our definition of low presence compounds is done based on a numerical criterion.

On the other hand, Table 2.1 in Chapter 2 shows that the PSS condition, though satisfied by the majority of biochemical compound in the metabolic pathway, is not general. This issue is also found in the BP model. Indeed, we may find experimentally determined pathways containing low presence unbalanced compounds. In the Results section we show how to deal with this issue (by neglecting Equation (3.10) for certain low presence compounds).

In the BP model each reaction active in the pathway has at least one active balanced compound as an output, except any reaction producing the target compound T. Should a reaction be in the pathway and not satisfy this condition it can only be producing high presence compounds, which by definition are freely available anyway. Hence we impose the constraint:

$$\sum_{c=1,\, p_{cr}\geq 1}^{C} b_c \geq z_r \qquad p_{Tr}=0;\ r=1,\ldots,R \tag{3.11}$$

We need to consider the issues of cycles (a closed path in the directed graph representation, e.g. C6-R5-C3-R2-C5-R3-C6 in Figure 3.4) in the pathway. Cycles do exist in metabolic pathways, but in our approach some types of cycles are allowed, others are disallowed.

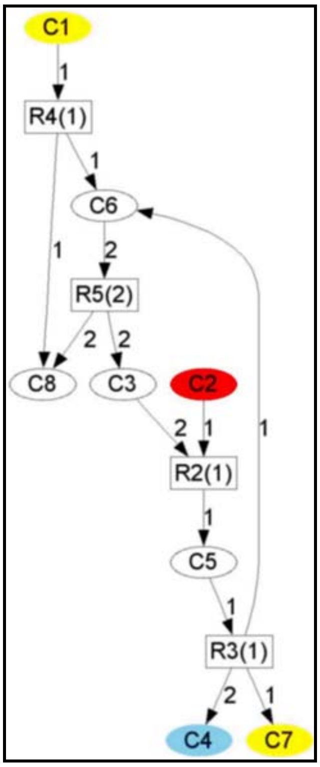

Figure 3.4: A possible pathway that contains a cycle.

Each reaction in our database of R reactions has a specified direction associated with it. Define the set B={(α,β)| reaction α and reaction β are the reverse of each other, α<β}. In order to disallow a cycle around a reaction and its reverse we impose the constraint:

$$z_\alpha + z_\beta \leq 1 \qquad \forall(\alpha,\beta)\in B \tag{3.12}$$

Considering a pathway as a directed graph we define a c-cycle to be an alternating sequence of c active balanced compounds and c active reactions that starts and ends at the same compound and within which no compound/reaction is repeated except at the start/end of the sequence. An example 3-cycle (C6-R5-C3-R2-C5-R3-C6) can be seen in Figure 3.4. In the BP model we regard a c-cycle in a metabolic pathway as allowable if and only if:

- the source compound and the target compound are the same (S=T) and the c-cycle involves that compound; or
- the c-cycle involves exactly one high presence balanced compound

If the above conditions are not met then the c-cycle is disallowed.

The first of these conditions is a logical one. If S=T then the pathway must be a cycle by definition and so must be allowed. The second of these conditions is based on examination of known pathways. In a random sample of 25 pathways (taken from http://biocyc.org/ECOLI/, but excluding the forty pathways dealt with here) we found 5 pathways where there was an allowable c-cycle, but only one pathway where there was a disallowed c-cycle.

To illustrate this second condition if and only if exactly one of the three balanced compounds (C3, C5 and C6) in the 3-cycle (C6-R5-C3-R2-C5-R3-C6) in Figure 3.4 is a high presence compound would the 3-cycle be allowed, otherwise it would be disallowed.

If a c-cycle is disallowed a constraint must be imposed to prevent it appearing in the pathway. To illustrate this consider the case c=3. A 3-cycle involves 3 balanced compounds and 3 reactions. Any 3 reactions α,β,λ for which there exist 3 compounds d,e,h for which: d is an input for α ($n_{d\alpha}>0$); e is an output from α ($p_{e\alpha}>0$) and e is an input for β ($n_{e\beta}>0$); h is an output from β ($p_{h\beta}>0$) and h is an input for λ ($n_{h\lambda}>0$); and d is an output from λ ($p_{d\lambda}>0$); gives rise to a potential 3-cycle (d-α-e-β-h-λ-d). If this 3-cycle is disallowed (it does not satisfy the conditions given above) then the constraint $b_d + z_\alpha + b_e + z_\beta + b_h + z_\lambda \leq 5$ prevents it from appearing. In general the constraint required to prevent a c-cycle from appearing is: *sum of the b variables for the compounds in the c-cycle plus sum of the z variables for the reactions in the c-cycle ≤ 2c-1.*

3.2.4 Objective

Above we have set out a series of variables and constraints (a mathematical model) that we believe can be used to recover a metabolic pathway. It is likely that there is more than one feasible solution to the above mathematical model and so to arrive at a pathway we propose an objective that is to be optimised. Our computational results (reported below) indicate that two factors are of importance in terms of an optimisation objective: the total number of reactions involved in the pathway and the number of excess molecules of Adenosine Triphosphate (ATP).

The total number of reactions involved in the pathway ($\sum_{r=1}^{R} z_r$) should be minimised. This makes biological and evolutionary sense as minimising the number of reactions involved reduces the "complexity" of the pathway. Broadly speaking we would expect that the fewer the reactions involved in a pathway the fewer the enzymes that will be needed by an organism to catalyse the reactions in the pathway. Moreover we would expect that the more reactions involved in a pathway the greater the chance that it may be disrupted, for example should an enzyme not be present due to a genetic defect. Other authors (Meléndez-Hevia and Isidoro, 1985; Meléndez-Hevia and Torres, 1988; Meléndez-Hevia, 1990; Meléndez-Hevia *et al.*, 1994, 1996; Mittenthal *et al.*, 1998; Ebenhöh and Heinrich, 2003) have also emphasised minimisation of the number of reactions involved in a metabolic pathway.

Denoting ATP as compound 1 for simplicity the number of excess molecules of ATP ($\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r$) should be maximised. This makes biological and evolutionary sense as ATP is a key metabolic compound. ATP has been termed the cell's energy currency and is the universal carrier of chemical energy in the cells of all living organisms from bacteria and fungi to plants and animals including humans. It captures the chemical energy released by the combustion of nutrients and transfers it to reactions that require energy. Previous work examining optimality criteria associated with the structure of metabolic pathways

(Meléndez-Hevia *et al.*, 1996, 1997; Heinrich *et al.*, 1997; Stephani and Heinrich, 1998; Stephani *et al.*, 1999; Heinrich and Ebenhöh, 2001) has also focused on the optimisation of (net) ATP production.

Maximising excess ATP:

- if ($\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r$) > 0 produces as many "spare" molecules of ATP as possible for use in other pathways
- if ($\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r$) < 0 uses as few "spare" molecules of ATP (generated in other pathways) as possible in the pathway from S to T.

Attempting to minimise one factor (total number of reactions) whilst simultaneously maximising another (excess ATP) involves a tradeoff. Whilst this tradeoff can be treated in a number of ways (e.g. see Heinrich *et al.*, 1991) in this thesis we examine the two extreme cases of this tradeoff:

$$\text{minimise } M_3(\sum_{r=1}^{R} z_r) - (\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r) \tag{3.13}$$

$$\text{maximise } M_3(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r) - (\sum_{r=1}^{R} z_r) \tag{3.14}$$

where M_3 is a large positive constant. Objective (3.13) gives primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP, whilst objective (3.14) gives primary weight to maximising excess ATP and secondary weight to minimising the total number of reactions. Note here that non-extreme cases of this tradeoff, i.e. $\lambda_1(\sum_{r=1}^{R} z_r) - \lambda_2(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r)$, $\lambda_1>0, \lambda_2>0$, were not explored computationally.

3.2.5 Overview

The BP model (optimise (3.13) or (3.14) subject to (3.1)-(3.12) plus c-cycle constraints) is an integer linear program. Algorithmically such programs are solved by linear programming based tree search. Modern software packages to perform this task, such as ILOG CPLEX, 2005, which we used, are well developed and highly sophisticated.

One computational point here deals with our treatment of c-cycles. We imposed constraints to prevent all disallowed 2-cycles directly and solved the integer program as given above. The solution obtained was then checked to see whether it contained any disallowed c-cycles (for any c>2). Finding a cycle in the directed graph composed of balanced compounds and active reactions is (algorithmically) an easy task, and checking to see whether a c-cycle is allowed or not is trivial. If any disallowed c-cycles were found then constraints to eliminate them (as discussed above) were added and the process repeated until a solution without any disallowed c-cycles was found.

3.3 Results

We have used the metabolic network of *E.Coli* (the best studied organism in the biological world) presented by Reed *et al.*, 2003, which is available from http://systemsbiology.ucsd.edu/In_Silico_Organisms/E_coli/E_coli_reactions and comprises 880 cytosolic reactions and 613 compounds. A cytosolic reaction is one occurring in the cytosol, which essentially defines the medium where metabolism is carried out. A full list of reactions/compounds can be found in Appendices A and B.

We applied the BP model to the forty *E.Coli* experimentally determined pathways shown in Table 3.2. The pathways used were taken from Keseler *et al.*, 2005; Nelson and Cox, 2005 and http://biocyc.org/ECOLI/. A detailed description of the experimentally determined pathways examined can be found in Appendix C.

Pathway Number	Pathway name	Low presence unbalanced compounds
1	Gluconeogenesis	-
2	Glycogen	-
3	Glycolysis	-
4	Proline biosynthesis	-
5	Ketogluconate metabolism	-
6	Pentose phosphate	g3p
7	Salvage pathway deoxythymidine phosphate	ura, thym
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	fad, fadh2, accoa
9	NAD biosynthesis	dhap, o2, h2o2,prpp
10	Arginine biosynthesis	akg, asp-L,fum.,ac, accoa,cbp
11	Sperdimine biosynthesis	urea, 5mta, ametam
12	Threonine Degradadation to synthetise propionate	for
13	Serine biosynthesis	akg
14	Histidine biosynthesis	akg, aicar, gln-L
15	Tirosine biosynthesis	akg
16	Coenzyme A biosynthesis	ctp, cmp, cyst-L
17	Pantothenate biosynthesis	akg, thf, mlthf,ala-B
18	Tetrahydrofolate biosynthesis	chor, gln-L,for, gcald
19	Riboflavin and FMN and FAD biosynthesis	db4p, for
20	Heme Biosynthesis	o2, fe2,frdp
21	De novo sinthesis of pyrimidine ribonucletides	asp-L, gln-L, q8, q8h2,prpp
22	De novo sinthesis of pyrimidine deoxyribonucletides	dhf, mlthf,trdox, trdrd
23	Phenylethylamine degradation	o2, h2o2
24	Rhamnose degradation	dhap
25	Fucose degradation	dhap
26	Entner-Doudoroff Pathway	-
27	Anaerobic Respiration	oaa, accoa, h2
28	Arginine degradation	akg, succoa
29	Proline degradation	fad,fadh2
30	Glycolate degradation	accoa
31	Phospholipid Biosynthesis	q8, q8h2
32	Biosynthesis of cysteine	-
33	Allantoin degradation	-
34	Deoxycytidine degradation	urea
35	Phenylalanine Biosynthesis	acald, ura
36	Glyoxylate Cycle	dhap
37	Propionate Degradation	suc, accoa
38	Glutamate Biosynthesis Cycle	oaa
39	Biotin Synthesis	akg
40	Glycerol Degradation	amet, cys-L,amob

Table 3.2: Metabolic pathways considered and low presence unbalanced compounds

The fundamental limitation of the BP model arises with pathways 6-25, 27-31 and 34-40 in that they contain some low presence unbalanced compounds, as shown in Table 3.2. Due to equation (3.10), the BP model forces low presence compounds to be balanced. Hence for these pathways we did not force these compounds to be balanced (i.e. we excluded them from equation (3.10)). In other words the BP model, for these pathways, requires pathway knowledge with respect to low presence unbalanced compounds incorporated into the mathematical model.

To illustrate this limitation, Figure 3.5 shows the Anaerobic Respiration pathway, which convert one molecule of *Pyruvate* (pyr) into one molecule of *2-Oxoglutarate* (akg). The number in brackets after the compound label is the percentage presence, δ_c. For example, the percentage presence of *Oxaloacetate* (oaa), δ_{oaa}, is 1.4%. As $\delta_{oaa} \leq \Delta$, with Δ = 4%, then oaa is a low presence compound. According to equation (3.10), low presence compounds must be in aggregate (net) terms balanced. However, the Anaerobic Respiration pathway presents a case in which oaa is, in aggregate (net) terms, unbalanced. Indeed, one molecule of oaa is consumed in R272, as can be seen in Figure 3.5. In order to recover this pathway, the BP model needs, when applied for this particular pathway, to exclude the low presence balancing constraint, equation (3.10), for this particular compound. This also applies to any other low presence unbalanced compounds in the pathway. In the Anaerobic Respiration pathway we have only one low presence unbalanced compound, oaa. Accordingly equation (3.10) was removed for oaa compound when the BP model was applied to the Anaerobic Respiration pathway. This procedure was repeated for each experimentally determined pathway. Table 3.2 shows the low presence unbalanced compounds for each pathway. For example, equation (3.10) was removed for ura and thym compounds when the BP model was applied to Salvage pathway deoxythymidine phosphate (Pathway 7).

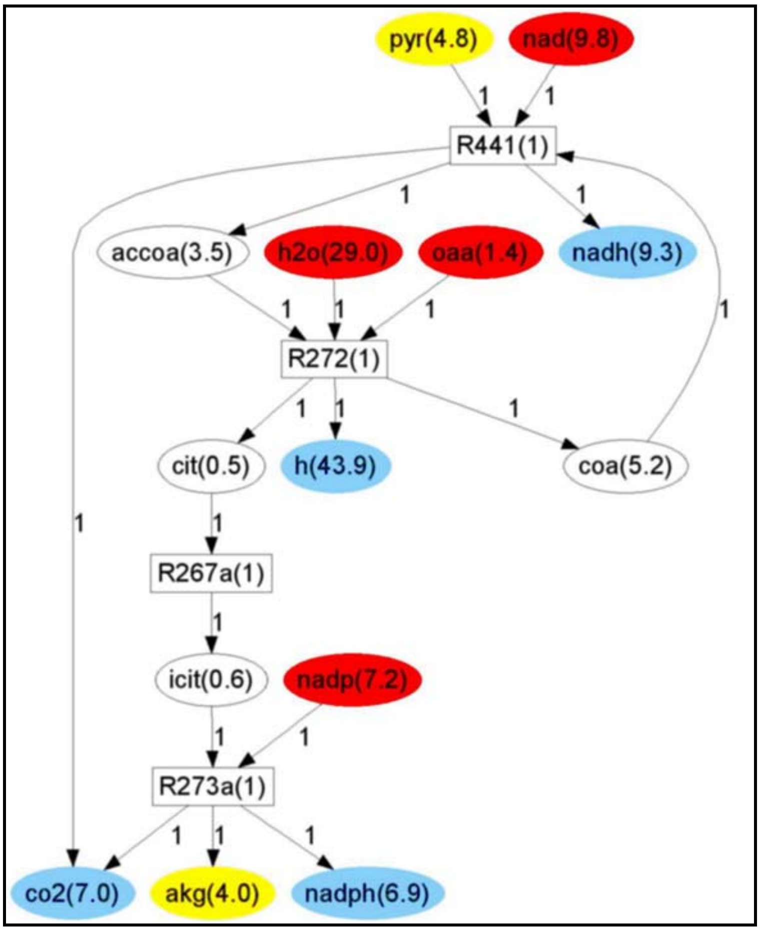

Figure 3.5: Anaerobic respiration pathway

Thus, the BP model needs to know beforehand the low presence unbalanced compounds for each experimentally determined pathway. Clearly this is a limitation since most experimentally determined pathways contain low presence unbalanced compounds. The Improved Beasley-Planes (IBP) model, described in Chapter 6, directly addresses this issue.

Despite this limitation however we believe that the BP model is important because, as our results below indicate, given this knowledge we can recover a large number of the experimentally determined pathways when we apply the BP model. **This indicates that we have reason to believe that disparate observed experimentally determined pathways have a common underlying mathematical model.**

3.3.1 Structural recovery of experimentally determined pathways

We mean here by structural recovery that, once the BP model is solved, the solution is precisely the same as the experimentally determined metabolic pathway, both in terms of the reactions/compounds involved in the pathway and its inherent stoichiometry (reaction ticks). Note that the BP model needs Q_S, Q_T the number of source and target molecules to be specified.

Table 3.3 indicates that for 38 of our 40 experimentally determined pathways one (or both) of our objectives do recover the structure of the pathway. Statistically this is a highly significant result (significant at the 0.001% level), as we will show in a later subsection.

With respect to computation time the average computation time over the eighty cases shown in Table 3.3 was 4.6 seconds, no case requiring more than 30 seconds (1.86Ghz pc, 2GB RAM). For thirty-eight of the forty pathways in Table 3.3 optimising using objective (3.14) took longer than optimising using objective (3.13), on average five times longer.

In forty-eight of the fifty-one “yes” cases in Table 3.3 there is a unique pathway providing the optimal objective function value and in only three cases is there an alternative pathway providing the same optimal objective function value.

As we have a significant number of constraints in the BP model the question arises as to the relevance of the objective adopted. In the limit for example there may be only one unique solution satisfying the constraints, and if so the objective adopted becomes irrelevant. We have investigated this issue and have found that in all 51 “yes” cases in Table 3.3 we have more than one solution satisfying the constraints.

Pathway number	Pathway name	Pathway recovered?	
		Objective 3.13	Objective 3.14
1	Gluconeogenesis	yes	no
2	Glycogen	yes	no
3	Glycolysis	yes	yes
4	Proline biosynthesis	yes	no
5	Ketogluconate metabolism	yes	no
6	Pentose phosphate	yes	no
7	Salvage pathway deoxythymidine phosphate	yes	no
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	no	yes
9	NAD biosynthesis	yes	no
10	Arginine biosynthesis	yes	no
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	yes
14	Histidine biosynthesis	yes	no
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	no
17	Pantothenate biosynthesis	yes	no
18	Tetrahydrofolate biosynthesis	yes	no
19	Riboflavin and FMN and FAD biosynthesis	no	no
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	no
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	no
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	no
25	Fucose degradation	yes	no
26	Entner-Doudoroff Pathway	yes	yes
27	Anaerobic Respiration	yes	yes
28	Arginine degradation	yes	no
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	no
31	Phospholipid Biosynthesis	yes	no
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	no
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	yes	no
37	Propionate Degradation	yes	no
38	Glutamate Biosynthesis Cycle	yes	no
39	Biotin Synthesis	no	no
40	Glycerol Degradation	yes	no
Number of "yes" entries		37	14

Table 3.3: Structural Recovery

Equation (3.9) explicitly excludes solutions in which reactions in the pathway produce any of the source compound (or consume any of the target compound). If we amend the BP model (which is trivially done) to allow such solutions then, with respect to Table 3.3, we degrade the results slightly, failing to recover pathways 1, 6, 9, 27 and 37 for objective (3.13) and pathways 26, 27 and 34 for objective (3.14). In a random sample of 25 pathways (taken from http://biocyc.org/ECOLI/, but excluding the forty pathways dealt with here) we found only one pathway in which equation (3.9) was violated (and that was for a pathway where the source compound was itself a high presence compound).

If we do not impose the constraint on allowable c-cycles then, with respect to Table 3.3, we have a mix of situations. Objective (3.13) now fails to recover pathway 3 and 9. However, objective (3.13) now recovers pathway 19 and 39, which were not recovered by the BP model (see Table 3.3) as they contains at least one disallowed cycle. Thus, objective (3.13) preserves 37 recovered pathways. With respect to objective (3.14), we degrade the results significantly, as no pathway is now recovered.

As our approach is linked to elementary flux modes (and extreme pathways) the question arises as to the results we would obtain were we to apply an elementary flux modes based approach. Here such an approach would be to optimise (3.13) or (3.14) subject to (3.1)-(3.10), i.e. excluding equations (3.11), (3.12) and c-cycle constraints. If we do this then, with respect to Table 3.3, we degrade the results significantly. Objective (3.13) now only recovers thirty-three pathways (including pathway 19 and pathway 39) and objective (3.14) fails to recover any pathway. Note here that we neglect here the non-decomposability condition, as we do not have a linear mathematical statement of such condition.

3.3.2 Q_S,Q_T recovery of experimentally determined pathways

In the BP model it is necessary to specify the number of molecules of the source and target compounds (Q_S,Q_T) involved in the pathway (equation (3.8)). For the results shown in Table 3.3 these values have (obviously) been taken as equal to those associated with the experimentally determined pathway. For example, the Anaerobic Respiration pathway shown in Figure 3.7 has (Q_S,Q_T)=(1,1). In this section, we present results as to the BP model when applied to a number of different (Q_S,Q_T) pairs (Q_S,Q_T≤6), so that the dominant pair is determined in terms of the objective function. In the case that the dominant pair is precisely that appearing in the experimentally determined pathway, then the BP model does recover the (Q_S,Q_T) pair observed in the experimentally determined pathway. Such analysis was exclusively carried out in those pathways in which the BP model achieves structural recovery, i.e. those pathways having a "yes" in Table 3.3. As the BP model present two different objective functions, (3.13) and (3.14), the criterion for selecting the dominant pair was modified according to objective function optimised. To illustrate this, we show below results obtained for the Anaerobic respiration pathway (pathway 27), whose structure was recovered for objective function (3.13) and (3.14) as shown in Table 3.3 and thus, (Q_S,Q_T) analysis must be carried out for both objectives.

The Anaerobic respiration pathway in Table 3.3 has (Q_S,Q_T)=(1,1), requiring 4 reactions and consuming no molecules of ATP. For this pathway Table 3.4 shows for a number of different (Q_S,Q_T) pairs (Q_S,Q_T≤6) the number of reactions and the excess ATP when the BP model is solved using objective (3.13). Situations where the BP model indicated that no feasible solution exists are indicated by a 'X'. In other words in these cases no values for the decision variables in the BP model exist which satisfy all the constraints of that model for the particular (Q_S,Q_T) pair examined. Objective (3.13) gives primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP. In order to identify the dominant (Q_S,Q_T) pair with respect to this objective we examine all entries in the table. Let E represent the set of all feasible (Q_S,Q_T) pairs in the table. We apply the following procedure:

- eliminate repeats from E. An entry is a repeat if it involves the same number of reactions but precisely k (≥2, integer) times as many source/target/excess ATP molecules. A repeat essentially corresponds to the same reaction set but with the ticks multiplied by a factor of k. In Table 3.4, for example, the entries for

(Q_S,Q_T)=(4,2) and (Q_S,Q_T)=(6,3) are a repeat of the entry for (Q_S,Q_T)=(2,1) with k=2 and k=3 respectively. In addition, the pairs seen down the diagonal are all repeats of the entry for (Q_S,Q_T)=(1,1). After elimination of repeats the entries left are (Q_S,Q_T)=(1,1); (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (2,3); (2,4); (2,5); (2,6); (3,4); (3,5); (3,6); (4,3); (4,5); (4,6); (5,6); (6,4); (6,5).

- eliminate from E any entries that do not involve the minimum number of reactions. In Table 3.4 the minimum number of reactions is 4. Thus, all the remaining entries aside from (Q_S,Q_T)=(1,1) are eliminated from E.

- choose from E the entry which involves the maximum excess ATP, ties broken by minimum number of source molecules used, and then further broken if necessary by maximum number of target molecules produced. As there is only one entry with the minimum number of reactions, it is not necessary to tie breaks here.

For this pathway the BP model indicates that the pair (Q_S,Q_T)=(1,1) dominates all other cases. This is indicated by the * superscript on that entry in the Table 3.4. Hence in this case the BP model recovers the (Q_S,Q_T)=(1,1) pair observed in the experimentally determined pathway.

(number of reactions, excess ATP)		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number of molecules Q_S of source compound	1	(4,0)*	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)
	2	(6,-1)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)
	3	X	X	(4,0)	(5,0)	(5,0)	(5,0)
	4	X	(6,-2)	(7,-2)	(4,0)	(5,0)	(5,0)
	5	X	X	X	X	(4,0)	(5,0)
	6	X	X	(6,-3)	(7,-3)	(7,-3)	(4,0)

Table 3.4: **BP model solution for objective (3.13), expressed as (number of reactions, excess ATP), for varying Q_S and Q_T for Anaerobic respiration pathway**

Amending that procedure to identify the dominant (Q_S,Q_T) pair for objective (3.14) is easily done. The results for the Anaerobic Respiration pathway with objective (3.14) can be seen in Table 3.5. As objective (14) gives primary weight to maximising excess ATP and secondary weight to minimising the total number of reactions, the procedure for identifying the dominant pair with objective (3.14) is:

- eliminate repeats from E. In Table 3.5 the entries for (Q_S,Q_T)=(4,2) and (Q_S,Q_T)=(6,3) are a repeat of the entry for (Q_S,Q_T)=(2,1) with k=2 and k=3 respectively. In addition, the pairs seen down the diagonal are all repeats of the entry for (Q_S,Q_T)=(1,1). After elimination of repeats the entries left are (Q_S,Q_T)=(1,1); (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (2,3); (2,4); (2,5); (2,6); (3,4); (3,5); (3,6); (4,3); (4,5); (4,6); (5,6); (6,4); (6,5).

- eliminate from E any entries that do not involve the maximum excess ATP. Since excess ATP for all the remaining entries in Table 3.5 is zero, no entry is eliminated in this step.

- choose from E the entry which involves the minimum number of reactions, ties broken by minimum number of source molecules used, and then further broken if necessary by maximum number of target molecules produced. In the example Table 3.5, the entry for (Q_S,Q_T)=(1,1) presents the minimal number of reactions.

For this pathway the BP model indicates that the pair (Q_S,Q_T)=(1,1) dominates all other cases. Hence in this case the BP model recover the (Q_S,Q_T)=(1,1) pair observed in the experimentally determined pathway.

(number of reactions, excess ATP)		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number of molecules Q_S of source compound	1	(4,0)*	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)
	2	(7,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)
	3	X	X	(4,0)	(5,0)	(5,0)	(5,0)
	4	X	(7,0)	(8,0)	(4,0)	(5,0)	(5,0)
	5	X	X	X	X	(4,0)	(5,0)
	6	X	X	(7,0)	(8,0)	(8,0)	(4,0)

Table 3.5: BP model solution for objective (3.14), expressed as (number of reactions, excess ATP), for varying Q_S and Q_T for Anaerobic respiration pathway

We have repeated the analysis shown in Table 3.4 and Table 3.5 for those cases in which the BP model recovers the pathway structure (see Appendix C for details of this analysis). The summary of this analysis can be seen in Table 3.6. Our judgment in that for thirty-seven of the forty pathways the BP model (either objective (3.13) or (3.14)) recovers the (Q_S,Q_T) pair observed in the experimentally determined pathway. Statistically this is a highly significant result (significant at the 0.001% level), as shown in the following subsection.

Pathway number	Pathway name	(Q_S,Q_T) recovered?	
		Objective 3.13	Objective 3.14
1	Gluconeogenesis	yes	-
2	Glycogen	yes	-
3	Glycolysis	yes	yes
4	Proline biosynthesis	yes	-
5	Ketogluconate metabolism	yes	-
6	Pentose phosphate	no	-
7	Salvage pathway deoxythymidine phosphate	yes	-
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	-	yes
9	NAD biosynthesis	yes	-
10	Arginine biosynthesis	yes	-
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	no
14	Histidine biosynthesis	yes	-
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	-
17	Pantothenate biosynthesis	yes	-
18	Tetrahydrofolate biosynthesis	yes	-
19	Riboflavin and FMN and FAD biosynthesis	-	-
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	-
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	-
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	-
25	Fucose degradation	yes	-
26	Entner-Doudoroff Pathway	yes	no
27	Anaerobic Respiration	yes	yes
28	Arginine degradation	yes	-
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	-
31	Phospholipid Biosynthesis	yes	-
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	-
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	yes	-
37	Propionate Degradation	yes	-
38	Glutamate Biosynthesis Cycle	yes	-
39	Biotin Synthesis	-	-
40	Glycerol Degradation	yes	-
Number of “yes” entries		36	12

Table 3.6: (Q_S,Q_T) Recovery

3.3.3 Statistical significance

We address here the issue of the statistical significance of the results we have obtained. We will deal with:

- recovering the pathway structure (Table 3.3); and
- recovering the (Q_S,Q_T) pair associated with the pathway (Table 3.6)

separately.

3.3.3.1 Structural recovery

Table 3.3 shows that out of 80 cases we have a "yes" entry in 51 cases (indicating that we recovered the experimentally determined pathway), and a "no" entry in 29 cases (indicating that we failed to recover the experimentally determined pathway). With regard to the issue of statistical significance then, as an analogy, if we had conducted 80 throws of a coin and observed 51 heads and 29 tails then most likely we would have nothing of particular significance if we were testing whether the coin was fair or not. But for a fair coin the probability of a head is known to be ½.

In order to conduct a hypothesis test to judge the statistical significance of our results we need to know the probability (say ρ) that the BP model, when solved, will (by chance) recover a known pathway; i.e. we would like to conduct the hypothesis test:

H_0: results from the BP model arise due to chance
probability of a success ("yes" entry in Table 3.3) = ρ

versus

H_1: results from the BP model do not arise from chance
probability of a success > ρ

In order to carry through this hypothesis test on our results we need to know ρ.

To proceed let us ignore the issue of reactions ticks (pathway stoichiometry) for convenience. For our database each compound is involved in (on average) 4.5 reactions. As an approximation therefore (for a reasonable value of K) there are of the order of 4.5^K different pathways from the source compound S to the target compound T that involve exactly K reactions. This is an approximation since a reaction may produce more than one compound (each of which in turn may be involved in 3.5 other reactions).

For the forty pathways we have examined the number of reactions involved in the experimentally determined pathway varies from a minimum of 2 to a maximum of 10, with an average of 4.825. Adopting an average value of 4 as a deliberate under-estimate so as to not bias any calculations in our favour an estimate of the number of possible pathways involving exactly 4 reactions is 4.5^4, which is approximately 410.

Only one of these 410 possible pathways corresponds to the experimentally determined pathway. If the BP model were (for example) simply making a random choice from this set of possible pathways then it is clear that the probability of achieving a "yes" entry in Table 3.3 is very low. We however achieve 51 "yes" entries.

Of course this value of 410 is purely an estimate from average values and so adopting a value for ρ of 1/410 = 0.0024 for hypothesis testing may be misleading. Here we shall make a very conservative assumption and assume that 1/410 overestimates the true value of ρ by two orders of magnitude, i.e. we shall use a value of ρ of 1/4.1 = 0.24 for hypothesis testing. Hence we have the hypothesis test:

H_0: results from the BP model arise due to chance

probability of a success = ρ = 0.24

versus

H_1: results from the BP model do not arise from chance probability of a success > ρ

The BP model recovers the experimentally determined pathway in 51 cases out of 80 in Table 2, so a sample probability of 51/80 = 0.64 of success. The test statistic for this one-sided hypothesis test is (sample probability – ρ)/√[ρ(1-ρ)/(sample size)] = (0.64-ρ)/√[ρ(1-ρ)/80] = (0.64-0.24)/√[(0.24)(1-0.24)/80] = 8.32. This is a statistically highly significant result. At the 0.001% level for example the critical value is 4.27, and our test statistic far exceeds this, so H_0 would be rejected and we would conclude that the results from the BP model do not arise from chance. This fact shows that the results obtained for the BP model for recovering the pathway structure are statistically significant at the 0.001% level, as noted above.

3.3.3.2 Q_S,Q_T recovery

Table 3.6 shows that the BP model can recover the specific values of (Q_S,Q_T) associated with the experimentally determined pathway (where Q_S is the number of molecules of the source compound S consumed and Q_T is the number of molecules of the target compound T produced). This analysis was performed by solving the BP model for all different (Q_S,Q_T) pairs (Q_S,$Q_T \leq 6$) and determining the (Q_S,Q_T) pair that was dominant.

At first sight, if we examine all (Q_S,Q_T) pairs, where Q_S,$Q_T \leq 6$ so 36 different pairs in total, we might believe that there is a probability of 1/36 of (by chance) choosing the single correct pair associated with the experimentally determined pathway. However this ignores the issue of repeats, which might (in some cases) have the same number of reactions and the same excess ATP. Referring back to the (Q_S,Q_T) discussion carried out in Section 3.3.2, an entry is a repeat if it involves the same number of reactions but precisely k (≥2, integer) times as many source/target/excess ATP molecules. Clearly, a potential repeat will involve, at least, k times as many source and target molecules. This is the case, for example, of (Q_S,Q_T)=(2,2), which is a potential repeat of (Q_S,Q_T)=(1,1).

The matrix below shows for all (Q_S,Q_T) pairs, where $Q_S,Q_T \leq 6$, those pairs that are potential repeats (denoted by a "R") and those that are not (denoted by a "✓").Over the 36 cases shown there are 23 that are not potential repeats. Hence we have an initial estimate of the probability of choosing the single correct pair associated with the experimentally determined pathway as 1/23.

		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number of molecules Q_S of source compound	1	✓	✓	✓	✓	✓	✓
	2	✓	R	✓	R	✓	R
	3	✓	✓	R	✓	✓	R
	4	✓	R	✓	R	✓	R
	5	✓	✓	✓	✓	R	✓
	6	✓	R	R	R	✓	R

Table 3.7: Matrix of repeats

However this estimate of 1/23 depends upon the maximum values examined for Q_S and Q_T (here 6). Over all the forty experimentally determined pathways we examined the maximum Q_S value is 3 and the maximum Q_T value is 2. If we restrict our investigation to $Q_S \leq 3$ and $Q_T \leq 2$ then we have just five cases that are not potential repeats and hence an initial estimate of the probability of choosing the single correct pair associated with the experimentally determined pathway as 1/5.

So, for the purposes of statistical testing, we shall assume that there is a probability of 1/5 of, by chance, choosing the correct (Q_S,Q_T) pair associated with the experimentally determined pathway. Hence we have the hypothesis test:

H_0: results from the BP model arise due to chance probability of a success = ρ =1/5, where ρ is the probability that the BP model will by chance recover the (Q_S,Q_T) pair associated with the experimentally determined pathway

versus

H_1: results from the BP model do not arise from chance

probability of a success > ρ

The BP model recovers the (Q_S,Q_T) pair associated with the experimentally determined pathway for 48 pathways out of 80, see Table 3.6, so a sample probability of 48/80 = 0.6 of success. The test statistic for this one-sided hypothesis test is (sample probability – ρ)/√[ρ(1-ρ)/(sample size)] = (0.6-ρ)/√[ρ(1-ρ)/10] = (0.6-(1/5))/√[(1/5)(4/5)/80] = 8.94. This is a statistically highly significant result. At the 0.001% level for example the critical value is 4.27, and our test statistic far exceeds this, so H_0 would be rejected and we would conclude that the results from the BP model do not arise from chance. This fact shows that the results obtained for the BP model for recovering the (Q_S,Q_T) pair associated with the pathway are statistically significant at the 0.001% level, as noted above.

3.3.4 Sensitivity analyis relating to Δ

In the BP model it is necessary to specify the user defined input parameter Δ which determines whether a compound is a low presence, or a high presence, compound. We conducted a sensitivity analysis as to how the results change as Δ changes. This can be seen in Table 3.7, where we have summarized the number of “yes” entries that we obtained in the equivalent of Table 3.3 and Table 3.6 for varying Δ values. It is clear from this table that over a fairly wide range of Δ values a significant number of “yes” entries are obtained.

Note here that the value of Δ=4% associated with the results presented in Table 3.3 and Table 3.6 was originally chosen based on limited computational experience with a number of pathways. It was not chosen via systematic enumeration of results for all pathways for a range of Δ values and then selection of the best Δ value. As can be seen from Table 3.7 we could improve the results presented in Table 3.3 and Table 3.6 were we to use Δ=5% for example.

	Number of “yes” entries			
	Pathway structure recovered?		(Q_S,Q_T) recovered if pathway structure recovered?	
Value of Δ (%)	Objective (3.13)	Objective (3.14)	Objective (3.13)	Objective (3.14)
2.5	30	5	30	4
3	33	11	33	9
3.5	33	11	33	9
4	37	14	36	12
4.5	37	29	36	21
5	37	31	36	24
5.5	35	30	34	24

Table 3.8: Sensitivity analysis relating to Δ

3.4 Neglected issues

The BP model neglects three issues: bioenergetics (Gibbs free energy), enzymes and cofactors/coenzymes. Mathematically all of these can be easily incorporated into the BP model. However, the data available for the 880 reactions considered was not sufficient to enable any of these issues to be implemented computationally. We describe below how to extend our model to deal with these issues.

Let G_r be the Standard Gibbs free energy involved in one tick of reaction r (r=1,…,R). The Gibbs free energy provides a measure about the directionality and spontaneity of a particular reaction. Consider, for example, the following reversible reaction α: A + B ↔ C + D. If the Standard Gibbs free energy is less than zero, i.e. $G_\alpha<0$, then A and B will necessarily tend to be converted into C and D. Similarly, if the Standard Gibbs free energy is greater than zero, i.e. $G_\alpha >0$, then C and D will necessarily tend to be converted into A and B. In addition, if the Standard Gibbs free energy is equal to zero, i.e. $G_\alpha =0$, then the reaction has reached equilibrium.

This idea can be easily applied for metabolic pathways. A particular pathway will be energetically feasible if and only if the sum of Standard Gibbs free energy of the active reactions (more precisely the ticks of the reactions) involved in the pathway is less than zero. This can be expressed in the equation below.

$$\sum_{r=1}^{R} \mathrm{G_r t_r} < 0 \tag{3.15}$$

In addition, it might be of interest to determine the metabolic pathway with minimum Standard Gibbs free energy net value. This can be expressed in the objective function below.

$$\text{minimise} \sum_{r=1}^{R} \mathrm{G_r t_r} \tag{3.16}$$

The main reason as to why Standard Gibbs free energies were not included in the BP model was the lack of data for each particular reaction in the metabolic network when the BP model was built. We examined the group contribution methodology presented by of Mavrouniotis (1990; 1991) to estimate Standard Gibbs free energies values. However, results reported in those works did not cover every biochemical reaction in the metabolic network. Full Standard Gibbs data are now available for *E.Coli* in Feist *et al.*, 2007.

An additional point to note here is that it would be more appropriate to utilise real Gibbs free energy, H_r (r=1,…,R), since Standard Gibbs free energy assumes 1 molar concentration for each compound in the cell/organism, which clearly is not true. We give below the chemical formula to calculate real Gibbs free energy, H_α, for the example reaction α.

$$H_\alpha = G_\alpha + RT\frac{[C]\cdot[D]}{[A]\cdot[B]} \qquad (3.17)$$

where R is the Boltzmann constant; T is the absolute temperature; [C] is the cellular concentration of compound C; [D] is the cellular concentration of compound D; [A] is the cellular concentration of compound A; [B] is the cellular concentration of compound B.

The discipline of metabolomics is meant to provide cellular concentrations of compounds in large scale. However, we are still far from obtaining the concentrations for each particular compound. Despite this fact, we think that Gibbs Energy provide an interesting link between computational and experimental methods.

As far as the issue of enzymes is concerned, let E be the total number of enzymes. Let m_{er} be 1 if enzyme e (e=1,…,E) catalyses reaction r (r=1,…,R), 0 otherwise. This notation is general and allows us to have more than one enzyme catalysing a particular reaction. In addition, a particular enzyme might catalyse a number of different reactions.

We need a binary zero-one variable x_e = 1 if we make use of enzyme e (e=1,…,E) in the pathway, 0 otherwise.

We need constraints relating a reaction to the enzyme needed for the reaction. These are:

$$\sum_{e=1}^{E} m_{er}x_e \geq z_r \quad r=1,\ldots,R \qquad (3.18)$$

$$\sum_{r=1}^{R} m_{er}z_r \geq x_e \quad e=1,\ldots,E \qquad (3.19)$$

The first constraint ensures that if a reaction occurs, then at least one of enzymes that catalyses that reaction must be active. The second constraint ensures that if an enzyme is active, then at least one of the reactions catalysed by that enzyme must be active.

It would be an interesting objective function to minimise the number of the enzymes involved in the pathway.

$$\text{minimise} \sum_{e=1}^{E} \mathrm{x_e} \tag{3.20}$$

The main cause for neglecting enzymes in the BP model was that the metabolic network of *E.Coli* presented by Reed *et al.*, 2003, contains a large number of reactions whose catalysing enzyme is still unknown.

Finally, as we explained in the Chapter 2, a cofactor is generally defined as a biochemical compound that fulfils the same specific and secondary function in a considerable number of reactions. However, the list of cofactors for a given metabolic network has not been defined unambiguously. The model described later in Chapter 6 directly deals with this issue.

3.5 Conclusions

We have presented our initial mathematical optimisation model named the BP model so as to recover experimentally determined metabolic pathways. The BP model showed excellent performance, as the pathway structure and (Q_S,Q_T) pair were recovered in 37 out of 40 experimentally determined pathways using one of the two objectives proposed. However, the model needs to know beforehand the unbalanced low presence compounds in the experimentally determined pathway. This constitutes a major drawback for predicting novel (unknown) metabolic pathways. The IBP (Improved Beasley-Planes) model described in Chapter 6 is meant to avoid that prior pathway knowledge. Despite this issue, the results presented here for the 40 *E.Coli* pathways shows that the BP model is more accurate than previous stochiometric approaches, namely elementary flux modes and extreme pathways, for determining biologically meaningful metabolic pathways. In addition, the BP model is more applicable, since it overcomes the combinatorial explosion suffered by previous approaches by using an optimization approach. Moreover it indicates that there is reason to believe there is a general mathematical model underlying the many different experimentally determined pathways seen.

Chapter 4

Pathway disruption, an application of the BP model

Chapter 3 showed the effectiveness of the BP model for recovering experimentally determined pathways. In this chapter we illustrate how the BP model can be used to investigate the disruption of metabolic pathways. In particular, we focus on the Glycolysis pathway, a key ATP producer pathway. Interestingly, our results accords with work done from a non-mathematical (biochemical/medical) perspective.

4.1 Introduction

In previous chapters we explained that in order to search for meaningful metabolic pathways in a metabolic network, experimentally determined pathways provide an appropriate starting point. Despite the issue of low presence unbalanced compounds, the BP model showed high success in recovering experimentally determined (known) pathways from the metabolic network. This fact does give us a degree of confidence that, when we apply the BP model to an unknown situation, the pathway predicted by the BP model will have biological significance.

To illustrate this, assume that Figure 4.2 shows the experimentally determined pathway converting C1 into C7, given our example metabolic network shown in Figure 4.1. Suppose that, once we solve the BP model for this example pathway, we achieve recovery. Suppose now that (due to a genetic disease, for example) the pathway shown in Figure 4.2 presents an enzymatic deficiency, e.g. the enzyme catalysing R1. Such deficiency prevents this pathway from being active. In the literature the phrase "knocked out" is often used to indicate a reaction is disabled/unable to be performed. This situation represents a case in which the active pathway converting C1 into C7 is unknown (under the assumption that the organism continues to transform C1 into C7). Since the BP model obtains biologically significant metabolic pathways, if we apply the BP model with reaction R1 "knocked out", i.e., $z_{R1}= 0$, then the pathway predicted by the BP model may have biological significance.

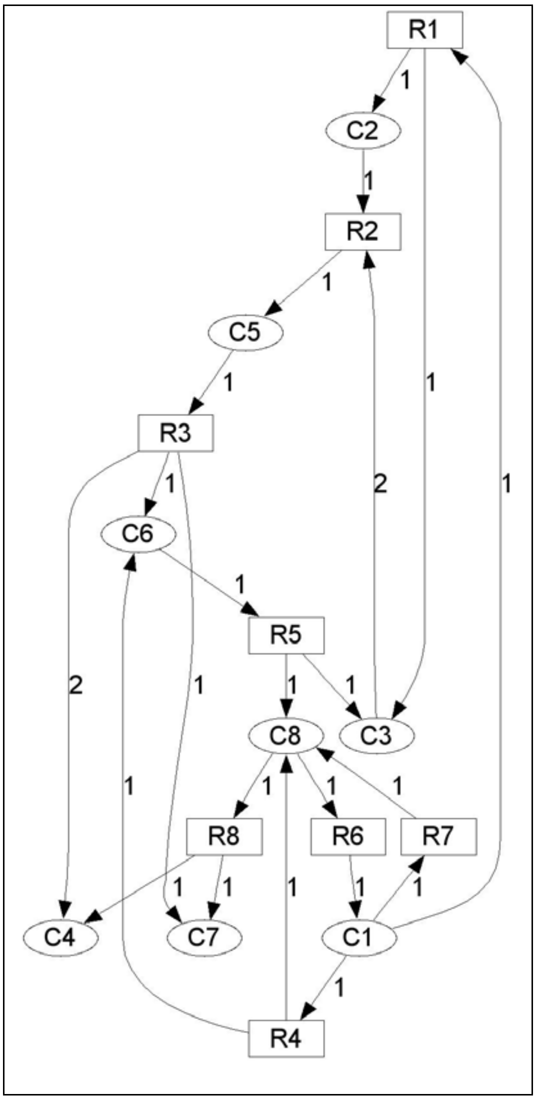

Figure 4.1: Previous Figure 1.2 in Chapter 1

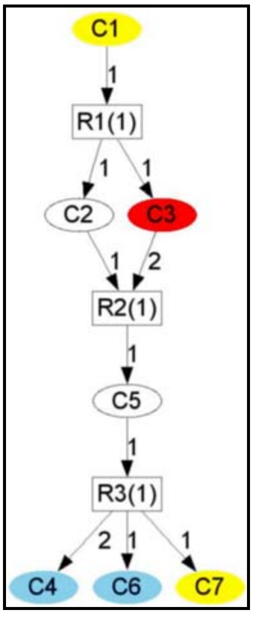

Figure 4.2: Experimentally determined pathway converting C1 into C7

In this chapter the BP model is applied to the unknown situation that arises when one or more reactions are knocked out and an organism must adapt by utilising previously unutilised pathways. Note here that we specifically refer to reaction knockout as a convenient shorthand way of saying inhibit the enzyme(s) that catalyses a particular reaction. Enzyme inhibition can take place either using gene-based inhibition of enzyme production, or by pharmacological means. In addition, note that as enzymes catalyse reactions in both directions when we refer to reaction knockout we implicitly mean knockout a reaction and its reverse (if it exists). The reason as to why we might be interested in reaction knockout is that we wish to have a means to deliberately disrupt a metabolic pathway. This may be, for example, because we wish to disable/kill an organism utilising that pathway.

Knockout approaches given in the literature typically build upon flux balance analysis (FBA), (Kauffman *et al.,* 2003; Lee *et al.*, 2006; Price *et al.*, 2004). The starting point for FBA is a known set of biochemical (enzyme catalysed) reactions that can take place in an organism, i.e. the metabolic network of a given organism. Given the metabolic network the basic assumption in FBA is that a steady state applies and the majority of metabolites in the organism must be in balance. These balance constraints are the essential FBA constraints and they can be written in a mathematical form using reaction flux vectors. As a result of this balance assumption linear programming can be applied to find the effect (for example) on organism growth (biomass production) if a particular reaction is knocked out (deleted, eliminated, so it has zero flux). These network-based approaches to reaction knockout however make no use of the information that certain sets of reactions are commonly recognised as grouped together into pathways. Hellerstein, 2007, has recently argued the need for a pathway-based approach to reaction knockout. We present below a detailed review of mathematical approaches in the literature to reaction knockout.

In this chapter we present a novel knockout approach based upon the BP model. Though the BP model also includes FBA constraints, it fundamentally stands for a pathway-based approach. We apply our approach to disrupt the operation of a given experimentally determined metabolic pathway. In particular we will focus here on Glycolysis. This pathway is currently of interest due to a recent revival of interest in the Warburg effect, that cancer cells utilise Glycolysis, and hence disrupting that pathway may be of benefit in fighting cancer (Garber, 2004; Xu *et al.*, 2005; Bui and Thompson, 2006; Fantin *et al.*, 2006; Gatenby and Gillies, 2004; Mathupala *et al.*, 2006).

4.2 Review of network-based approaches to reaction knockout

Firstly, we would note here that we focus our discussion on reaction knockout, whereas often the literature uses the phrase "gene knockout". The difficulty with focusing directly on gene knockout is that it neglects the effect of isoenzymes (isozymes), where two or more enzymes catalyse the same reaction. By focusing on reaction knockout we automatically account for isoenzymes. Moreover in the literature it is not uncommon to see that although authors may use the phrase "gene knockout" the mathematical/computational details of their procedure make it clear that they are actually investigating reaction knockout (e.g. by setting flux for a reaction to zero in a FBA based approach).

As described above, FBA has been applied in the literature as a basis for investigating knockout. For example, Edwards and Palsson, 2000a, used FBA to investigate knockout in *E. coli* MG1655; Edwards and Palsson, 2000b, knockout in *E. coli* K-12. Burgard and Maranas, 2001, used FBA together with mixed-integer linear programming to investigate the maximum number of knockouts possible for *E. coli* whilst maintaining a specified level of biomass production; they also investigated knockin, i.e. reaction addition, the reverse of knockout.

Burgard *et al.*, 2003, presented a bilevel optimisation approach, OptKnock, where the outer optimisation objective is to maximise a given flux, and the inner optimisation objective is FBA based (maximise biomass). In their approach the number of knockouts chosen is bounded above by a prespecified value. Fong *et al.*, 2005, applied OptKnock to *E. coli* to identify knockouts associated with increasing lactate production. Pharkya and Maranas, 2006, extended OptKnock to OptReg, where as well as reaction knockout they include options relating to reactions being repressed/activated (down/up regulated, having flux values much lower/higher than their steady state (FBA based) flux values).

Pharkya *et al.*, 2004, presented OptStrain, that first considers knockin, then knockout. In their procedure maximal yield for a target product is first calculated when it is possible to utilise any of a large set of possible reactions (this set including not only reactions native to the organism, but also other reactions that are non-native). Then the minimum number of non-native reactions that provide maximal yield are calculated (using mixed-integer linear programming, with explicit enumeration of alternative optimal solutions). Finally OptKnock (Burgard *et al.*, 2003) is used to identify knockouts in the organism composed of native reactions and those non-native reactions identified at the previous step as members of a non-native minimal set.

Although most FBA knockout work has been done with *E. coli* work with other organisms has also been presented in the literature. For example Borodina *et al.*, 2005, considered *Streptomyces coelicolor*; Duarte *et al.*, 2004, and Deutscher *et al.*, 2006, considered the yeast *Saccharomyces cerevisiae*; Thiele *et al.,* 2005, considered *Helicobacter pylori.*

Knockout (and knockin) approaches based upon FBA however effectively assume that the entire flux vector can be changed from its initial state (before knockout/knockin) to an entirely new state (after knockout/knockin), in effect a complete "rerouting" of the fluxes in the organism. Here the literature has been predominantly concerned with knockout and various authors have argued that although FBA knockout approaches may give long-term evolutionary insight into how an organism might eventually adapt to a knockout, it is less effective at predicting the immediate response of an organism to a knockout.

For this reason approaches aimed at focusing on immediate flux changes after knockout have appeared in the literature. Minimisation of metabolic adjustment (MOMA), Segrè *et al.*, 2002, minimises the Euclidean distance between the flux vector before knockout and the flux vector after knockout. In MOMA the Euclidean objective adopted tries to ensure that there are only "small" changes between the flux vector before knockout and the flux vector after knockout. Alper *et al.*, 2005, used MOMA to investigate multiple knockouts in *E. coli* when focusing on lycopene biosynthesis.

In regulatory on/off minimisation of metabolic flux (ROOM), Shlomi *et al.,* 2005, mixed-integer linear programming is used to find the flux after knockout that minimises the number of "significant" flux changes compared with the flux before knockout (where a key conceptual and computational issue is how large a flux change has to be before it is classified as "significant"). Shlomi *et al.,* 2005, with respect to *E. coli*, give an example where the number of significant flux changes after knockout are 12 with ROOM, 317 with MOMA and 119 with FBA.

One theme encountered in FBA based knockout is that of classifying reactions as "essential" or not. For example, essential reactions may be defined as those whose knockout renders the organism ineffective (e.g. unable to grow at all, or at best grow very slowly, as compared with the organism before knockout). Work of this kind can be seen, for example, in Borodina *et al.*, 2005; Burgard and Maranas, 2001; Deutscher *et al.*, 2006, Segrè *et al.*, 2002; Shlomi *et al.,* 2005.

4.3 Pathway-based knockout approach

In this section we present our approach to knockout based upon the BP model for metabolic pathways. To motivate our approach suppose that, for biomedical reasons, we wish to disrupt the operation of a known metabolic pathway by knocking out reactions within it. Clearly within a pathway there are many options for reaction knockout. For example if the pathway involves K reactions then there are K choices for single reaction knockout; K(K-1)/2 choices for two reaction knockout; K(K-1)(K-2)/6 choices for three reaction knockout; etc. Of these many choices which one should we adopt? The reason why we might be interested in more than single reaction knockout with respect to a pathway is clear – the organism may adapt to a single knockout via a limited rerouting of flux within the pathway (whilst also perhaps utilising other reactions not in the pathway), the more reactions associated with the pathway that are knocked out, the more difficult this becomes.

In our approach, we examine all of these possible knockout choices and select the best one. Whilst this might at first sight seem computationally expensive the fact that we are adopting a pathway based approach means that K is relatively small. Over the forty pathways we consider the maximum number of reactions in a pathway is K=10. Hence there are only (at most) 10 choices for single reaction knockout; K(K-1)/2 = 10(10-1)/2 =45 choices for two reaction knockout; K(K-1)(K-2)/6 = 10(10-1)(10-2)/6 = 120 choices for three reaction knockout. Approaches in the literature rarely go beyond single or two reaction knockout. As reported below examining a single choice is not computationally expensive and so our approach is computationally feasible. Moreover, as illustrated below, since we explicitly examine all choices we are able not only to identify the best choice, but also the second-best, third-best, etc, i.e. to produce a ranked list of choices. This ranking enables skilled professionals to bring to bear considerations that cannot be easily incorporated into a mathematical model on a restricted range of choices in order to make a final considered judgment with regard to the choice to select.

In order to illustrate our approach we discuss below both single reaction knockout, and two reaction knockout, in *E. coli* when our aim is to disrupt the Glycolysis pathway, a key ATP producing pathway, which transforms D-glucose into pyruvate (see Figure 4.3).

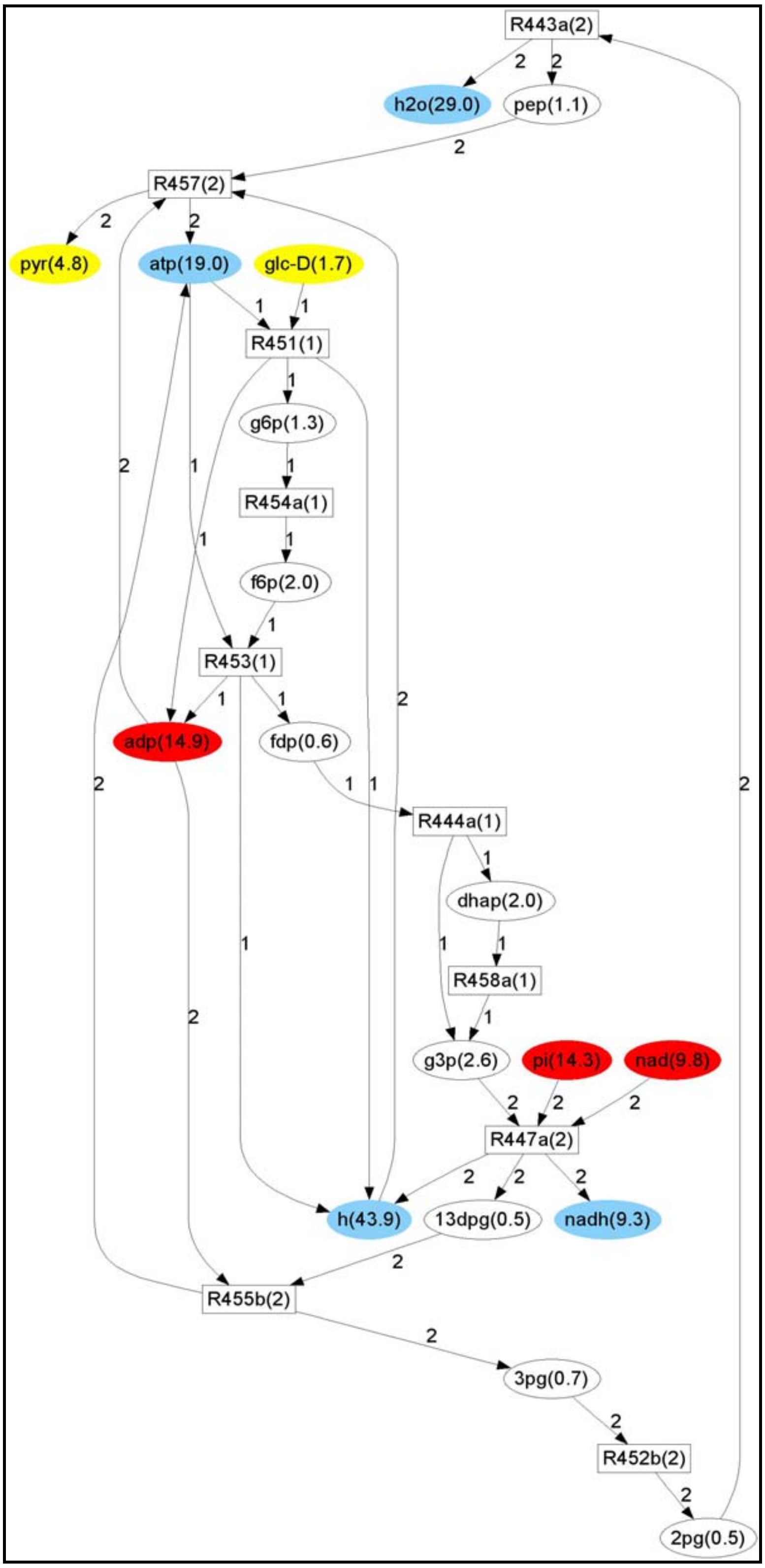

Figure 4.3: Glycolysis pathway

4.3.1 Single reaction knockout

In order to decide the best reaction to knockout we adopt the procedure below:

For each active reaction R1 in the pathway under consideration:

- apply the BP model when reaction R1 is knocked out for each (Q_S,Q_T) pair (Q_S,$Q_T \leq 6$), where Q_S represents the number of molecules of source compound, Q_T the number of molecules of target compound. This was carried out for each of the ten reactions involved in the Glycolysis pathway. Table 4.1, for example, shows results when reaction R444 was deleted.

		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number	1	X	(10,1)*	X	X	X	X
of	2	X	X	X	(10,2)	X	X
molecules	3	X	X	X	X	X	(10,3)
Q_S of	4	X	X	X	X	X	X
source	5	X	X	X	X	X	X
compound	6	X	X	X	X	X	X

Table 4.1: Results when R444a was deleted

- identify the dominant (Q_S,Q_T) pair with respect to our objective (primary weight on minimising the total number of reactions and secondary weight on maximising excess ATP) for each. This pathway represents for the best alternative pathway given the knockout of the reaction considered. Let (a,b) be the (number of reactions, excess ATP) involved in this alternative pathway. In the above example, the pairs seen down the diagonal are all repeats of each other, doubling and then tripling the number of source and target molecules (and excess ATP). The dominant pair is (Q_S,Q_T)=(1,2) when reaction R444 is deleted, thus (a,b)= (10,1). Situations where the Beasley-Planes model indicated that no feasible solution exists are indicated by a 'X'.

Now over all reactions R1 considered choose the reaction that maximises the total number of reactions (a), ties broken by minimising total excess ATP (b), and further broken if necessary by minimising the number of other pathways in which R1 is involved.

Here we are choosing the worst alternative pathway out of the cases considered, according to the BP objective of giving primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP. The reaction whose knockout corresponds to this worst alternative pathway is the best reaction to knockout to disrupt the functioning of the pathway, because the aim behind reaction knockout is to make the functioning of the subsequent pathway after reaction knockout as difficult as possible. In the event that there are two (or more) reactions that tie for consideration here then we tie-break by choosing the reaction involved in as few other pathways as possible. Note here that in this procedure we do not constrain the pathway after reaction knockout to consume/produce the same number of molecules of source/target compound as in the original (undisrupted) pathway.

Table 4.2 gives for each reaction the dominant (Q_S,Q_T) pair and the corresponding (number of reactions, excess ATP)= (a,b) in the Glycolysis pathway. Note that the Glycolysis pathway itself has (number of reactions, excess ATP) = (10,2) and, as we would expect, the pathways we have after single reaction knockout are worse with respect to these (combined) characteristics than the original pathway.

Reaction deleted	Dominant (Q_S,Q_T) pair	(Number of reactions, excess ATP)
R444a	(1,2)	(10,1)
R443a	-	Infeasible
R447a	-	Infeasible
R451	-	Infeasible
R452b	-	Infeasible
R453	(1,2)	(10,1)
R454a	(1,2)	(10,1)
R457	(1,2)	(14,-1)
R455b	-	Infeasible
R458a	(1,2)	(10,1)

Table 4.2: Best alternative pathway for each reaction in Glycolysis pathway

Considering the above table, and ignoring the five infeasible cases for the moment, we can see that from the set of reactions {R444a, R453, R454a, R457, R458a} the worst alternative pathway is that corresponding to the knockout of R457, since that involves 14 reactions and consumes one molecule of ATP. This is worse than any other (feasible) case - recall that our objective is to give primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP.

It is clear from all ten entries in the above table that, if we wish to best disrupt the Glycolysis pathway, we should choose one of the five reactions {R443a, R447a, R451, R452b, R455b}, as the Beasley-Planes model indicates that with any of these reactions knocked out it will not be possible to transform D-glucose to pyruvate.

To make a choice from this reaction set we examine the number of pathways in which these reactions appear. We have, in our work, 40 pathways. Of these 40 pathways R443a, R447a, R452b and R455b each appear in two pathways (Gluconeogenesis and Glycolysis), whilst R451 only appears in Glycolysis. Since in deleting a reaction we cannot be blind to the effect on other (known) pathways it appears that the best choice for deletion is R451.

In the light of the above table, and of the discussion as to "tie-breaking" between reactions by looking at the number of known pathways in which reactions are involved, we can form an ordered list of reactions as below. Here the reactions are ordered in increasing order of their attractiveness for knockout, in terms of (number of reactions, excess ATP), ties broken by number of known pathways the reaction is involved in. Another way to view this ordered list is that we are ranking reactions in terms of their potential for disrupting the pathway when disabled.

Reaction deleted	(Number of reactions, excess ATP)	Number of known pathways the reaction is involved in
R451	Infeasible	1
R443a, R447a, R452b, R455b	Infeasible	2
R457	(14,-1)	1
R453	(10,1)	1
R444a, R454a	(10,1)	2
R458a	(10,1)	3

Table 4.3: Ranked list of reactions in increasing potential for disruption

In the table above the first reaction listed, R451, has the highest potential for disrupting the pathway when knocked out. We then have {R443a, R447a, R452b, R455b}, then R457, then R453, then {R444a ,R454a} and finally R458a. Here, reactions grouped together are those that our approach regards as being of equal potential.

Of course, given the above table different orderings are possible. For example we might wish to place more weight on not interfering with other pathways and so rank in order of the last column in the above table. Whatever the ranking criteria adopted it is clear that the approach presented above does provide a structure to deciding which reactions might be possible targets for deletion in terms of disrupting a pathway.

In summary, when we apply this procedure to our chosen Glycolysis pathway we find that the best reaction to knockout is R451, namely *atp* + *glc-D* → *adp* + *g6p* + *h*. This reaction is catalysed by the enzyme (protein) hexokinase and so our approach indicates that this enzyme would be a good target in order to disrupt Glycolysis. Although this suggestion for targeting hexokinase has come from a purely mathematical analysis it is interesting to note that other workers (Mathupala *et al.*, 2006; Xu *et al.*, 2005), working purely from a biochemical/medical perspective, have also recently drawn attention to targeting hexokinase in terms disrupting Glycolysis.

4.3.2 Two reaction knockout

One issue in relation to our mathematical approach to disrupting a metabolic pathway is the purpose of the pathway. Taking Glycolysis as an example is the (primary) purpose of this pathway to transform D-glucose into pyruvate or to produce/consume another compound (e.g. to produce ATP)?

In terms of the disruption approach considered above we have regarded the purpose of the pathway as being related to transformation of the source compound into the target compound. Our approach assumes that the organism will attempt to continue that transformation after disruption. However, if the primary purpose of the pathway is to produce/consume another compound, the organism may switch to an already existing pathway that performs the same function (albeit possibly with less effectiveness).

If the purpose of the pathway is to produce/consume another compound then the disruption approach given above is still applicable, but slightly more complicated. Suppose for example we assume that the purpose of the Glycolysis pathway is to produce ATP – hence if that pathway is disrupted the organism will utilise another (known) pathway to produce ATP. Here therefore, in an attempt to disrupt ATP production, we face the problem of simultaneously disrupting two known pathways.

Our approach is easily extendable to the problem of simultaneously disrupting two (or even more) pathways. Considering simultaneous disruption of two pathways, and focusing on single reaction knockout in each pathway, then we simply apply the BP model to both pathways a number of times, each time with two reactions (one from one pathway, the other from the other pathway) knocked out.

With respect to *E. coli* an alternative to Glycolysis for ATP production is the Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle shown in Figure 4.4.

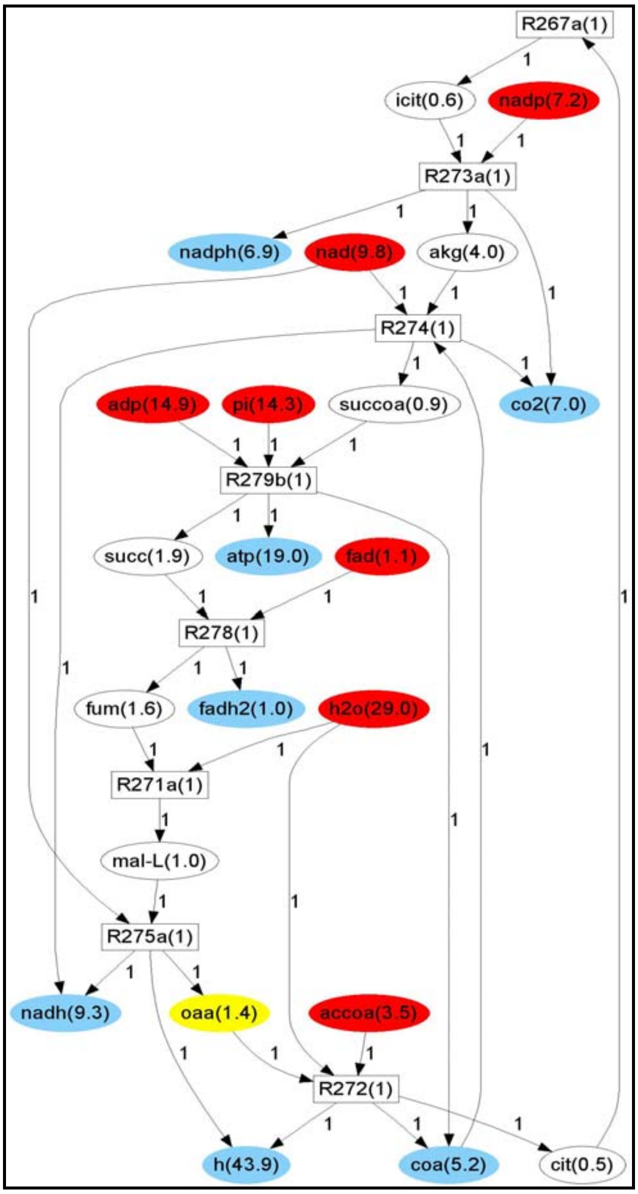

Figure 4.4: TCA Cycle pathway

In more detail therefore our approach here is:

For each reaction pair, one reaction R1 from Glycolysis, the other reaction R2 from the TCA cycle:

- apply the BP model to identify the best alternative pathway to Glycolysis when both reactions R1 and R2 are knocked out using objective (3.13). Let (a,b) be the (number of reactions, excess ATP) involved in this alternative pathway.

- apply the BP model to identify the best alternative pathway to the TCA cycle when both reactions R1 and R2 are knocked out using objective (3.14). Let (c,d) be the (number of reactions, excess ATP) involved in this alternative pathway.

Now over all pairs of reactions R1 and R2 considered choose the pair that maximises the total number of reactions (a+c), ties broken by minimising total excess ATP (b+d) and further broken if necessary by minimising the number of other pathways in which R1 and R2 are involved. Here we are choosing the worst alternative pathways according to the BP objective of giving primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP. The pair of reactions whose knockout corresponds to these worst alternative pathways is the best pair of reactions to knockout to disrupt ATP production by the Glycolysis and TCA pathways.

When we apply this procedure we find that there are eighty pairs of reactions (one from Glycolysis, one from TCA) that need to be examined. Although this might seem large the BP model requires (on average) only 4.6 seconds (1.86Ghz pc, 2GB RAM) each time it is solved and so total computation time is not excessive. Over these eighty reaction pairs we find that the best reaction pair to knockout involves R451 *atp* + *glc-D* → *adp* + *g6p* + *h* from Glycolysis and R273 *icit* + *nadp* ↔ *akg* + *co2* + *nadph* from TCA. Here this first Glycolysis reaction is as discussed above, this second TCA reaction is catalysed by isocitrate dehydrogenase (IDH). Work presented from a biochemical (rather than a mathematical) perspective has emphasised the importance of this reaction in TCA (Kabir and Shimizu, 2004).

Again, since we have examined all choices, we can produce a ranked list of all eighty reaction pairs in terms of their potential for (in this instance) disrupting ATP production by the Glycolysis and TCA pathways. Table 4.4 show results for the eighty pairs of reactions under study in decreasing disruption potential. Here, for example, there are two reaction pairs in equal second place in this ranked list {R447*;* R455} from Glycolysis and R273 from TCA. Next in this ranked list we have {R443*;* R452} from Glycolysis and R273 from TCA. Next in this ranked list we have R451 from Glycolysis and {R271*;* R274*;* R278*;* R279}from TCA and so onwards. The last place in this ranked list is R458 from Glycolysis and {R267*;* R272} from TCA. Note that we found in total 19 ranks for the disruption of ATP production by the Glycolysis and TCA pathways.

Reaction knocked out in Glycolysis	Reaction knocked out in TCA Cycle	(a,b)	(c,d)	Number of known pathways the pair is involved	Disruption Rank
R451	R273a	infeasible	(18,2)	1	1
R447a	R273a	infeasible	(18,2)	2	2
R455b	R273a	infeasible	(18,2)	2	2
R443a	R273a	infeasible	(13,1)	2	3
R452b	R273a	infeasible	(13,1)	2	3
R451	R271a	infeasible	(3,0)	0	4
R451	R274	infeasible	(3,0)	0	4
R451	R278	infeasible	(3,0)	0	4
R451	R279b	infeasible	(3,0)	0	4
R443a	R274	infeasible	(3,0)	1	5
R443a	R278	infeasible	(3,0)	1	5
R443a	R279b	infeasible	(3,0)	1	5
R447a	R271a	infeasible	(3,0)	1	5
R447a	R274	infeasible	(3,0)	1	5
R447a	R278	infeasible	(3,0)	1	5
R451	R275a	infeasible	(3,0)	1	5
R457	R275a	infeasible	(3,0)	1	5
R452b	R271a	infeasible	(3,0)	1	5
R452b	R274	infeasible	(3,0)	1	5
R452b	R278	infeasible	(3,0)	1	5
R452b	R279b	infeasible	(3,0)	1	5
R455b	R271a	infeasible	(3,0)	1	5
R455b	R274	infeasible	(3,0)	1	5
R455b	R278	infeasible	(3,0)	1	5
R455b	R279b	infeasible	(3,0)	1	5
R443a	R275a	infeasible	(3,0)	2	6
R447a	R275a	infeasible	(3,0)	2	6
R451	R267a	infeasible	(3,0)	2	6
R451	R272	infeasible	(3,0)	2	6
R452b	R275a	infeasible	(3,0)	2	6
R455b	R275a	infeasible	(3,0)	2	6
R443a	R267a	infeasible	(3,0)	3	7
R443a	R272	infeasible	(3,0)	3	7
R447a	R267a	infeasible	(3,0)	3	7
R447a	R272	infeasible	(3,0)	3	7
R452b	R267a	infeasible	(3,0)	3	7

R452b	R272	infeasible	(3,0)	3	7
R455b	R267a	infeasible	(3,0)	3	7
R455b	R272	infeasible	(3,0)	3	7
R443a	R271a	infeasible	(3,1)	1	8
R447a	R279b	infeasible	(3,1)	1	8
R453	R273a	(10,1)	(18,2)	1	9
R444a	R273a	(10,1)	(18,2)	2	10
R454a	R273a	(10,1)	(18,2)	2	10
R458a	R273a	(10,1)	(18,2)	3	11
R457	R271a	(14,1)	(3,0)	0	12
R457	R274	(14,1)	(3,0)	0	12
R457	R278	(14,1)	(3,0)	0	12
R457	R279b	(14,1)	(3,0)	0	12
R457	R273a	(14,1)	(3,0)	1	13
R457	R267a	(14,1)	(3,0)	2	14
R457	R272	(14,1)	(3,0)	2	14
R453	R271a	(10,1)	(3,0)	0	15
R453	R274	(10,1)	(3,0)	0	15
R453	R278	(10,1)	(3,0)	0	15
R453	R279b	(10,1)	(3,0)	0	15
R444a	R271a	(10,1)	(3,0)	1	16
R444a	R274	(10,1)	(3,0)	1	16
R444a	R278	(10,1)	(3,0)	1	16
R444a	R279b	(10,1)	(3,0)	1	16
R453	R275a	(10,1)	(3,0)	1	16
R454a	R271a	(10,1)	(3,0)	1	16
R454a	R274	(10,1)	(3,0)	1	16
R454a	R278	(10,1)	(3,0)	1	16
R454a	R279b	(10,1)	(3,0)	1	16
R444a	R275a	(10,1)	(3,0)	2	17
R453	R267a	(10,1)	(3,0)	2	17
R453	R272	(10,1)	(3,0)	2	17
R454a	R275a	(10,1)	(3,0)	2	17
R458a	R271a	(10,1)	(3,0)	2	17
R458a	R274	(10,1)	(3,0)	2	17
R458a	R278	(10,1)	(3,0)	2	17
R458a	R279b	(10,1)	(3,0)	2	17
R444a	R267a	(10,1)	(3,0)	3	18
R444a	R272	(10,1)	(3,0)	3	18
R454a	R267a	(10,1)	(3,0)	3	18
R454a	R272	(10,1)	(3,0)	3	18
R458a	R275a	(10,1)	(3,0)	3	18
R458a	R267a	(10,1)	(3,0)	4	19
R458a	R272	(10,1)	(3,0)	4	19

Table 4.4: Ranked list of two reactions knockout in decreasing potential for disruption

4.4 Conclusions

In this chapter we have discussed how the BP model can be used to investigate pathway disruption (reaction knockout). Our approach is a pathway-based approach, that (as the Beasley-Planes model) builds on flux balance analysis. It involves explicit enumeration and evaluation of all knockout choices and hence is able not only to choose the optimal (best) knockout but also to provide an explicit ranking of all possible knockout choices. Note that our approach is general. Certainly we could apply the same approach as in this chapter using a different mathematical model for metabolic pathways (with a known objective in terms of factors like number of reactions and excess of ATP).

We have shown that for the Glycolysis pathway the prediction from our knockout approach in terms of the enzyme/reaction to target to best disrupt the pathway accords with work done from a non-mathematical (biochemical/medical) perspective. Moreover, as illustrated for two key ATP producing pathways, it is applicable to the problem of simultaneously disrupting two (or more) pathways. Predicting how one might best disrupt metabolic pathways via mathematics, before undertaking any laboratory work, is clearly of value.

Chapter 5

Analysis of path finding approaches

As described in Chapter 2, path finding approaches to metabolic pathways adopt a graph theory approach to the problem of determining biologically meaningful metabolic pathways. Although path finding approaches are often regarded as a promising concept for analysing metabolic pathways little validation has been carried out. In this chapter the effectiveness of using compound node connectivities in a path finding approach is examined. In addition, an approach to path finding based upon integer programming is presented.

5.1 Introduction

The BP model needs a prior classification of the biochemical compounds, namely dividing compounds into low presence and high presence compounds. Low presence compounds must be in aggregate (net) terms balanced. By balanced we are referring to a stoichiometric related balance, namely a compound is balanced if the total number of molecules consumed by the reactions involved in the pathway is equal to the total number of molecules produced by the reactions involved in the pathway. Whereas, high presence compounds are stoichiometrically unconstrained, i.e. they could be in aggregate (net) terms produced to excess, consumed (freely available) or balanced.

These balancing related constraints, equation (3.10) of the BP model, reduce the difficulty of finding biologically meaningful metabolic pathways. In addition, the majority of the biochemical compounds active in a metabolic pathway satisfy equation (3.10), as noted in Chapter 2. However, there exist experimentally determined pathways which have low presence unbalanced compounds, as described in Chapter 3. These pathways, as not satisfying equation (3.10), will never be determined by the BP model without prior knowledge. In order to fix this problem, the BP model did not force such low presence compounds to be balanced. However this strategy cannot be extended for unknown metabolic pathways, as the BP model needs to know beforehand the low presence unbalanced compounds of the metabolic pathway. This fact constitutes the major limitation of the BP model. This limitation also applies to stoichiometric approaches (Schilling *et al.*, 2000; Schuster *et al.*, 2000), as described in Chapter 2.

In contrast to the BP model and stoichiometric approaches, path finding approaches do not include stoichiometric constraints and thus avoid the limitation described above. This is the reason as to why we analyse path finding approaches in detail.

Recalling Chapter 2, path finding approaches focus on the fact that there is a (directed) path (containing no cycles) from the source compound to the target compound in experimentally determined metabolic pathways. We refer to this directed path as the metabolic path for a particular experimentally determined metabolic pathway. This path may not be unique, in particular when the pathway is branched. To illustrate this, suppose that, given our example metabolic network shown in Figure 5.1, the subgraph shown in Figure 5.2 is the experimentally determined metabolic pathway that converts C1 into C7. Here, for example, we have the two metabolic paths shown in Figure 5.3.

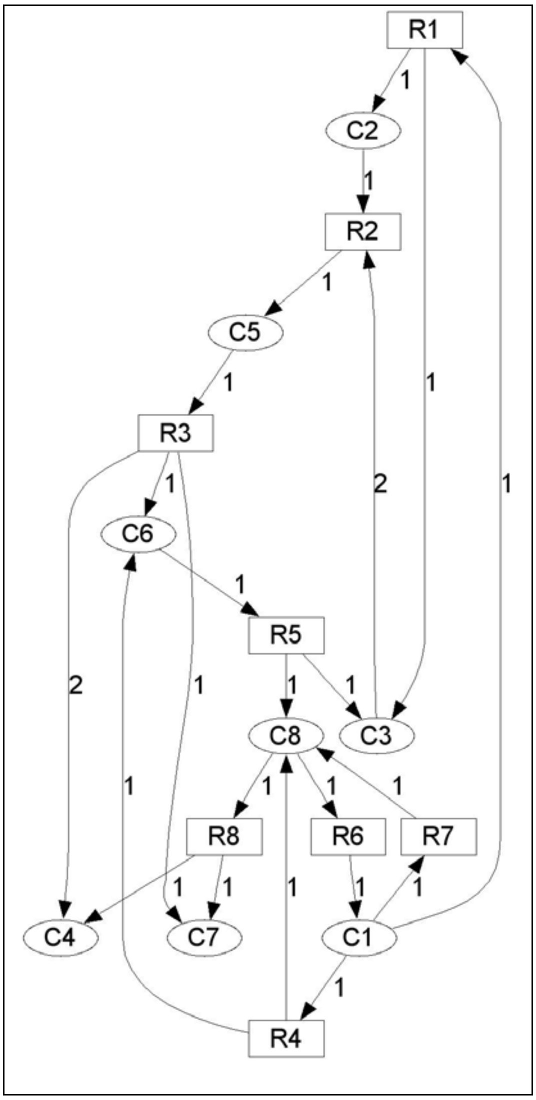

Figure 5.1: Previous Figure 1.2 in Chapter 1

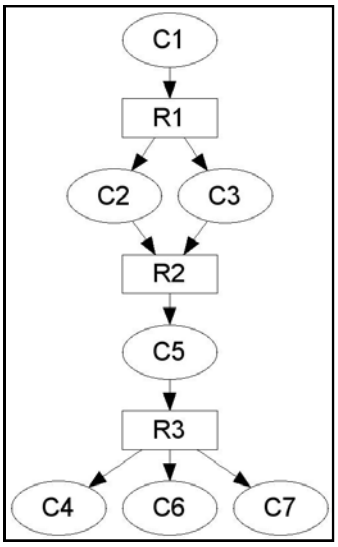

Figure 5.2: Experimentally determined pathway converting C1 into C7

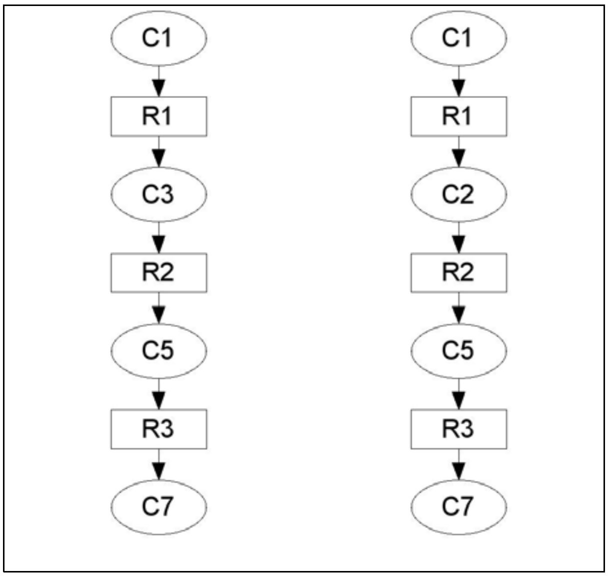

Figure 5.3: Two metabolic paths associated with the pathway shown in Figure 5.2

Note, as in Figure 5.2 and 5.3, the difference between a metabolic pathway and a metabolic path. The metabolic pathway contains all the reactions and compounds involved in the pathway. The metabolic path is a directed path from the source compound to the target compound in the metabolic pathway and may (as in both the metabolic paths seen in Figure 5.3) contain only a subset of intermediate reactions/compounds.

Since experimentally determined pathways do have at least one directed path from the source compound to the target compound, it is plausible to assume that unknown metabolic pathways will also contain directed paths from the source compound to the target compound. Thus, the key assumption behind path finding approaches is that finding directed paths between the source compound and the target compound in the entire metabolic network will give insight into the intermediate reactions/compounds used in unknown metabolic pathways.

Clearly the number of metabolic paths from a source compound to a target compound is very high for a given metabolic network and not every path will have biological significance. Accordingly path finding approaches have evolved to define a suitable distance metric and then find k-shortest paths (with k small) from the source compound to the target compound in the metabolic network. The approaches described in Chapter 2 (Küffner *et al.*, 2000; Arita *et al.*, 2000; McShan *et al.*, 2003; Dooms *et al.*, 2005;

Rahman *et al.*, 2005; Croes *et al.*, 2005, 2006) propose different distance metrics aimed to provide biological significance to the k-shortest paths.

Typically the effectiveness of a path finding approach is examined by seeing how well it performs with respect to a known metabolic pathway. Note here that, in terms of the metabolic pathway, we need only focus on the reactions involved in the metabolic path, as for each reaction we know the set of compounds involved. For example, both of the metabolic paths shown in Figure 5.3 involve reactions R1, R2 and R3. Since, from Figure 5.2, these are the only reactions involved in the metabolic pathway then, for this example, knowledge of either of the metabolic paths shown in Figure 5.3 would give us complete insight into the underlying metabolic pathway as shown in Figure 5.2. In particular note here that in path finding approaches we do not seek stoichiometric information as to the pathway (e.g. number of reaction ticks, number of the molecules of source/target compound involved).

In this chapter we examine the effectiveness of a distance metric based on compound node connectivities, as initially proposed by Croes *et al.*, 2005; 2006, in ten *E.Coli* metabolic pathways. The work of Croes *et al.*, 2005; 2006, is unusual in that they only consider paths from a source reaction to a target reaction (which we denote as the R-R case). In our analysis, we also consider paths from a source compound to a target compound (which we denote as the C-C case). Moreover we present results for higher values of k (up to k=10) than Croes *et al.*, 2005; 2006, (they considered up to k=5) so as to see the benefit of increasing the number of shortest paths considered.

One important point to note here is that the ten *E.Coli* pathways examined in this chapter are precisely the first ten pathways studied in Chapter 3. Pathways 11-40 were excluded from the analysis carried out in this chapter as we think pathways 1-10 are sufficient to evaluate the scope and effectiveness of path finding approaches.

It is clear from our reading of the various papers discussed in Chapter 2 (Küffner *et al.*, 2000; Arita *et al.*, 2000; McShan *et al.*, 2003; Dooms *et al.*, 2005; Rahman *et al.*, 2005;

Croes *et al.*, 2005, 2006) that authors have taken an approach to calculating paths based upon algorithms such as breadth-first and depth-first search. Such algorithms, although relatively easy to code, are often computationally ineffective (especially for paths that involve many nodes). As such a detailed examination of papers discussed above often reveals some choice being made so as to limit computational effort. For example:

- Dooms *et al.*, 2005, constraint programming, impose a limit of the size of the metabolic network

- Croes *et al.*, 2006, depth-first search, impose an upper limit on the number of nodes in the path and the total length of the path

Such choices, whilst being necessary for computational reasons, do mean that the paths found may not (in fact) be optimal, i.e. there may exist shorter paths that have been missed because of these heuristic choices.

In addition algorithms such as breadth-first and depth-first search do not produce paths in increasing distance order, i.e. they do not first find the (k=1) shortest path; then the (k=2) second shortest path; etc. Rather the entire search algorithm must be allowed to finish enumerating paths in the directed graph (many of which will be irrelevant) before all of the k-shortest paths are known.

In this chapter we are concerned with finding k-shortest paths between a source compound and a target compound in the metabolic network, where in each path no node appears more than once. In order to do this we present below an integer programming approach that produces paths in increasing distance order. Although other approaches to finding k-shortest paths are available, e.g. see Guerriero *et al.*, 2001, we believe using integer programming does have advantages over previous approaches used for calculating k-shortest paths in the metabolic pathway literature, in terms of:

- producing paths in increasing distance order; and

- guaranteeing that the paths found will be optimal.

We should stress here that we are aware that in the Operational Research literature there are effective algorithms presented for optimally computing k-shortest paths in increasing distance order. We have not implemented them in this thesis as we initially believed we might build on k-shortest paths and add additional constraints, so an integer programming approach could offer more flexibility than a specialised algorithm (in fact this is not done in this chapter, but is done in Chapter 6).

5.2 Integer programming approach

5.2.1 Formulation

In our approach we have a metabolic network of R reactions (where each reaction has a specified direction so a reversible reaction contributes two different reactions to the total number R), which collectively involve C different compounds. Let m_{cr} have the value 1 if compound c is an input compound for reaction r, 0 otherwise. Let d_{rc} have the value 1 if compound c is an output compound from reaction r, 0 otherwise. Let $W_c = \sum_{r=1}^{R} \max(m_{cr}, d_{rc})$ be the connectivity of compound c, i.e. the number of reactions in which the compound appears in the database of reactions. Since no compound is both input and output from the same reaction, i.e. we cannot have $m_{cr}=d_{rc}=1$ for c=1,…, C, r=1,…, R, W_c can also be viewed as the sum of the in-degree and out-degree of compound c in the directed graph representation, specifically $W_c = \sum_{r=1}^{R} (m_{cr}+d_{rc})$.

Suppose we are seeking the shortest path from a source node S to a target node T in our directed graph representation where, for ease of exposition, we assume below that S and T are compound nodes. Amending the formulation given below if S and T are reaction nodes is easily done.

5.2.1.1 Variables

We need to decide the arcs involved in the metabolic path, so our zero-one (binary, integer) variables are:

- u_{cr} = 1 if the arc from compound node c to reaction node r is in the metabolic path; 0 otherwise
- v_{rc} = 1 if the arc from reaction node r to compound node c is in the metabolic path; 0 otherwise

If m_{cr}=0, i.e. the arc does not exist, then we fix u_{cr} to 0; similarly if d_{rc}=0 we fix v_{rc} to 0. This enables us to present the constraints below in a simplified form.

5.2.1.2 Constraints

The constraints are:

$$\sum_{r=1}^{R} u_{Sr} = \sum_{r=1}^{R} v_{rT} = 1 \qquad (5.1)$$

$$\sum_{r=1}^{R} v_{rS} = \sum_{r=1}^{R} u_{Tr} = 0 \qquad (5.2)$$

Equation (5.1) ensures that one arc leaves S and one arc enters T. Equation (5.2) that no arc enters S and no arc leaves T.

$$\sum_{c=1}^{C} u_{cr} = \sum_{c=1}^{C} v_{rc} \qquad r=1,\ldots,R \qquad (5.3)$$

$$\sum_{r=1}^{R} v_{rc} = \sum_{r=1}^{R} u_{cr} \qquad c=1,\ldots,C \ c\neq S,T \qquad (5.4)$$

Equation (5.3) ensures that the number of arcs entering a reaction node is equal to the number leaving. Equation (5.4) fulfils the same condition for compound nodes.

$$\sum_{c=1}^{C} u_{cr} \leq 1 \qquad r=1,\ldots,R \qquad (5.5)$$

$$\sum_{r=1}^{R} v_{rc} \leq 1 \qquad c=1,\ldots,C\ c\neq S,T \qquad (5.6)$$

Equations (5.5) and (5.6) ensure that no reaction/compound node is revisited in the path.

We need constraints to prevent cycles appearing. Referring back to Figure 5.1, if we are seeking a path from C1 to C7 then Figure 5.4, where we do have a path from C1 to C7 but also a cycle R2→C5→R3→C6→R5→C3→R2, is a valid solution to the constraints presented so far above.

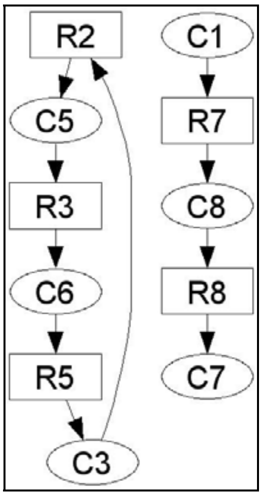

Figure 5.4: An example cycle

Note here on a technical issue that if we are seeking just the (k=1) shortest path then cycles will not appear as the distance metric (W_c) we use is non-negative. However because we intend to use our formulation to find k-shortest paths (for k≥2) cycles may appear.

Cycle elimination constraints are standard in the literature. As an illustration, for the cycle with six arcs (R2→C5, C5→R3, R3→C6, C6→R5, R5→C3, C3→R2) shown in Figure 5.4 the cycle elimination constraint is ($v_{R2,C5} + u_{C5,R3} + v_{R3,C6} + u_{C6,R5} + v_{R5,C3} + u_{C3,R2}$) ≤ 5. In general the constraint to eliminate a cycle is: (sum of the v_{rc} and u_{cr} variables for arcs appearing in the cycle) ≤ (number of arcs in the cycle - 1). Computationally cycle elimination constraints are added as and when cycles appear in solutions (identifying a cycle in a directed graph is algorithmically an easy task). Adding constraints to eliminate cycles as and when they appear is standard computational practice (since adding constraints to prevent any cycle at all appearing entails adding a very large number of constraints). After adding a cycle elimination constraint we resolve the problem, this process of adding cycle elimination constraints and resolving being repeated until no cycles exist in the solution.

Other authors dealing with path finding in metabolic pathways (e.g. McShan *et al.*, 2003; Croes *et al.*, 2005, 2006) constrain paths so as to exclude a reaction and its reverse. This is easily done within our integer programming approach. Let B be the set $\{(\alpha,\beta)|$ reaction node α and reaction node β are the reverse of each other, $\alpha<\beta\}$. If a reaction r is in the path then it must be true that $\sum_{c=1}^{C} u_{cr} = 1$. So to prevent reactions α and β from both being in the path we have:

$$\sum_{c=1}^{C} u_{c\alpha} + \sum_{c=1}^{C} u_{c\beta} \leq 1 \qquad \forall(\alpha,\beta)\in B \tag{5.7}$$

5.2.1.3 Objective

The objective function is to minimise the total connectivity of the compounds involved in the path, i.e.

$$\text{minimise } W_S + \sum_{c=1,c\neq S,T}^{C} W_c \sum_{r=1}^{R} (v_{rc} + u_{cr})/2 + W_T \tag{5.8}$$

where $\sum_{r=1}^{R} (v_{rc} + u_{cr})/2$ will have the value one if compound c is in the path, and the value zero if compound c is not in the path. Note that the objective function includes the connectivity of the source (W_S) and target compound (W_T). The choice of this objective was based on empirical evidence presented in Croes *et al.*, 2005, 2006, that the intermediate compounds in experimentally determined pathways present a low degree of connectivity. Although this can be explained on the basis of evolution, we think more research needs to be done to clarify this aspect of metabolic pathways.

5.2.2 Solution elimination constraints

In order to find the k-shortest path, we need to add further constraints to eliminate the (k-1)-shortest paths from the set of solutions. To illustrate this suppose we are interesting in finding the (k=2) second shortest path. Let U_{cr}^{1} and V_{rc}^{1} be the solution for the (k=1) shortest path. We need to eliminate this shortest path from the set of solutions. To do this we add the following constraint to our formulation:

$$\sum_{c=1}^{C}\sum_{r=1,U_{cr}^{1}=0}^{R} u_{cr} + \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{1}=0}^{R} v_{rc} + \sum_{c=1}^{C}\sum_{r=1,U_{cr}^{1}=1}^{R} (1\text{-}u_{cr}) + \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{1}=1}^{R} (1\text{-}v_{rc}) \geq 1 \qquad (5.9)$$

This constraint ensures that the Hamming distance between:

- the (k=1) shortest path solution, and
- the path solution found after this constraint is added and the problem resolved

is at least one, which means that there is a difference of at least one arc between the two solutions. Therefore, if we add this constraint to our formulation and resolve, we will find a new path solution, which is different from the previous solution (the 1-shortest path). Because we are minimising (c.f. equation (5.8)) this new path will be the "next best" path in objective function terms - so it must be the 2-shortest path.

In the general case, in order to find the k-shortest path, we have to include k-1 solution elimination constraints as below related to the (k-1)-shortest paths:

$$\sum_{c=1}^{C}\sum_{r=1,U_{cr}^{K}=0}^{R} u_{cr} + \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{K}=0}^{R} v_{rc} + \sum_{c=1}^{C}\sum_{r=1,U_{cr}^{K}=1}^{R} (1\text{-}u_{cr})$$

$$+ \sum_{c=1}^{C}\sum_{r=1,V_{rc}^{K}=1}^{R} (1\text{-}v_{rc}) \geq 1 \qquad K=1,\ldots,k\text{-}1 \qquad (5.10)$$

where U_{cr}^{K} and V_{rc}^{K} are the solution for the K-shortest path.

5.2.3 Overview

The formulation given above for finding the k-shortest metabolic path is an integer linear (zero-one, binary) program. Algorithmically such programs are solved by linear programming based tree search, which guarantees that the solution found will be optimal. Here, as in Chapter 3, we used Cplex.

Hence to summarise we have presented above a formulation for finding k-shortest paths to which standard software can be applied that:

- produces paths in increasing distance order; and
- guarantees that the paths found will be optimal.

5.3 Results

5.3.1 Introduction

We have used the metabolic network of *E.Coli* (the best studied organism in the biological world) presented by Reed *et al.*, 2003, which is available from http://systemsbiology.ucsd.edu/In_Silico_Organisms/E_coli/E_coli_reactions and comprises 880 cytosolic reactions and 613 compounds. A cytosolic reaction is one occurring in the cytosol, which essentially defines the medium where metabolism is carried out. A full list of reactions/compounds can be found in Appendices A and B.

The approach to finding k-shortest paths given above was applied to ten *E.Coli* experimentally determined pathways. The pathways used were taken from Keseler *et al.*, 2005; Nelson and Cox, 2005 and http://biocyc.org/ECOLI/. A detailed description of the experimentally determined pathways can be found in Appendix C.

Two different cases were considered: the reaction to reaction (R-R) case, where paths are computed from the first reaction to the last reaction in the pathway (such as was considered in Croes *et al.*, 2005, 2006); and the compound to compound (C-C) case, where paths are computed from source compound to target compound in the pathway.

In order to judge the effectiveness of our path finding approach we will compare each path found with a single metabolic path associated with each experimentally determined pathway. As noted above, there may be more than one metabolic path associated with a pathway. In this event we choose from amongst the possible metabolic paths just one path against which to compare our results. In order to address this issue, the associated metabolic path is defined here as the shortest path (under the distance metric as described above) that links the initial compound (reaction) and the final compound (reaction) of the pathway via balanced intermediate compounds. Details as to the associated metabolic paths can be found in Appendix C. To illustrate this issue, Figure 5.5 shows Gluconeogenesis pathway (Pathway 1). The source and target compounds (pyr and g6p respectively in this pathway) are coloured yellow. Compounds coloured blue are produced to excess; compounds coloured red are freely available; and compounds shown in white are balanced.

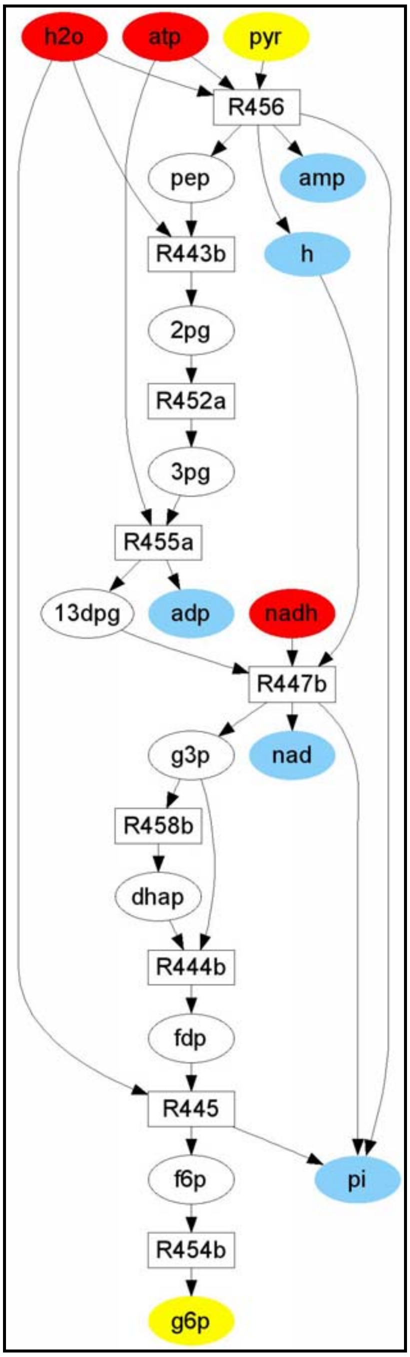

Figure 5.5: Gluconeogenesis pathway

With regard to the issue of multiple metabolic paths there are four metabolic paths, for the C-C case, in the above pathway. These are:

- pyr→R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p →R458b→dhap→R444b→fdp→R445→f6p→R454b→g6p

- pyr→R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p →R444b→fdp→R445→f6p→R454b→g6p

- pyr→R456→h→R447b→g3p→R458b→dhap→R444b→fdp→R445→f6p→ R454b→g6p

- pyr→R456→h→R447b→ g3p →R444b→fdp→R445→f6p→R454b→g6p

In order to associate a single metabolic path with this pathway we choose (from amongst these four paths in this particular instance) the shortest path (under the distance metric as described above) via balanced intermediate compounds. Balanced compounds are shown in white in the above pathway. When this is done we have the following metabolic path: pyr→R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p →R444b→fdp→R445→f6p→R454b→g6p.

For the R-R case there are also four metabolic paths in the above pathway. These are:

- R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p →R458b→dhap→R444b→fdp→R445→f6p→R454b

- R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p →R444b→fdp→R445→f6p→R454b

- R456→h→R447b→g3p→R458b→dhap→R444b→fdp→R445→f6p→ R454b

- R456→h→R447b→ g3p →R444b→fdp→R445→f6p→R454b

In order to associate a single metabolic path with this pathway we choose (from amongst these four paths in this particular instance) the shortest path (under the distance metric as described above) via balanced intermediate compounds. When this is done we have the following metabolic path:

R456→pep→R443b→2pg→R452a→3pg→R455a→13dpg→R447b→ g3p
→R444b→fdp→R445→f6p→R454b.

The computed paths were evaluated using the same criteria as in Croes *et al.*, 2005, 2006. These criteria, detailed below, essentially measure the degree of correspondence between any computed path and a path (the metabolic path) that represents the metabolic pathway.

In order to compare the computed path and the metabolic path, Croes *et al.*, 2005, 2006 defined the following correspondence values which indicate, numerically, correspondence between the computed path and the metabolic path:

- **True positives (TP):** The total number of reactions and compounds found in the computed path that are also in the metabolic path. The source and target nodes, whether reaction or compound, are not considered.
- **False positives (FP):** The total number of reactions and compounds found in the computed path that are not in the metabolic path.
- **False negatives (FN):** The total number of reactions and compounds found in the metabolic path that are not in the computed path.
- **Sensitivity (Sn):** = TP/ (TP + FN), is the fraction of the reactions and compounds in the metabolic path (excluding source and target) that are in the computed path.
- **Positive Predictive Value (PPV):** = TP/ (TP + FP), is the fraction of the reactions and compounds in the computed path (excluding source and target) that are in the metabolic path.
- **Accuracy (Ac):** = (Sn + PPV)/2, is the average of the previous two values.

Sensitivity, positive predictive value and accuracy are all defined such that higher values represent closer correspondence between the computed path and the metabolic path. If the computed path corresponds exactly to the metabolic path then Sn=PPV=Ac=1 (equivalently FP=FN=0).

As noted above, typically the effectiveness of any path finding approach is examined by seeing how well it performs (for example as evaluated by the above correspondence values) with respect to a known metabolic pathway. In other words given the source and target compound, and the entire metabolic network, how well does a particular path finding approach do at discovering the reactions and compounds involved in a known metabolic path or pathway?

Note here that we have adopted (as detailed above) the same correspondence values as defined in Croes *et al.*, 2005, 2006 but we should be clear that their approach is flawed. This is because they include compounds in their correspondence values. As mentioned in the introduction section we need only focus on reactions (since for each reaction we know the set of compounds involved). As many reactions involve more than one input/output compound the correspondence measures used by Croes *et al.*,2005, 2006 could classify a computed path as less than perfect even if it contains exactly the same set of reactions as the metabolic path (due to different compounds being involved in the metabolic path and the computed path). For example, referring to Figure 5.3, suppose the metabolic path is C1→R1→C3→R2→C5→R3→C7 (the left-hand path shown in Figure 5.3), but the computed path is C1→R1→C2→R2→C5→R3→C7 (the right-hand path shown in Figure 5.3). This will give TP=6, FP=1, FN=1, Sn=6/7, PPV=6/7, Ac=6/7. Yet both paths contain precisely the same reactions, which is the key feature. Clearly the correspondence values defined by Croes *et al.*, 2005, 2006 are inappropriate. Better correspondence measures would drop all mention of compounds in the values defined above. Despite this flaw we, for reasons of consistency of comparison with the results presented previously in Croes *et al.*, 2005, 2006, will present our results below using the correspondence measures that include compounds as defined above.

5.3.2 Results for an example known pathway - Gluconeogenesis

Table 5.1 shows the correspondence values for each of the k-shortest paths (k=1,2,…,10) computed from the initial reaction to final reaction (the R-R case) in the Gluconeogenesis pathway. For k=1, i.e. the shortest path, there is a low level of correspondence between the metabolic path and the computed shortest path. For k=2, i.e. the second shortest path, correspondence increases (sensitivity, positive predictive value and accuracy all increase). Note though that as we increase k we find different paths and so there is no guarantee that correspondence increases with increasing k. For k=4, for example, the correspondence values decrease – so the 4-shortest path corresponds less well to the metabolic path than the 3-shortest path. In fact the correspondence between the 4-shortest path and the metabolic path is less than for the (k=1) shortest path. It can be seen from Table 5.1 that for k=6 the solution is precisely the same as the Gluconeogenesis metabolic path.

k shortest path k	True positives (TP)	False positives (FP)	False negatives (FN)	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
1	5	2	8	0.385	0.714	0.549
2	10	1	3	0.769	0.909	0.839
3	10	1	3	0.769	0.909	0.839
4	3	4	10	0.231	0.429	0.330
5	3	4	10	0.231	0.429	0.330
6	13	0	0	1	1	1
7	2	13	11	0.154	0.133	0.144
8	2	13	11	0.154	0.133	0.144
9	2	13	11	0.154	0.133	0.144
10	10	3	3	0.769	0.769	0.769

Table 5.1: Correspondence values for the first ten shortest paths in the Gluconeogenesis pathway in the R-R case

Table 5.2 shows the correspondence values for each of the k-shortest paths (k=1,2,…,10) computed from the source compound to the target compound (the C-C case) in the Gluconeogenesis pathway. It can be seen that correspondence is markedly less than for the R-R case, and for no value of k examined is the computed k-shortest path the same as the Gluconeogenesis metabolic path.

k shortest path k	True positives (TP)	False positives (FP)	False negatives (FN)	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
1	2	15	13	0.133	0.118	0.125
2	7	2	8	0.467	0.778	0.622
3	0	21	15	0	0	0
4	0	27	15	0	0	0
5	2	23	13	0.133	0.080	0.107
6	2	21	13	0.133	0.087	0.110
7	0	18	15	0	0	0
8	0	23	15	0	0	0
9	2	23	13	0.133	0.080	0.107
10	0	20	15	0	0	0

Table 5.2: Correspondence values for the first ten shortest paths in the Gluconeogenesis pathway in the C-C case

Table 5.3 shows correspondence values for the best correspondence path (as measured by maximum accuracy) amongst all of the first k-shortest paths for a number of different values of k for the Gluconeogenesis pathway. For k=5 in the C-C case, for example, the best correspondence path out of the first five shortest paths has accuracy 0.622. Examining Table 5.2 we can see that this path was the second shortest path. Because here we take the maximum accuracy path from amongst the first k-shortest paths correspondence increases as we increase k.

Case	k	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
R-R	**1**	0.385	0.714	0.549
	5	0.769	0.909	0.839
	10	1	1	1
C-C	**1**	0.133	0.118	0.125
	5	0.467	0.778	0.622
	10	0.467	0.778	0.622

Table 5.3: Values for the best correspondence path among the first k-shortest paths for k=1,5,10 in the Gluconeogenesis pathway

5.3.3 Results for ten known pathways – correspondence values

In our analysis we have examined the ten *E.Coli* experimentally determined pathways (including Gluconeogenesis) shown in Table 5.4. One complication arose with pathways 6 (Pentose phosphate) and 10 (Arginine biosynthesis) in the R-R case in that the definition of the first or last reaction turned out to be ambiguous, there being two different options for the first reaction (pathway 10) or for the last reaction (pathway 6), as can be seen in Appendix C. Consequently, we computed two different metabolic paths in the R-R case for each pathway.

As far as the C-C case is concerned, one minor issue relates to pathway 8, the TCA cycle. In this pathway the source compound and the target compound are the same. The usual definition of a path is that the initial and final nodes are different (whereas in a cycle the initial and final nodes are the same). Hence in order to deal with this pathway we treated the source/target compound as two different compounds, one relating to being used as input to a reaction, the other relating to being used as output from a reaction.

Pathway number	Pathway name
1	Gluconeogenesis
2	Glycogen
3	Glycolysis
4	Proline biosynthesis
5	Ketogluconate metabolism
6a	Pentose phosphate
6b	Pentose phosphate
7	Salvage pathway deoxythymidine phosphate
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle
9	NAD biosynthesis
10a	Arginine biosynthesis
10b	Arginine biosynthesis

Table 5.4: Experimentally determined pathways examined

Detailed results for all of the pathways shown in Table 5.4 can be found in Appendix C.

Table 5.5 shows the same information as Table 5.3, but averaged over all the pathways considered. In Table 5.5 we, for example, have that the average accuracy for the shortest path (k=1) is 0.830 in the R-R case, but only 0.449 in the C-C case. If we take, for each of the pathways examined, the maximum accuracy (best correspondence) path over the first five shortest paths this accuracy increases to 0.966 in the R-R case and 0.818 in the C-C case.

Case	k	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
R-R	**1**	0.813	0.847	0.830
	5	0.948	0.983	0.966
	10	0.968	0.991	0.979
C-C	**1**	0.400	0.499	0.449
	5	0.755	0.882	0.818
	10	0.755	0.882	0.818

Table 5.5: Values for the best correspondence path among the first k-shortest paths for k=1,5,10 averaged over all pathways

Table 5.5 indicates high correspondence for the R-R case. These results are in accordance with the results presented in Croes *et al.*, 2005, 2006. However, one of the deficiencies of the Croes *et al.*, 2005, 2006 work is that no results are presented for the C-C case. This is especially important as, in the literature, metabolic pathways are typically viewed as relating to transforming one compound into another – not as relating to going from one reaction to another. It is clear from Table 5.5 that correspondence is poor for the C-C case. Even taking the first ten shortest paths correspondence (average maximum accuracy) is only 0.818 – less that the correspondence achieved for the shortest path (k=1) in the R-R case.

Poor maximum accuracy in the C-C case is especially found in those metabolic paths where the number of intermediate reactions/compounds between the source compound and the target compound is high. Figure 5.6 plots, for each pathway in the C-C case, the maximum accuracy over all ten shortest paths against the number of intermediate reactions/compounds in the metabolic path. As can be seen from Figure 5.6, maximum accuracy declines as the metabolic path involves more intermediate reactions/compounds. One further point to be made from Table 5.5 is that the results are not significantly improved in either the R-R or C-C cases by moving from the first five shortest paths to the first ten shortest paths. In other words the computation of more shortest paths beyond the first five is of little (average) benefit.

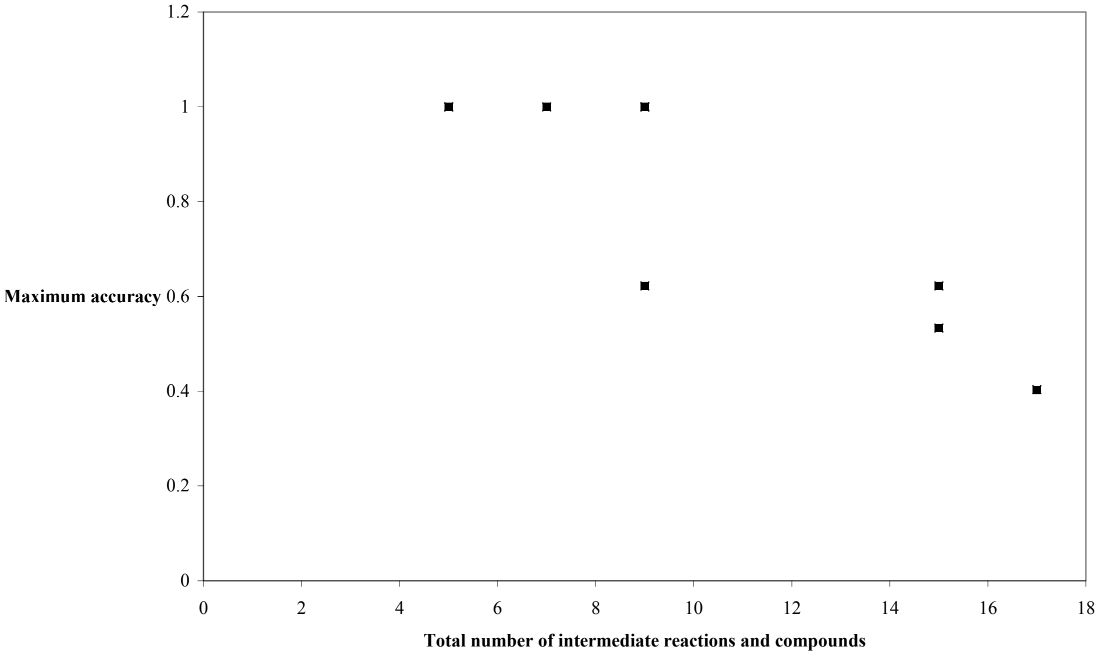

Figure 5.6: Maximum accuracy over all ten k-shortest paths for each pathway as against number of intermediate reactions/compounds for the C-C case

5.3.3 Results for ten known pathways – metabolic path recovery

Whilst Table 5.5 gives an insight into accuracy we believe that it is appropriate to also tabulate whether, or not, a computed shortest path corresponds *exactly* to the metabolic path (which we term "recovering" the path). Clearly recovering a metabolic path is the ideal case (and corresponds to an accuracy (Ac) of one). Table 5.6 indicates for the R-R case whether, or not, we recover the metabolic path amongst the first k-shortest paths for k=1,5,10.

Pathway number	Pathway name	Metabolic path recovered?		
		k 1	k 5	k 10
1	Gluconeogenesis	no	no	yes
2	Glycogen	yes	yes	yes
3	Glycolysis	no	yes	yes
4	Proline biosynthesis	yes	yes	yes
5	Ketogluconate metabolism	yes	yes	yes
6a	Pentose phosphate	yes	yes	yes
6b	Pentose phosphate	yes	yes	yes
7	Salvage pathway deoxythymidine phosphate	no	yes	yes
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	no	no	no
9	NAD biosynthesis	yes	yes	yes
10a	Arginine biosynthesis	yes	yes	yes
10b	Arginine biosynthesis	yes	yes	yes
Number of “yes” entries (maximum 12)		8	10	11

Table 5.6: Metabolic path recovery amongst the first k-shortest paths for k=1,5,10 for the R-R case

Pathway number	Pathway name	Metabolic path recovered?			BP model	
					Metabolic pathway recovered?	
		k 1	k 5	k 10	Objective (3.13)	Objective (3.14)
1	Gluconeogenesis	no	no	no	yes	no
2	Glycogen	yes	yes	yes	yes	no
3	Glycolysis	no	no	no	yes	yes
4	Proline biosynthesis	no	no	no	yes	no
5	Ketogluconate metabolism	yes	yes	yes	yes	no
6	Pentose phosphate	no	yes	yes	yes	no
7	Salvage pathway deoxythymidine phosphate	no	yes	yes	yes	no
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	no	no	no	no	yes
9	NAD biosynthesis	no	yes	yes	yes	no
10	Arginine biosynthesis	yes	yes	yes	yes	no
Number of “yes” entries (maximum 10)		3	6	6	9	2

Table 5.7: Metabolic path recovery amongst the first k-shortest paths for k=1,5,10 for the C-C case and comparison with the BP model

Table 5.7 presents the same information as Table 5.6 but for the C-C case. It also shows the results for the BP model for these ten *E.Coli* experimentally determined metabolic pathways.

Table 5.6 indicates that for the R-R case the (k=1) shortest path recovers the metabolic path in 8 out of 12 pathways – this figure rising to recovering 11 of the 12 paths if we consider k=10. On the other hand, Table 5.7 indicates that for the C-C case the (k=1) shortest path recovers the metabolic path in only 3 out of 10 pathways – this figure rising to recovering 6 of the 10 paths if we consider k=10.

Comparing the results shown in Table 5.7 (k=10) with the results for the BP model (taking the best of both objectives), we have a mix situations: some whether both approaches achieve recovery (e.g. pathway 2); some whether both approaches do not achieve recovery (e.g. pathway 8); and some where the path finding approach presented here does not achieve recovery and the BP model does (e.g. pathway 1). It appears clear that BP model produces more accurate results. Note here however that path finding approaches do not need any prior pathway knowledge.

5.3.4 Discussion

In essence the path finding approach to metabolic pathways given above rests on the hypothesis that insight into a metabolic pathway can be obtained by finding k-shortest paths using compound node connectivities as a distance metric. Our results partially support this hypothesis.

It is clear that for the R-R case, where our results are in accordance with the results presented previously in Croes *et al.*, 2005, 2006, this hypothesis is valid. We have high correspondence values and recover the metabolic path for nearly all pathways examined. However for the C-C case the validity of the hypothesis is more questionable. Correspondence values are not as good as for the R-R case and we recover far fewer metabolic paths.

Clearly the lack of success for the C-C case, as opposed to the R-R case, could be due to a number of factors. It may be that k-shortest paths (whatever the distance metric used) is not an appropriate concept for analysing metabolic pathways. It is clear that the literature is divided as to whether, or not, making use of shortest paths is of value with respect to metabolic pathways. Stoichometric approaches (Schilling *et al.*, 2000; Schuster *et al.*, 2000) do not use shortest paths. However if utilising shortest paths was inappropriate then we would not have expected the results for the R-R case to be any better than the results for the C-C case. But in fact we find that the results for the R-R case are better than the results for the C-C case.

It could be, of course, that k-shortest paths are an appropriate concept for analysing metabolic pathways, but we have used an inappropriate distance metric. The distance metric used in this paper related to the connectivity of the compounds involved in the path. Other possibilities (obviously) exist. For example we might use a distance metric based on reactions. Such a metric, for example, could be related to the amount of chemical change that takes place at each reaction, or to energetic considerations such as the Gibbs free energy for each reaction. Alternatively a distance metric that takes both compounds and reactions into account may be appropriate.

Finally we would note here that the key difference between the C-C case and the R-R case relates to the fact that in the C-C case we have to choose two extra reactions in the path: one reaction having the source compound as an input compound, the other reaction having the target compound as an output compound. In the R-R case these two reactions are specified. This might imply that taking the current distance metric (which is compound based, but which is successful for the R-R case), and amending it for the C-C case with reaction terms that relate only to any reactions having the source compound as an input compound, or having the target compound as an output compound, might be a profitable approach.

5.4 Conclusions

In this chapter, the effectiveness of using compound node connectivities in a path finding approach to metabolic pathways has been examined. We found that finding k-shortest paths using a distance metric based on compound node connectivities performed well when a metabolic path was regarded as being from a source reaction to a target reaction. The same approach performed less well when a metabolic path was regarded as being from a source compound to a target compound. An approach to path finding based upon integer programming was also presented that produces k-shortest paths in increasing distance order and guarantees that the paths found will be optimal.

Chapter 6

The Improved Beasley-Planes model

In this chapter we present the Improved Beasley-Planes (IBP) model for recovering experimentally determined pathways. Using path finding approaches as a building block, the IBP model also incorporates elements from the BP model. Although the IBP model results are slightly worse than the BP model results, the IBP model overcomes the issue of having to know which are the low presence unbalanced compounds, which represents the major limitation of the BP model.

6.1 Introduction

In Chapter 5 we showed that path finding approaches present a significant advantage with respect to the BP model as no prior pathway knowledge is needed. The main assumption behind path finding approaches is that metabolic pathways do contain at least one directed path from the source compound to the target compound. Clearly this assumption applies for any metabolic pathway, whether known or unknown. The key decision relates to the choice of a proper distance metric so as to provide biological significance to the sought k-shortest paths. In Chapter 5 we examined a distance metric based on total connectivity, as initially proposed by Croes *et al*., 2005, 2006. Though the results presented do not show complete success for recovering experimentally determined metabolic pathways, we consider that path finding approaches can be utilised as a building block for constructing a more refined approach.

In this chapter we present the Improved Beasley-Planes (IBP) model for recovering experimentally determined pathways. As in path finding approaches, the IBP model starts from the idea that metabolic pathways do contain at least one directed (metabolic) path from the source compound to the target compound. However the IBP model also addresses pathway stoichiometry, namely including stoichiometric constraints, as previously done in the BP model. In essence the IBP model views a metabolic pathway as a finite set of metabolic paths from the source compound to the target compound that satisfy different (logical) stoichiometric constraints. Interestingly, the IBP model provides a link between path finding approaches (Croes *et al*., 2005, 2006) and stoichiometric approaches (Schilling *et al.*, 2000; Schuster *et al.*, 2000). To the best of our knowledge, no approach in the literature has to date combined both types of approaches. This fact constitutes a significant advance in the field from the modelling point of view.

To illustrate the IBP model, given our example metabolic network in Figure 6.1, let us assume that Figure 6.2 shows the experimentally determined pathway that converts C1 into C7. This pathway contains two different metabolic paths, namely C1→R1→C3→R2→C5→R3→C7 and C1→R1→C2→R2→C5→R3→C7. Both paths provide the precise set of reactions involved in the pathway. However, they do not give information as to pathway stochiometry, e.g. balanced compounds, the number of ticks of a particular reaction, number of source compound molecules consumed, etc. The IBP model introduces stoichiometric constraints in the paths so as to recover pathway stoichiometry.

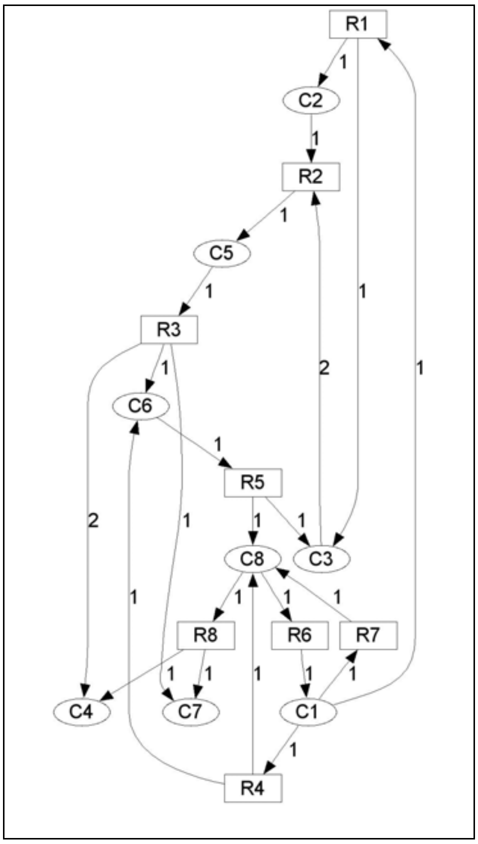

Figure 6.1: Previous Figure 1.2 in Chapter 1

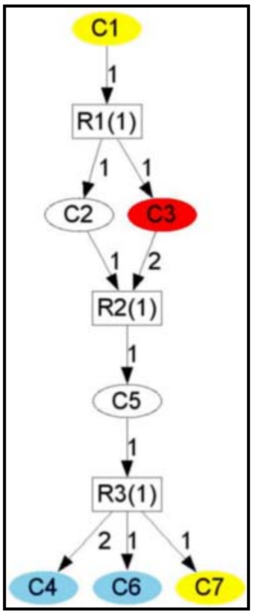

Figure 6.2: A possible pathway converting C1 into C7

In addition, the IBP model also considers the issue of branched metabolic pathways. Figure 6.3 shows an example branched metabolic pathway. In order to recover the complete set of reactions involved in the pathway, we need at least two metabolic paths. These might be: C1→R1→C2→R2→C5→R3→C7 and C1→R7→C8→R8→C7. This contrasts with Figure 6.2 where a single metabolic path contained the complete set of reactions in the pathway. Accordingly, a single path representation (as is usually done in path finding approaches) turns out to be limited for branched metabolic pathways. It is for that reason that we extend the IBP model into a set of metabolic paths.

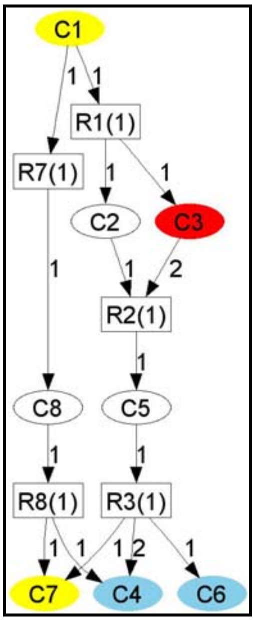

Figure 6.3: A possible branched metabolic pathway converting C1 into C7

In order to provide biological significance to the metabolic paths, the IBP model introduces additional constraints related to high presence compounds, inorganic compounds and cofactors, which do not typically appear as intermediate compounds in experimentally determined metabolic pathways. These compound sets are properly defined in the mathematical model section below. Moreover, the IBP model discusses two novel objectives functions. Whilst one objective is similar to objective (3.13) of the BP model presented in Chapter 3, the other objective is related to objective (5.8) of the path finding approach presented in Chapter 5. We also show the relationship between the objectives.

We present below the Improved Beasley-Planes optimisation model for recovering experimentally determined metabolic pathways.

6.2 Mathematical model

In the IBP model we have a metabolic network of R reactions (where each reaction has a specified direction so a reversible reaction contributes two different reactions to the total number R) which collectively involve C different compounds. Suppose we are seeking a pathway that transforms Q_S molecules of source compound S into Q_T molecules of target compound T and contains a maximum number of K metabolic paths from the source compound to the target compound.

Technically the IBP model is an integer linear program. We first describe the variables and constraints related to the pathway stoichiometry. Secondly, we describe the variables and constraints related to the directed metabolic paths. Subsequently the linking constraints between pathway stoichiometry and the metabolic paths are presented. In addition, we present constraints related to compound sets, i.e. high presence compounds, inorganic compounds and cofactors. Finally the objective function of the IBP model is described.

As the reader will note some of this material below (e.g. variables, constraints) echoes that seen in Chapter 3 and Chapter 5. However for ease of understanding of the IBP model we have repeated that material in this chapter.

6.2.1 Variables and constraints related to metabolic pathway stoichiometry

A metabolic pathway is a set of enzyme-catalysed biochemical reactions that transforms Q_S molecules of source compound S into Q_T molecules of target compound T. Figure 6.2 shows a possible metabolic pathway that converts one molecule of C1 into one molecule of C7.

Thus, a reaction may, or may not, be active in the pathway. We have the following binary (zero-one) variable:

$z_r = 1$ if reaction r is active in the pathway, 0 otherwise ($r=1,\ldots,R$)

and the associated tick variable:

t_r the number of ticks of reaction r in the pathway (this must be an integer variable (≥ 0) with value 0 if the reaction not active)

We need a constraint relating the number of ticks of a reaction to the zero-one variable signifying whether the reaction is active or not, this is:

$$t_r \leq M_1 z_r \qquad r=1,\ldots,R \qquad (6.1)$$

where M_1 is a large positive constant that represents the maximum number of ticks of any reaction. If the reaction does not tick then it must be inactive, so we have the constraint:

$$z_r \leq t_r \qquad r=1,\ldots,R \qquad (6.2)$$

As in the BP model, the IBP model involves variables relating to whether compounds are balanced (or not). A balanced compound is one where the number of molecules needed (consumed) is equal to the number produced. A compound which is balanced can either be active (number of molecules needed = number produced > 0) or inactive (number of molecules needed = number produced = 0) in the pathway. Considering Figure 6.2, for example, the active balanced compounds are C2 and C5.

Let n_{cr} be the number of molecules of compound c needed as input for one tick of reaction r and p_{cr} be the number of molecules of compound c produced as output by one tick of reaction r. For each compound c (c=1,…,C) define:

- $b_c=1$ if for compound c the number of molecules needed is equal to the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r = \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If $b_c=1$ compound c is ***balanced***.

- $e_c=1$ if for compound c the number of molecules needed is less than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r < \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If $e_c=1$ compound c is ***produced to excess***, since we have "spare" molecules of the compound to be disposed of (in other pathways).

- $f_c=1$ if for compound c the number of molecules needed is greater than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r > \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If $f_c=1$ compound c must be ***freely available***, since we need “spare” molecules of the compound that have come from other pathways.

Considering Figure 6.2, for example, compound C4 is produced to excess (denoted by the blue colouring) and compound C3 is freely available (denoted by the red colouring).

We have the constraint:

$$b_c + e_c + f_c = 1 \qquad c=1,...,C \qquad (6.3)$$

In order to link the variables e_c and f_c to the number of molecules of each compound produced we need the constraints.

$$e_c \geq (\sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r)/M_2 \qquad c=1,\dots,C \qquad (6.4)$$

$$e_c \leq 1 + (\sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r -1)/ M_2 \qquad c=1,\dots,C \qquad (6.5)$$

$$f_c \geq (\sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r)/M_2 \qquad c=1,\dots,C \qquad (6.6)$$

$$f_c \leq 1 + (\sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r -1)/M_2 \qquad c=1,\dots,C \qquad (6.7)$$

where M_2 is a large positive constant.

We need constraints specifying that the required number of molecules of the source compound S (Q_S) and target compound T (Q_T) are involved. These are:

$$\sum_{r=1}^{R} n_{Sr}t_r = Q_S \quad \text{and} \quad \sum_{r=1}^{R} p_{Tr}t_r = Q_T \qquad (6.8)$$

If the source compound and target compound are different then we produce none of the source compound and consume none of the target compound, i.e.

$$\sum_{r=1}^{R} p_{Sr}t_r = \sum_{r=1}^{R} n_{Tr}t_r = 0 \qquad \text{if } S \neq T \qquad (6.9)$$

6.2.2 Variables and constraints related to metabolic paths

We define a metabolic path to be a directed path from the source compound S to the target compound T. Figure 6.4 shows an example metabolic path contained in the example metabolic pathway shown in Figure 6.2 (repeated below in Figure 6.4 for convenience). Note the difference between a metabolic pathway and a metabolic path. The metabolic pathway contains all the reactions and compounds involved in the pathway, whilst the metabolic path typically contains only a subset of intermediate reactions/compounds.

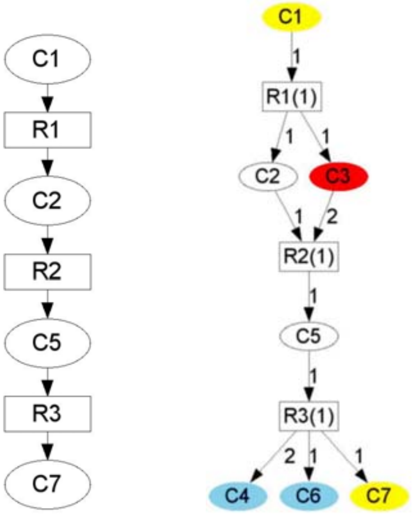

Figure 6.4: A metabolic path from C1 to C7 in the pathway shown in Figure 6.2

Let K be the maximum number of directed metabolic paths from the source compound to the target compound. We need to decide the arcs involved in the K metabolic paths, so our zero-one (binary, integer) variables are:

- u_{crk} = 1 if the arc from compound node c to reaction node r is in metabolic path k; 0 otherwise (c=1,…,C; r=1,…,R; k=1,…,K)

- v_{rck} = 1 if the arc from reaction node r to compound node c is in metabolic path k; 0 otherwise (r=1,…,R; c=1,…,C; k=1,…,K)

Let m_{cr} have the value 1 if compound c is an input compound for reaction r, 0 otherwise. If m_{cr}=0, i.e. the arc does not exist, then we fix $u_{crk}\forall k$ to 0. Let d_{rc} have the value

1 if compound c is an output compound from reaction r, 0 otherwise. Similarly if d_{rc}=0, then we fix v_{rck} $\forall$k to 0. So u_{crk}= 0 $\forall$k if m_{cr} =0 r=1,…,R; c=1,…,C and v_{rck}= 0 $\forall$k if d_{rc} =0 r=1,…,R; c=1,…,C.

Note here that allowing a maximum of K metabolic paths does not imply that the solution will contain K different paths. It is possible for the solution to contain K copies of exactly the same metabolic path. Rather the formulation given allows distinct metabolic paths to exist if other constraints in the problem are best satisfied by having multiple distinct metabolic paths.

The constraints related to the K metabolic paths are:

$$\sum_{r=1}^{R} m_{Sr}u_{Srk} = 1 \text{ and } \sum_{r=1}^{R} d_{rT}v_{rTk} = 1 \qquad k=1,\ldots,K \qquad (6.10)$$

$$\sum_{r=1}^{R} d_{rS}v_{rSk} = 0 \text{ and } \sum_{r=1}^{R} m_{Tr}u_{Trk} = 0 \qquad k=1,\ldots,K,\ S\neq T \qquad (6.11)$$

Equation (6.10) here ensures that one arc leaves S and one arc enters T for each of the K metabolic paths. Equation (6.11) here ensures that no arc enters S and no arc leaves T for each of the K metabolic paths.

$$\sum_{c=1}^{C} m_{cr}u_{crk} = \sum_{c=1}^{C} d_{rc}v_{rck} \qquad r=1,\ldots,R;\ k=1,\ldots,K \qquad (6.12)$$

$$\sum_{r=1}^{R} d_{rc}v_{rck} = \sum_{r=1}^{R} m_{cr}u_{crk} \qquad c=1,\ldots,C\ c\neq S,T;\ k=1,\ldots,K \qquad (6.13)$$

Equation (6.12) ensures that the number of arcs associated with metabolic path k entering a reaction node is equal to the number leaving. Equation (6.13) fulfils the same condition for compound nodes.

$$\sum_{c=1}^{C} m_{cr}u_{crk} \leq 1 \qquad r=1,\ldots,R;\ k=1,\ldots,K \qquad (6.14)$$

$$\sum_{r=1}^{R} d_{rc}v_{rck} \leq 1 \qquad c=1,\ldots,C\ c\neq S,T;\ k=1,\ldots,K \qquad (6.15)$$

Equation (6.14) and (6.15) ensure that no reaction/compound node is revisited in metabolic path k.

As in Chapter 5, we need to prevent cycles appearing in the solution. Consider metabolic path k once we have solved the IBP model (as considered so far above). We may have a cycle for the non-zero variables (u_{crk},v_{rck}) associated with this path. If S ≠T a cycle defines a path of successive arcs associated with non-zero variables (u_{crk},v_{rck}) in the metabolic network that starts and ends at the same compound, whilst if S=T a cycle defines a path of successive arcs associated with non-zero variables (u_{crk},v_{rck}) in the directed network that starts and ends at the same compound and does not contain compound S. Figure 6.5 shows an example cycle in the case S ≠T.

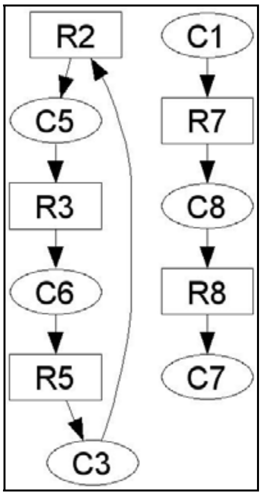

Figure 6.5: An example cycle

The constraints to eliminate a cycle are: (sum of the v_{rck} and u_{crk} variables for arcs appearing in the cycle) ≤ (number of arcs in the cycle - 1) k=1,…,K. Note that this cycle elimination constraint applies for all metabolic paths, irrespective of the metabolic path in which it was discovered. Technically this is because we are allowing K paths. If we detect a cycle in a Cplex solution associated with a specific metabolic path then unless we eliminate that cycle in all K paths simultaneously we may well find computationally that we use Cplex K times, each time discovering what is essentially the same cycle, but with a different metabolic path, k, label. As described in Chapter 5, in our computational implementation cycle elimination constraints are added as and when cycles appear in solutions (identifying a cycle in a directed graph is algorithmically an easy task).

As a final comment, we did not explore here the incorporation of constraints relating to "labelling" of the K metabolic paths (so as to uniquely define the first path, the second path, etc). This possibility might be a fruitful one to attempt, since these constraints might reduce computational time and eliminate equivalent optimal solutions.

6.2.3 Linking constraints between metabolic pathway stoichiometry and the K metabolic paths

We need to link the variables related to the metabolic pathway stoichiometry (z_r, t_r, b_c) and the variables related to the metabolic paths (u_{crk} and v_{rck}).

Firstly we need to relate the appearance of a reaction node in a metabolic path to the variables z_r signifying whether or not a reaction is present in the pathway. This is done by:

$$z_r \geq u_{crk} \qquad m_{cr}=1;\ r=1,\ldots,R;\ c=1,\ldots,C;\ k=1,\ldots K \qquad (6.15)$$

$$z_r \geq v_{rck} \qquad d_{rc}=1;\ r=1,\ldots,R;\ c=1,\ldots,C;\ k=1,\ldots K \qquad (6.16)$$

These constraints ensure that if an arc is in any of the K metabolic paths then z_r is forced to be one. If the arcs are not used then these constraints are inactive. Similarly if z_r is zero then these constraints ensure that no arc associated with reaction r can be used in any metabolic path.

In addition, we can reasonably impose the constraint that if a reaction is present in the pathway, then it must lie on one of the K metabolic paths. This constraint is true for all the pathways examined if we have a sufficiently high K value. To some extent this constraint addresses an issue of scope, i.e. reactions that are not on a metabolic path are outside the scope of our pathway. This is supported by the recent work of Ihmels *et al.*, 2004, which suggests that metabolic flow is driven by linear pathways. This constraint is:

$$\sum_{k=1}^{K}\sum_{c=1}^{C} m_{cr}u_{crk} \geq z_r \qquad r=1,\ldots,R \qquad (6.17)$$

which ensures that if a reaction r is active then we have at least one arc coming into that reaction associated with one of the K metabolic paths. Note that this constraint allows a reaction to be on more than one metabolic path.

It may be that for S and T there is a single reaction that has S as an input compound and T as an output compound. If such cases we may, perhaps, find a pathway that comprises just this reaction. If such a reaction exists then we exclude it from ever appearing in the pathway, i.e.

$$z_r = 0 \qquad \text{if } n_{Sr} \geq 1 \text{ and } p_{Tr} \geq 1 \;\; r=1,\ldots,R \qquad (6.18)$$

With respect to the compound nodes in the paths, we impose the constraint that if an intermediate compound c is on a metabolic path then the compound must be balanced. This constraint is:

$$b_c \geq u_{crk} \qquad r=1,\ldots,R;\; c=1,\ldots,C;\; c\neq S,T;\; k=1,\ldots K \qquad (6.19)$$

$$b_c \geq v_{rck} \qquad r=1,\ldots,R;\; c=1,\ldots,C;\; c\neq S,T;\; k=1,\ldots K \qquad (6.20)$$

which ensure that if an arc associated with an intermediate ($c\neq S,T$) compound is used in any of the K metabolic paths then the compound must be balanced. If the arcs are not used then these constraints are inactive. Similarly if b_c is zero then these constraints ensure that the arc associated with compound c cannot be used in any metabolic path.

Note here equations (6.19) and (6.20) are much less restrictive than equation (3.10) of the BP model or even the PSS condition in elementary flux modes (Schuster *et al.*, 2000) and extreme pathways (Schilling *et al.*, 2000). Indeed equation (3.10) and PSS condition define a set of compounds (low presence compounds and internal compounds respectively) that must necessarily be balanced in every metabolic pathway. This clearly does not apply for every metabolic pathway and the problem of unbalanced low presence compounds arises. Here equations (6.19) and (6.20) exclusively state that intermediate compounds on a metabolic path must be balanced, but do not explicitly state which compounds must be balanced. This potentially allows the same compound to be balanced in some pathways, whilst being consumed (or produced) in other pathways.

6.2.4 Compound set constraints

As in the BP model, we have found it necessary to distinguish between compounds that appear in a significant number of different reactions and compounds that appear in just a few reactions. The percentage presence δ_c of a compound c was previously defined as δ_c = 100(number of reactions in which c appears)/R = $100\sum_{r=1}^{R}\min(\max(p_{cr},n_{cr}),1)/R$. Compounds for which $\delta_c \leq \Delta$ (where Δ is an input parameter) we call low presence compounds. Compounds for which $\delta_c > \Delta$ we call high presence compounds. The set of high presence compounds is formally defined here as $D_1 = [c \mid \delta_c>\Delta, c=1\ldots,C]$. Table 6.1 shows the list of compounds belonging to D_1 with Δ=4%, the value used in the computational results reported later.

Compound	Percentage presence
Hydrogen ion	43.86
Water	28.98
Adenosine triphosphate	18.98
Adenosine diphosphate	14.89
Phosphate	14.32
Nicotinamide adenine dinucleotide	9.77
Nicotinamide adenine dinucleotide – reduced	9.32
Diphosphate	8.98
Nicotinamide adenine dinucleotide phosphate	7.16
Carbon dioxide	7.05
Nicotinamide adenine dinucleotide phosphate – reduced	6.93
L-Glutamate	5.91
Coenzyme A	5.23
Pyruvate	4.77
Ammonium	4.43
Adenosine monophosphate	4.43

Table 6.1: List of high presence compounds

We denote D_2 as the set of all (inorganic) compounds which do not involve carbon in their molecular makeup. D_2 is easily identified as we know the chemical formula for each compound. Hence formally D_2 = [c | compound c does not include carbon in its molecular makeup, c=1…,C]. Note that this subset does include some compounds which are high presence. For example water (H_2O) is a high presence compound, and it does not include carbon in its molecular makeup. Mathematically this means that we will have $D_1 \cap D_2 \neq \emptyset$. The complete list of inorganic compounds is shown in Table 6.2.

Iron ion	Fe
Hydrogen ion	H
Hydrogen	H2
Water	H2O
Hydrogen peroxide	H2O2
Hydrogen sulfide	H2S
Ammonium	NH4
Nitrite	NO2
Nitrate	NO3
Oxygen	O2
Superoxide anion	O2
Phosphate	HO4P
Diphosphate	HO7P2
Inorganic triphosphate	HO10P3
Selenide	HSe
Selenophosphate	H2O3PSe
Sulfite	O3S
Sulfate	O4S
Thiosulfate	O3S2

Table 6.2: List of inorganic compounds

We impose the constraint that if a compound is either in the set of high presence compounds D_1, or in the set of inorganic compounds D_2, then it cannot be on the metabolic path through a reaction provided it is possible for that reaction to have other compounds on such a path.

For a reaction r other compounds associated with r exist and can be on a metabolic path into r if and only if there exists an input compound c for reaction r (i.e. $n_{cr} \geq 1$) with $c \notin D_1 \cup D_2 - [S]$. For each reaction r that satisfies this condition we impose the constraint that no compound in $D_1 \cup D_2$ (but excluding source) can be on a metabolic path into r, i.e.

$$u_{crk}=0 \qquad \forall c \in D_1 \cup D_2\text{-}[S];\ k=1,\ldots K \qquad (6.21)$$

For a reaction r other compounds associated with r exist and can be on a metabolic path out of r if and only if there exists an output compound c for reaction r (i.e. $p_{cr} \geq 1$) with $c \notin D_1 \cup D_2$-[T]. For each reaction r that satisfies this condition we impose the constraint that no compound in $D_1 \cup D_2$ (but excluding target) can be on a metabolic path out of r, i.e.

$$v_{rck}=0 \qquad \forall c \in D_1 \cup D_2\text{-}[T];\ k=1,\ldots K \qquad (6.22)$$

In the IBP model we account for cofactors. Although the word cofactor is commonly used in the literature there appears to be no clear definition, certainly no definition based on numeric criteria. As described in Chapter 2, essentially two (organic) compounds are said to be cofactors if they commonly appear together, one as an input compound, the other as an output compound, in reactions and have similar chemical makeups. For the purposes of adopting a numeric definition of cofactors we conducted a graphical analysis.

We define the frequency $\Omega_{\alpha\beta}$ of a given pair of compounds (α,β) to be the frequency with which the pair of compounds (α,β) appear together on opposite sides of the same reaction (so $\Omega_{\alpha\beta} = \Omega_{\beta\alpha}$). Becker *et al.* (2006) used this parameter to find significant pairs, i.e. those with high frequency values. However, when trying to find cofactors, this parameter is not completely appropriate as $\Omega_{\alpha\beta}$ might lead to compounds with high connectivity (W_c), which are not necessarily cofactors. Recalling Chapter 5, the connectivity W_c of compound c is the number of reactions in which the compound appears in the metabolic network. For example, the pair Nadh-Coa (Nicotinamide adenine dinucleotide – reduced and Coenzyme A) has a frequency value of 8. In terms of frequency values this pair is significant as only eleven pairs have a higher frequency value. However this pair does not appear in the literature as cofactors. The reason we have a significant frequency here is that both compounds have a high total connectivity (82 for Nadh and 46 for Coa).

In order to avoid these misleading results, we introduce the relative frequency $\Omega^{\alpha\beta}$ for the compound α (and $\Omega^{\beta\alpha}$ for the compound β), formally $\Omega^{\alpha\beta}=\Omega_{\alpha\beta}/W_\alpha$ (and $\Omega^{\beta\alpha}=\Omega_{\alpha\beta}/W_\beta$

for compound β). Note here that unlike $\Omega_{\alpha\beta}$ ($=\Omega_{\beta\alpha}$) this is not symmetric, i.e. $\Omega^{\alpha\beta}\neq\Omega^{\beta\alpha}$ in general. The relative frequency takes values between 0 and 1. Clearly high values of the relative frequency might indicate a possible cofactor. In relation to the example pair described above, the relative frequency of Nadh with respect to Nadh-Coa pair is 0.1 and the relative frequency of Coa with respect to Nadh-Coa pair is 0.17, indicating that the pair Nadh-Coa is not significant. As we have two different values of relative frequency for a given pair (α,β) (precisely one for compound α and one for compound β), we used the average relative frequency to evaluate the significance of a pair, $(\Omega^{\alpha\beta} + \Omega^{\beta\alpha})/2$. In the example the average relative frequency for Nadh-Coa is 0.135, which is far from 1. Note here that we neglected inorganic compounds in the analysis, as usually cofactors include carbon in their molecular formula.

Figure 6.6 plots, for each particular pair (α,β), the average relative frequency of the pair against the frequency of the pair ($\Omega_{\alpha\beta}$). We chose as cofactors those pairs circled in red since they are far from the main set of points shown in Figure 6.6. Note that we did not include pairs with high average relative frequency but low frequency, e.g. $(\Omega^{\alpha\beta} + \Omega^{\beta\alpha})/2=1$ and $\Omega_{\alpha\beta}=2$. The reason is that cofactors usually have a significant frequency value $\Omega_{\alpha\beta}$. Similarly, pairs with low average relative frequency but high frequency, e.g. $(\Omega^{\alpha\beta} + \Omega^{\beta\alpha})/2=0.4$ and $\Omega_{\alpha\beta}=25$, are not considered, since cofactors must have a significant average relative frequency. The set of cofactors is listed in Table 6.3. These results agree with biochemistry textbooks (Nelson and Cox, 2005) but arrived at via a numerical consideration. Some of the compounds listed in Table 6.3 are high presence compounds (and so have $\delta_c > \Delta = 4\%$), e.g. atp-adp. However, we find cofactors which are not high presence compounds, e.g. q8- q8h2.

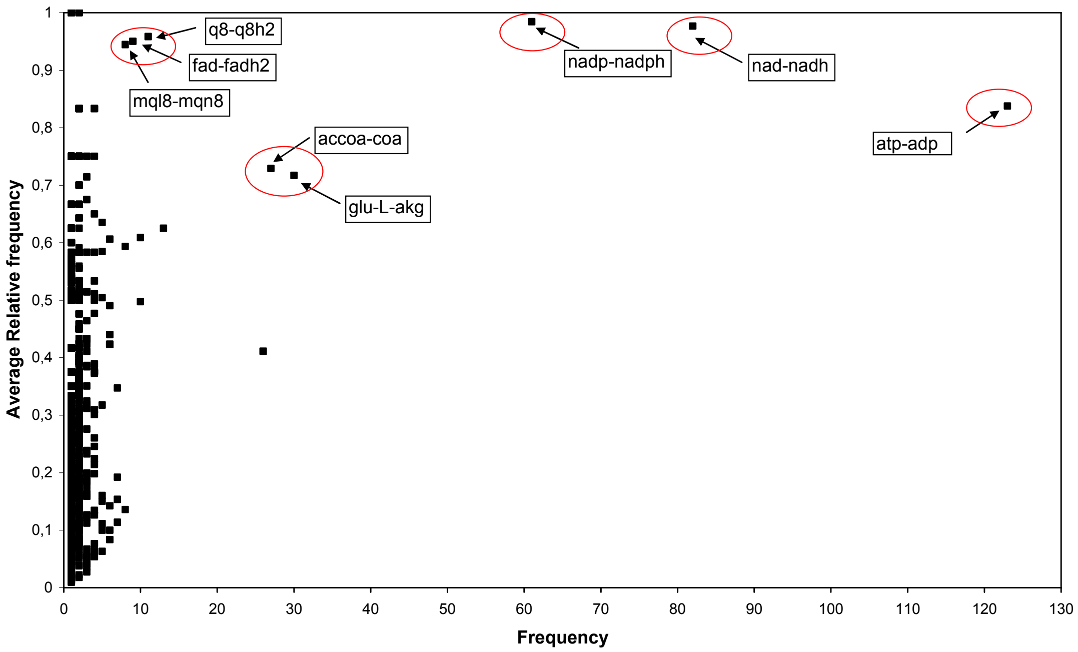

Figure 6.6: Graphical analysis of the pairs

atp	adp
nad	nadh
nadp	nadph
glu-L	akg
accoa	coa
q8	q8h2
fad	fadh2
mql8	mqn8

Table 6.3: List of cofactors

We denote D_3 as the set of cofactors, formally D_3= [(α,β) | compounds α and β are a cofactor pair] so D_3 is a set of compound pairs, not simply a set of compounds (cf D_1 and D_2 above). For convenience in presenting our constraints in a mathematical form below we adopt the convention that each compound pair appears twice in D_3 (e.g. if we had had a single cofactor pair composed of q8 and q8h2 then we would have D_3 = [(q8,q8h2), (q8h2,q8)].

If two compounds (α and β) are a cofactor pair then we impose the constraint that the compounds in this cofactor pair cannot be on the metabolic path through any reaction that involves them both provided it is possible for that reaction to have other compounds on such a path. Consider each cofactor pair $(\alpha,\beta)\in D_3$ in turn. Consider each reaction r in turn. If $(\alpha,\beta)\in D_3$ satisfy:

- $\alpha \neq S$ and $\beta \neq T$ (so the cofactor pair does not involve either the source compound or the target compound); and
- $n_{\alpha r} \geq 1$ and $p_{\beta r} \geq 1$ (so the cofactor pair is involved with reaction r with α as the input compound and β as the output compound); and
- for reaction r there exists an input compound λ (i.e. $n_{\lambda r} \geq 1$) with $\lambda \neq \alpha$, $\lambda \notin D_1 \cup D_2$-[S] and an output compound μ (i.e. $p_{\mu r} \geq 1$) with $\mu \neq \beta$, $\mu \notin D_1 \cup D_2$-[T] such that $(\lambda,\mu) \notin D_3$ (so the pair (λ,μ) is not itself a cofactor pair)

then we impose the constraint that the cofactor pair (α,β) cannot be on a metabolic path through r, i.e.

$$u_{\alpha r k}=0 \qquad \text{and} \qquad v_{r\beta k}=0 \qquad k=1,\ldots K \qquad (6.23)$$

Finally, for the sake of clarity, we denote D_4 as the set of main compounds. A main compound c is a low presence, organic compound not involved in any cofactor pair. Formally $D_4 = [c \mid c \notin D_1 \cup D_2 \text{ and } (c,\alpha) \notin D_3,\ \forall\, \alpha=1,\ldots, C, c=1\ldots,C]$. From the above equations, it can be observed that we encourage main compounds to appear as the intermediate compounds of the K metabolic paths. This is a common practice in the metabolic pathways literature. As noted in Chapter 5, Croes *et al.*, 2006 prevents high presence compounds from appearing in the k-shortest paths. Jeong *et al.*, 2000 showed that the list of high presence compounds is almost identical for a number of organisms. Ma and Zeng, 2003 removed cofactors and inorganic compounds from the metabolic network in their path finding analysis. They note that a cofactor might (or might not) be removed according to the reaction considered. However, the removal was done in a manual fashion (rather than mathematical). In addition, they noted that their list of cofactors and inorganic compound is very similar to the list of high presence compounds presented in Jeong *et al.*, 2000. A similar analysis can be found in Horne *et al.*, 2004.

6.2.5 Objective function

The results obtained for path finding approaches in Chapter 5 show that it is appropriate to assume that intermediate compounds in metabolic pathways are usually low presence compounds. This assumption is done on the basis of evolution. Indeed it seems reasonable to believe that metabolic pathways have evolved to accomplish their function in the cell via low presence compounds. Low presence compounds provide a degree of specificity to metabolic pathways. Note that the fact that metabolic pathways make use of specific compounds contrasts with the robustness shown for enzymes, especially due to the presence of isoenzymes. Starting from this assumption, we try and ensure low presence compounds are on metabolic paths and high presence compounds off such paths. Thus, we need a factor relating to the presence of compounds in metabolic paths.

We define N_{cr} to be the relative presence for compound c as input to reaction r, formally $N_{cr} = \delta_c/\min[\delta_b \mid n_{br}>0, b=1,\ldots,C]$ when $n_{cr}>0$. If a compound c has the lowest percentage presence over all input compounds for reaction r then N_{cr} will have the value 1. Such compound c with the lowest relative presence will be the most specific of the input compounds of the reaction r, since it participates in less biochemical reactions than any other input compound.

We define P_{rc} to be the relative presence for compound c as output from reaction r, formally $P_{cr} = \delta_c/\min[\delta_b \mid p_{br}>0, b=1,\ldots,C]$ when $p_{cr}>0$. If a compound c has the lowest percentage presence over all output compounds for reaction r then P_{cr} will have the value 1. Such compound c with the lowest relative presence will be the most specific of the output compounds of the reaction r, since it participates in less biochemical reactions than any other output compound.

In order to clarify the concept of the relative presence, Table 6.4 shows the relative presence values for reaction R1: $C1 \rightarrow C2 + C3$ in the example metabolic network in Figure 6.1. Note that the presence of C1 can be calculated as $\delta_{C1}=W_{C1}/R$, where W_{C1} is the total connectivity of C1. In the example, $W_{C1}= 4$ and R=8, namely $\delta_{C1}=4/8= 0.5$. Since only C1 appear as input compound of R1, then necessarily N_{C1R1}=1. Similarly, δ_{C2}=0.25 and δ_{C3}= 0.375, having that C2 is the most specific output compound of R1.

Input	**δ_c**	**min δ_c value**	**N_{cr}**
C1	0.5	0.5	1
Output	**δ_c**	**min δ_c value**	**P_{cr}**
C2	0.25	0.25	1
C3	0.375	0.25	1.5

Table 6.4: Relative presence values, P_{rc} and N_{cr}, for R1

Since we are trying to find paths converting the source compound into the target compound via specific input and output compounds, the factor we are interested in minimising is:

$$\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{r=1,n_{cr}>0}^{R} N_{cr}u_{crk} + \sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{r=1,p_{cr}>0}^{R} P_{cr}v_{rck} \qquad (6.24)$$

which is the weighted sum over all metabolic paths of the arcs in the path each being weighted by its relative presence. Note that equation (6.24) adds up 2K terms, one input and one output term for each of the K metabolic paths we choose. In order to provide an average value of relative presence for the metabolic pathway, we define the specificity of a pathway Ψ, using

$$\Psi = (\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{r=1,n_{cr}>0}^{R} N_{cr}u_{crk} + \sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{r=1,p_{cr}>0}^{R} P_{cr}v_{rck})/2K \qquad (6.25)$$

Interestingly, when trying to minimise the specificity, we are implicitly minimising the number of active reactions in the metabolic pathway ($\sum_{r=1}^{R} z_r$), which we, henceforth, refer as to the length of the pathway, L. We empirically demonstrate in the Results section that length and specificity are closely related.

For this reason we found it of interest to analyse two different objectives, namely one minimising the specificity of the pathways (Ψ) and other minimising the length of the pathway ($L = \sum_{r=1}^{R} z_r$).

In addition, we include a secondary term that minimises the number of unbalanced main compounds. The appearance of unbalanced main compounds usually implies that the pathway under study will need an additional pathway (in the global sense of all pathways in the organism) to balance such compounds. In order to minimise the dependency of the pathway under consideration with respect to other pathways, we need a factor relating to the number of unbalanced main compounds. We denote W the number of unbalanced main compounds (excluding the source and the target compound), formally W = $\sum_{c=1, c \neq S, c \neq T, c \in D_4}^{C}$ ($e_c + f_c$). The best case is obviously W=0, i.e. the pathway involves no unbalanced main compounds and thus the pathway is completely independent of other pathways.

Consequently, the IBP model considers two different objective functions:

$$\text{minimise } 1000\Psi + 100W \tag{6.26}$$

$$\text{minimise } 1000L + 100W \tag{6.27}$$

The values adopted for the weight of Ψ and L here (1000) and that of W (100) were decided empirically based on a few pathways. We suspect that no change is found when factor 1000 is increased (e.g. to 10000 or 100000). This would position our analysis in the extreme case, i.e. $M_3\Psi$ + W for objective (6.26) and M_3L + W for objective (6.27), where M_3 is a large positive constant. However, a more precise and extensive computational validation would be needed to confirm this suspicion. Note that whilst L and W must take integer values Ψ is, by its very nature (equation (6.25)), fractional. Although it is equivalent to divide each of (6.26) and (6.27) by 100 we leave them in the form given.

6.2.6 Overview

The IBP model (optimise (6.26) or (6.27) subject to (6.1)-(6.23) plus cycle constraints) is an integer linear program. Algorithmically such programs are solved by linear programming based tree search. Here, as in previous chapters, we used Cplex.

6.3 Results

As in previous chapters, we used the metabolic network of *E.Coli* presented by Reed *et al.*, 2003, which comprises 880 cytosolic reactions and 613 compounds. A full list of reactions/compounds can be found in Appendices A and B.

We applied the IBP model to the forty *E.Coli* experimentally determined pathways shown in Table 6.4. The pathways used were taken from Keseler *et al.*, 2005; Nelson and Cox, 2005 and http://biocyc.org/ECOLI/. A detailed description of the experimentally determined pathways can be found in Appendix C.

6.3.1 Structural recovery of experimentally determined pathways

As in Chapter 3, we mean here by structural recovery that, once the IBP model is solved, the solution is precisely the same as the experimentally determined metabolic pathway, both in terms of the reactions/compounds involved in the pathway and its inherent stoichiometry (reaction ticks).

The IBP model requires Q_S, Q_T the number of source and target molecules to be specified. In addition, the IBP model needs to specify the input parameter K, which defines the maximum number of metabolic paths. As we described in the Introduction section of this chapter, the reason for considering K metabolic paths was to meet the issue of branched metabolic pathways. In our forty experimentally determined pathways we only have one branched metabolic pathway, pathway 6 (Pentose Phosphate Pathway). This pathway can be recovered with two metabolic paths. Thus K=2 would be sufficient for recovering any of our forty pathway under study. However we applied the IBP model for K=1, 2,...,5 so as to see the performance of the IBP model under different values of K.

In addition, as explained above, the IBP model considers two different objectives. Objective (6.26) gives primary weight to the specificity (Ψ), whilst objective (6.27) gives primary weight to the length of the pathway (L). We present below results for both objectives.

Table 6.5 shows the performance of the IBP model for objective (6.26). The IBP model recovers 32 of our 40 experimentally determined pathways for K=1 and 33 pathways for K=2,…,5.

Pathwy number	Pathway name	Pathway recovered? Objective (6.26)	
		K=1	K=2,3,4,5
1	Gluconeogenesis	yes	yes
2	Glycogen	yes	yes
3	Glycolysis	no	no
4	Proline biosynthesis	yes	yes
5	Ketogluconate metabolism	yes	yes
6	Pentose phosphate	no	yes
7	Salvage pathway deoxythymidine phosphate	no	no
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	no	no
9	NAD biosynthesis	yes	yes
10	Arginine biosynthesis	yes	yes
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	yes
14	Histidine biosynthesis	yes	yes
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	yes
17	Pantothenate biosynthesis	yes	yes
18	Tetrahydrofolate biosynthesis	no	no
19	Riboflavin and FMN and FAD biosynthesis	yes	yes
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	yes
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	yes
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	yes
25	Fucose degradation	yes	yes
26	Entner-Doudoroff Pathway	no	no
27	Anaerobic Respiration	no	no
28	Arginine degradation	yes	yes
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	yes
31	Phospholipid Biosynthesis	yes	yes
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	yes
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	no	no
37	Propionate Degradation	yes	yes
38	Glutamate Biosynthesis Cycle	yes	yes
39	Biotin Synthesis	yes	yes
40	Glycerol Degradation	yes	yes
Number of “yes” entries		32	33

Table 6.5: Structural Recovery for K=1,…,5 and objective (6.26)

In addition, Table 6.6 shows the results of the IBP model for objective (6.27). The IBP model recovers in this case 29 of our 40 experimentally determined pathways for K=1 and 30 pathways for K=2,…,5.

Pathway number	Pathway name	Pathway recovered? Objective 6.27	
		K=1	K=2,3,4,5
1	Gluconeogenesis	yes	yes
2	Glycogen	yes	yes
3	Glycolysis	no	no
4	Proline biosynthesis	no	no
5	Ketogluconate metabolism	yes	yes
6	Pentose phosphate	no	yes
7	Salvage pathway deoxythymidine phosphate	no	no
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	no	no
9	NAD biosynthesis	no	no
10	Arginine biosynthesis	no	no
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	yes
14	Histidine biosynthesis	yes	yes
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	yes
17	Pantothenate biosynthesis	yes	yes
18	Tetrahydrofolate biosynthesis	no	no
19	Riboflavin and FMN and FAD biosynthesis	yes	yes
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	yes
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	yes
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	yes
25	Fucose degradation	yes	yes
26	Entner-Doudoroff Pathway	no	no
27	Anaerobic Respiration	no	no
28	Arginine degradation	yes	yes
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	yes
31	Phospholipid Biosynthesis	yes	yes
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	yes
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	no	no
37	Propionate Degradation	yes	yes
38	Glutamate Biosynthesis Cycle	yes	yes
39	Biotin Synthesis	yes	yes
40	Glycerol Degradation	yes	yes
Number of “yes” entries		29	30

Table 6.6: Structural Recovery for K=1,…,5 and objective (6.27)

In terms of recovery, the IBP model for K=1 differs from the IBP model for K=2,…,5 (using both objective (6.26) and objective (6.27)) in that pathway 6 was not recovered. Accordingly, we need K≥2 so as to recover pathway 6. In addition, Table 6.5 and Table 6.6 show that the IBP model has no effect (in terms of recovery) when the K value is increased from K=2 to K=3,4,5. Moreover, as seen in Tables 6.7 and 6.8, applying the IBP model for K>2 has an unnecessary increase in average computation time. For example, the average computation time over the forty pathways is 7.5 seconds for K=2 (1.86Ghz pc, 2GB RAM) and 69.6 seconds for K=5 when objective (6.26) is used. For that reason we henceforth fix K=2.

	K=1	K=2	K=3	K=4	K=5
Average computation time (seconds)	7.0	7.5	25.7	33.7	69.6
Minimum computation time (seconds)	2.7	3.0	3.3	3.7	4.0
Maximum computation time (seconds)	150.4	99.9	776.4	628.8	1,585.8

Table 6.7: Computation times for the forty pathways with objective (6.26)

	K=1	K=2	K=3	K=4	K=5
Average computation time (seconds)	6.3	11.5	42.2	92.1	127.6
Minimum computation time (seconds)	2.5	3.0	3.3	3.8	4.1
Maximum computation time (seconds)	63.3	127.1	815.3	1,834.2	1,871.2

Table 6.8: Computation times for the forty pathways with objective (6.27)

6.3.1.1 Analysis of the structural recovery for K=2

Tables 6.5 and 6.6 indicate that the IBP model recovers 33 of our 40 experimentally determined pathways using one (or both) of the objectives considered. Statistically this is a highly significant result at the 0.001% level. In no case does objective (6.27) achieve recovery and objective (6.26) does not. However, these results are slightly worse than the BP model, which achieves recovery in 38 pathways whether using objective (3.13) or objective (3.14), as noted in Table 6.9. Interestingly, we have a mix of situations: some where both approaches achieve recovery (e.g. pathway 1); some where the IBP model achieves recovery and the BP model does not (e.g. pathway 19); some where the IBP model does not achieve recovery and the BP model does (e.g. pathway 8).

Pathway number	Pathway name	Pathway recovered?	
		BP model	IBP model
1	Gluconeogenesis	yes	yes
2	Glycogen	yes	yes
3	Glycolysis	yes	no
4	Proline biosynthesis	yes	yes
5	Ketogluconate metabolism	yes	yes
6	Pentose phosphate	yes	yes
7	Salvage pathway deoxythymidine phosphate	yes	no
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	yes	no
9	NAD biosynthesis	yes	yes
10	Arginine biosynthesis	yes	yes
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	yes
14	Histidine biosynthesis	yes	yes
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	yes
17	Pantothenate biosynthesis	yes	yes
18	Tetrahydrofolate biosynthesis	yes	no
19	Riboflavin and FMN and FAD biosynthesis	no	yes
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	yes
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	yes
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	yes
25	Fucose degradation	yes	yes
26	Entner-Doudoroff Pathway	yes	no
27	Anaerobic Respiration	yes	no
28	Arginine degradation	yes	yes
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	yes
31	Phospholipid Biosynthesis	yes	yes
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	yes
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	yes	no
37	Propionate Degradation	yes	yes
38	Glutamate Biosynthesis Cycle	yes	yes
39	Biotin Synthesis	no	yes
40	Glycerol Degradation	yes	yes
Number of “yes” entries		38	33

Table 6.9: Comparison between the BP and the IBP model

In thirty-one of the thirty-three "yes" cases obtained in the IBP model for objective (6.26) (and K=2) there is a unique pathway associated with the optimal objective function value and in only two cases is there an alternative pathway providing the same optimal objective function value. However, when using objective (6.27), only twenty-two out of the thirty yes cases provide a unique optimal solution.

As we have a significant number of constraints in the IBP model the question arises as to the relevance of the objective adopted. In the limit for example there may be only one unique solution satisfying the constraints, and if so the objective adopted becomes irrelevant. We found that in all "yes" cases in Table 6.5 and Table 6.6 we have more than one solution satisfying the constraints. In addition, we found that in all 40 pathways (whether recovered or not) we have at least one solution satisfying the constraints. This fact contrasts with the BP model, where we found two cases (pathway 19 and pathway 39) for which the BP model obtains no feasible solution, i.e. no values for the decision variables in the BP model exist which satisfy all the constraints of the BP model. This observation shows that the set of constraints included in the IBP model, (6.1)-(6.23) are less restrictive than those included in the BP model. Thus, the objective adopted is clearly a more relevant issue for the IBP model. A later section analyses in detail the objective functions (6.26) and (6.27) defined in the IBP model.

As in the BP model, equation (6.9) explicitly excludes solutions in which reactions in the pathway produce any of the source compound (or consume any of the target compound). If we amend the IBP model to allow such solutions then, with respect to Table 6.5 and Table 6.6, we degrade the results slightly, failing to recover pathways 1, 10 and 37 for objective (6.26) and pathways 1 and 37 for objective (6.27).

Equation (6.17) defines the scope of the pathway, i.e. reactions active in the pathways must necessarily be contained in at least one of the K metabolic paths. If we remove equation (6.17) from the IBP model, then we degrade the results significantly for objective (6.26), recovering now only 14 pathways. However, the deletion of equation

(6.17) from the IBP model does not have much effect when using objective (6.27), since we now achieve recovery in 27 pathways.

Equations (6.21), (6.22) and (6.23) relate to high presence compounds, inorganic compounds and cofactors constraints. If we do not include these constraints in the IBP model then, we degrade the results slightly for objective (6.26), failing to recover now pathways 1, 9 and 10. However, a significant effect is found for objective (6.27), which now only recovers now 11 pathways. Note that in the 40 pathways under study we found that only pathway 27 violates these constraints.

6.3.2 Q_S,Q_T recovery of experimentally determined pathways

As in the BP model, the IBP model needs to specify the number of molecules of the source and target compounds (Q_S,Q_T) involved in the pathway (equation (6.8)). For the results shown in Table 6.5 and Table 6.6 these values have been taken as equal to those associated with the experimentally determined pathway. In this section, we present results as to the IBP model when applied to a number of different (Q_S,Q_T) pairs ($Q_S,Q_T \leq 6$), so that the dominant pair is determined in terms of the objective function. In the case that the dominant pair is precisely that appearing in the experimentally determined pathway, then the IBP model does recover the (Q_S,Q_T) pair observed in the experimentally determined pathway. Such analysis was exclusively carried out in those pathways in which the IBP model achieves structural recovery, i.e. those pathways having a “yes” in Table 6.5 and Table 6.6. As the IBP model present two different objective functions, (6.26) and (6.27), the criterion for selecting the dominant pair was modified according to objective function optimised. To illustrate this, we show below results obtained for the Gluconeogenesis pathway (pathway 1), whose structure was recovered for objective function (6.26) and (6.27) as shown in Table 6.5 and Table 6.6, respectively, and thus, (Q_S,Q_T) analysis must be carried out for both objectives.

The Gluconeogenesis pathway in Table 6.5 has (Q_S,Q_T)=(2,1). For this pathway Table 6.10 show for a number of different (Q_S,Q_T) pairs ($Q_S,Q_T \leq 6$) the specificity value (Ψ) and the number of unbalanced main compound (W) when the IBP model is solved using objective (6.26). Situations where the IBP model indicated that no feasible solution exists are indicated by a 'X'. Objective (6.26) gives primary weight to minimising the specificity (Ψ) and secondary weight to minimising the number of unbalanced main compounds (W). In order to identify the dominant (Q_S,Q_T) pair with respect to this objective we examine all entries in the table. Let E represent the set of all feasible (Q_S,Q_T) pairs in the table. We apply the following procedure:

- eliminate repeats from E. An entry is a repeat if it presents the same Ψ and W value but precisely k (≥2, integer) times as many source/target molecules. A repeat essentially corresponds to the same reaction set but with the ticks multiplied by a factor of k. In Table 6.10, for example, the entries for (Q_S,Q_T)=(4,2) and (Q_S,Q_T)=(6,3) are a repeat of the entry for (Q_S,Q_T)=(2,1) with k=2 and k=3 respectively. In addition, the pairs seen down the diagonal are all repeats of the entry for (Q_S,Q_T)=(1,1). After elimination of repeats the entries left, for example, in Table 6.10 are (Q_S,Q_T)=(1,1); (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (2,3); (2,4); (2,5); (3,1); (3,2); (3,3); (3,4); (3,5); (4,1); (4,3); (4,5); (4,6); (5,1); (5,2); (5,3); (5,4); (5,6); (6,1); (6,5).

- eliminate from E any entries that do not involve the minimum specificity value. In Table 3.4 the minimum number of reactions is 8.13. Thus, all the remaining entries aside from (Q_S,Q_T)=(1,1) are eliminated from E.

- choose from the remaining entries that which involves the minimum number of unbalanced main compounds, ties broken by minimum number of source molecules used, and then further broken if necessary by maximum number of target molecules produced. As there is only one entry with the minimum specificity in Table 6.10, for example, it is not necessary to tie breaks here.

For this pathway the IBP model indicates that the pair $(Q_S,Q_T)=(1,1)$ dominates all other cases either using objective (6.26). This is indicated by the * superscript on that entry in the Table 6.10. Hence in this case the IBP model for objective (6.26) does not recover the $(Q_S,Q_T)=(2,1)$ pair observed in the experimentally determined pathway.

(Ψ,W)		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number of molecules Q_S of source compound	1	(8.13,1)*	(9.02,4)	(8.96,3)	(9.02,4)	(8.96,3)	(9.02,4)
	2	(8.56,0)	(8.13,1)	(9.02,4)	(8.96,3)	(9.02,4)	(9.06,2)
	3	(9.09,3)	(9.19,1)	(8.13,1)	(9.02,4)	(8.96,3)	(9.02,4)
	4	(9.09,3)	(8.56,0)	(9.19,1)	(8.13,1)	(9.02,4)	(8.96,3)
	5	(9.09,3)	(9.09,3)	(9.19,1)	(9.19,1)	(8.13,1)	(9.02,4)
	6	(9.09,3)	(9.09,3)	(8.56,0)	(9.19,1)	(9.19,1)	(8.13,1)

Table 6.10: IBP model solution for objective (6.26) for varying Q_S and Q_T for Gluconeogenesis pathway

Amending that procedure to identify the dominant (Q_S,Q_T) pair for objective (6.27) is easily done. The results for the Gluconeogenesis pathway with objective (6.27) can be seen in Table 6.11. As objective (6.27) gives primary weight to minimising the length (L) of the pathway and secondary weight to minimising the number of unbalanced main compounds (W) in the pathway, the procedure for identifying the dominant pair with objective (6.27) is:

- eliminate repeats from E. In Table 6.11 the entries for $(Q_S,Q_T)=(4,2)$ and $(Q_S,Q_T)=(6,3)$ are a repeat of the entry for $(Q_S,Q_T)=(2,1)$ with k=2 and k=3 respectively. In addition, the pairs seen down the diagonal are all repeats of the entry for $(Q_S,Q_T)=(1,1)$. After elimination of repeats the entries left are $(Q_S,Q_T)=(1,1)$; (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (2,3); (2,5); (3,1); (3,2); (3,4); (3,5); (4,1); (4,3); (4,5); (5,1); (5,2); (5,3); (5,4); (5,6); (6,1); (6,5).

- eliminate from E any entries that do not involve the minimum L value. In Table 6.11 the minimum number of reactions is 7. Thus, all the remaining entries aside from (Q_S,Q_T)=(1,1) are eliminated from E.

- choose from the remaining entries that which involves the minimum number of unbalanced main compounds, ties broken by minimum number of source molecules used, and then further broken if necessary by maximum number of target molecules produced. As there is only one entry with the minimum L value in Table 6.10, for example, it is not necessary to tie breaks here.

For this pathway the IBP model indicates that the pair (Q_S,Q_T)=(1,1) dominates all other cases. Hence in this case the BP model for objective (6.27) does not recover the (Q_S,Q_T)=(1,1) pair observed in the experimentally determined pathway.

(L,W)		Number of molecules Q_T of target compound					
		1	2	3	4	5	6
Number of molecules Q_S of source compound	1	(7,1)*	(8,8)	(9,3)	(9,3)	(9,3)	(9,3)
	2	(9,0)	(7,1)	(9,3)	(8,8)	(9,3)	(9,3)
	3	(10,2)	(10,1)	(7,1)	(9,3)	(9,3)	(8,8)
	4	(10,2)	(9,0)	(10,1)	(7,1)	(9,3)	(9,3)
	5	(10,2)	(10,2)	(10,1)	(10,1)	(7,1)	(9,3)
	6	(10,2)	(10,2)	(9,0)	(10,1)	(10,1)	(7,1)

Table 6.11: IBP model solution for objective (6.27) for varying Q_S and Q_T for Gluconeogenesis pathway

We have repeated the analysis shown in Table 6.10 and Table 6.11 for those cases in which the IBP model recovers the pathway structure (see Appendix C). The summary of this analysis can be seen in Table 6.12. Our judgment is that for thirty of the forty pathways the IBP model (either objective (6.26) or (6.27)) recovers the (Q_S,Q_T) pair observed in the experimentally determined pathway. Statistically this is a highly significant result (significant at the 0.001% level).

Pathway number	**Pathway name**	**(Q_S,Q_T) recovered?**	
		Objective 6.26	Objective 6.27
1	Gluconeogenesis	no	no
2	Glycogen	yes	yes
3	Glycolysis	-	-
4	Proline biosynthesis	yes	-
5	Ketogluconate metabolism	yes	yes
6	Pentose phosphate	no	no
7	Salvage pathway deoxythymidine phosphate	-	-
8	Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle	-	-
9	NAD biosynthesis	yes	-
10	Arginine biosynthesis	no	-
11	Sperdimine biosynthesis	yes	yes
12	Threonine Degradadation to synthetise propionate	yes	yes
13	Serine biosynthesis	yes	yes
14	Histidine biosynthesis	yes	yes
15	Tirosine biosynthesis	yes	yes
16	Coenzyme A biosynthesis	yes	yes
17	Pantothenate biosynthesis	yes	yes
18	Tetrahydrofolate biosynthesis	-	-
19	Riboflavin and FMN and FAD biosynthesis	yes	yes
20	Heme Biosynthesis	yes	yes
21	De novo sinthesis of pyrimidine ribonucletides	yes	yes
22	De novo sinthesis of pyrimidine deoxyribonucletides	yes	yes
23	Phenylethylamine degradation	yes	yes
24	Rhamnose degradation	yes	yes
25	Fucose degradation	yes	yes
26	Entner-Doudoroff Pathway	-	-
27	Anaerobic Respiration	-	-
28	Arginine degradation	yes	no
29	Proline degradation	yes	yes
30	Glycolate degradation	yes	yes
31	Phospholipid Biosynthesis	yes	yes
32	Biosynthesis of cysteine	yes	yes
33	Allantoin degradation	yes	yes
34	Deoxycytidine degradation	yes	yes
35	Phenylalanine Biosynthesis	yes	yes
36	Glyoxylate Cycle	-	-
37	Propionate Degradation	yes	yes
38	Glutamate Biosynthesis Cycle	yes	yes
39	Biotin Synthesis	yes	yes
40	Glycerol Degradation	yes	yes
Number of “yes” entries		30	27

Table 6.12: (Q_S,Q_T) Recovery for the IBP model

Considering Table 6.5, 6.6 and 6.12 it is clear that objective (6.26) is slightly better that objective (6.27). For this reason we focus only on objective (6.26) and (Ψ,W) in the following sections.

6.3.3 Objective function

In the IBP model we consider two different objective functions. Whilst objective (6.26) gives primary weight to minimising the specificity (Ψ), objective (6.27) gives primary weight to minimising the length (L) of the pathways. Both objectives give secondary weight to minimising the number of unbalanced main compounds (W). In this sub-section, we are concerned with empirically showing that specificity and length (Ψ and L) are approximately equal, i.e. $\Psi \approx L$. In addition, we illustrate the reason as to why W was included in the objective function.

Table 6.5 and Table 6.6 show that the set of pathways recovered by objective (6.27) are also recovered by objective (6.26), which additionally recovers pathway 4, 9 and 10. This fact evidences a possible relationship between Ψ and L, since the secondary term (W) remains unchanged. Based on this empirical observation, we plot in Figure 6.7 the values of Ψ and L obtained for our forty experimentally determined pathways when we solve the IBP model for objectives (6.26) and (6.27) respectively. It can be observed that Ψ is very close to L in the majority of the pathways, the average value of Ψ-L among the forty pathways $E(\Psi-L)=0.15$ and the standard deviation $\sigma(\Psi-L) = 0.63$. Note also here that for linear pathways we always have $\Psi \geq L$, e.g. pathway 4. This does not apply for branched pathways, e.g. pathway 1, where $\Psi_1 < L_1$.

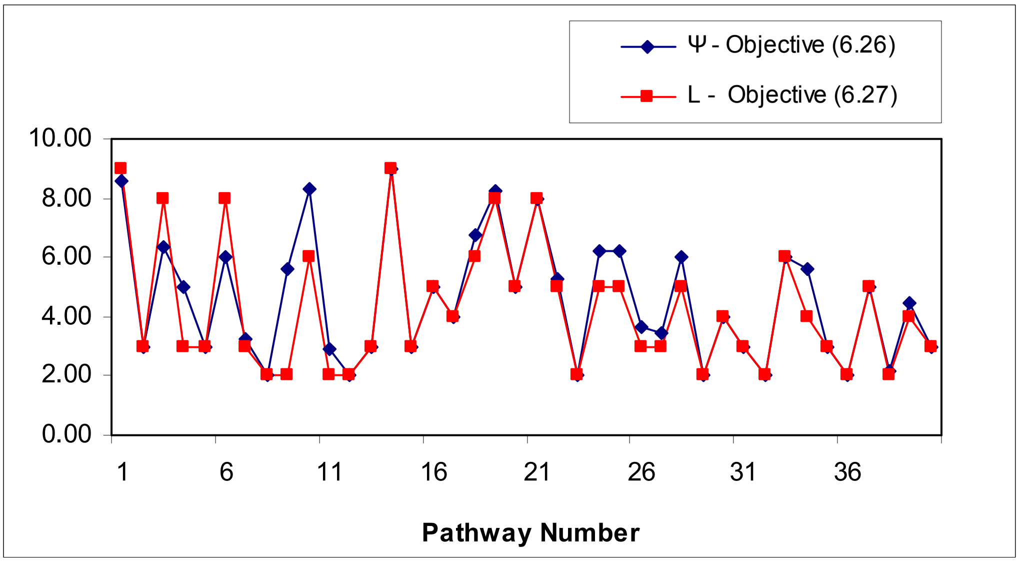

Figure 6.7: Correspondence between between Ψ and L

The maximum difference between Ψ and L in Figure 6.7 is found in pathway 9, where Ψ_9-L_9=3.60. Since objective (6.26) recovers pathway 9 and objective (6.27) does not, we expect Ψ_9 to be closer than L_9 to the actual length of pathway 9. We denote L^{edp} the actual length of the experimentally determined pathway, e.g. actual length of pathway 9, L_9^{edp} =5. Figure 6.8 plots the values of L^{edp} and L in our forty experimentally determined pathways. It can be observed that, in general, $L^{edp} \geq L$, since the solution of the IBP model for objective (6.27) identifies the pathway with the minimal length. In the case the IBP model for objective (6.27) achieves recovery, then $L^{edp} = L$. For pathway 9, we have L_9^{edp} - L_9 = 6. We averaged L^{edp} - L in our forty experimentally determined pathways and found that $E(L^{edp} - L)$=0.58 and $\sigma(L^{edp} - L)$= 1.25. As shown in Figure 6.9, we repeated this analysis between L^{edp} and Ψ, obtaining $E(L^{edp} - \Psi)$=0.25 and $\sigma(L^{edp} - \Psi)$ = 1.25. For pathway 9, in particular, $L_9^{edp} - \Psi_9$ = -0.60. We can then conclude that objective (6.26) is somewhat more accurate than objective (6.27) for recovering experimentally determined pathways.

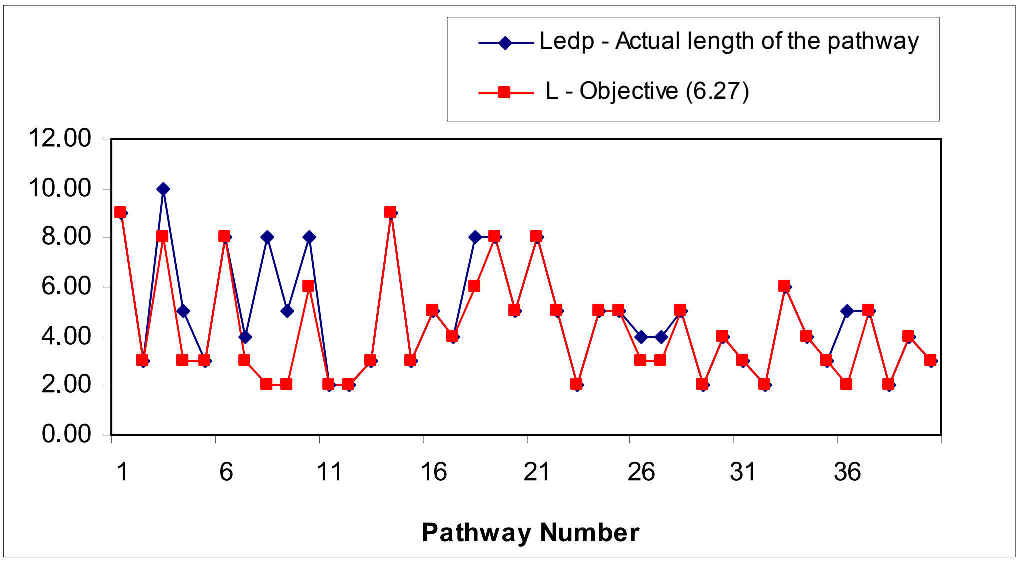

Figure 6.8: Correspondence between L and L^{edp}

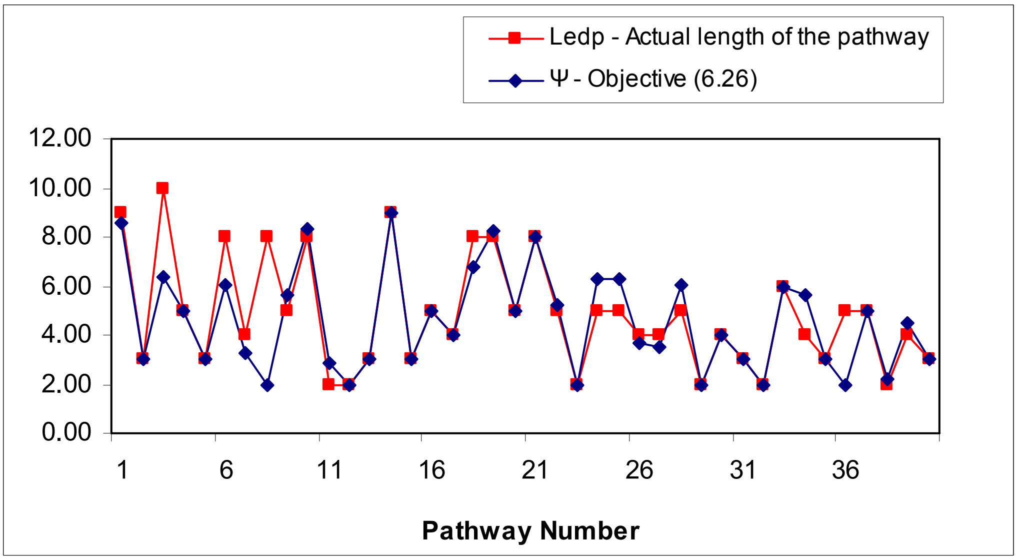

Figure 6.9: Correspondence between Ψ and L^{edp}

This last conclusion is confirmed by observations given above. First and foremost, objective (6.26) presents a higher recovery rate (both structure and (Q_S,Q_T)) than objective (6.27). In addition, objective (6.26) better characterises metabolic pathways, as generally a unique pathway is associated with the optimal solution. Objective (6.27) presents eight cases where more than one optimal solution was found.

With respect to the number of unbalanced main compounds W, a discussion is given below to evaluate the impact of W in the objective function of the IBP model. We exclusively use here objective (6.26) (with K=2) for this discussion. A similar analysis can be done for objective (6.27).

Figure 6.10 plots the value of (Ψ, W) when W is fixed to a constant value (W=0,1,2,3,4,5,6) for pathway 18. We recall here that pathway 18 was not recovered for objective (6.26). The solution from the IBP model corresponds to W=4, where we found the lowest Ψ value, as can be observed in Figure 6.10. Note that no solution exists when W is fixed to 0 or 1. Interestingly, when we fix W=3, the IBP model precisely corresponds to pathway 18, i.e. we achieve recovery. In addition, from Figure 6.10, it is easy to recognize that IBP model will not have a good performance for every W value. For example, when W=2, Ψ presents an elevated value with respect to its value for W=3. The fact that, by considering fixed W values, we have the experimentally determined pathway being amongst the solutions found does, we believe, indicate that W is a relevant factor to include in the IBP model.

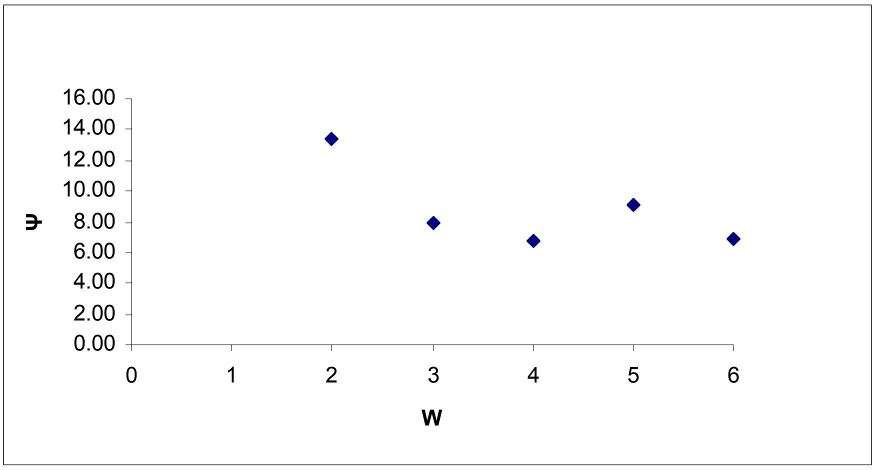

Figure 6.10: (Ψ, W) analysis for pathway 18

This analysis has been repeated for non-recovered pathways that do not constitute a cycle, namely pathway 3, 7, 18, 26, 27. Cyclic pathways are analysed separately in the following section. Table 6.13 shows the specificity (Ψ) for different W values, (W=0,1,2,3,4,5,6), e.g. Ψ=9.83 when W=1 in pathway 3. Note that we applied a time limit for the computation of some values of Table 6.13. Situations where the IBP model indicated that no feasible solution exists are indicated by ‘X’. We found the experimentally determined pathway in W=3 for pathway 18 and W=1 for pathway 26. Instead, pathway 3, 7 and 27 were not recovered for any W value. Note here that pathway 27 does not satisfy equation (6.23) of the IBP model, thus it cannot be recovered. That is the reason as to why we set infeasible in the solution (Ψ, W) value for pathway 27. In addition, the IBP model recognizes for pathway 7 that the best performance will be found for W=2. In summary, we think these results illustrate the relevance of including W in the objective function.

Pathway Number	W=0	W=1	W=2	W=3	W=4	W=5	W=6	Edp Solution (Ψ, W)	Recovered?
3	7.33	9.83	6.33	7.86	9.65	19.73	8.50	(9.57, 0)	no
7	X	X	3.26	X	X	X	X	(6.955,2)	no
18	X	X	13.38	8.00	6.75	9.16	6.89	(8,3)	yes
26	X	4.41	9.56	3.66	7.8	X	X	(4.41,1)	yes
27	5	3.47	7.25	6.31	6.32	X	X	X	no

Table 6.13: Summary of (Ψ, W) analysis for non-recovered and non-cyclic pathways

6.3.4 Non-recovered metabolic pathways

6.3.4.1 Cyclic pathways

The IBP model shows particular difficulties to recover cyclic pathways. We define a cyclic pathway to be a pathway satisfying S=T and $Q_S = Q_T$. In our forty experimentally determined pathways we have two cyclic pathways: pathway 8 (Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle) and pathway 36 (Glyoxylate Cycle). Note here that pathway 38 (Glutamate Biosynthesis Cycle) satisfies S=T, but $Q_S \neq Q_T$. For this reason, this pathway was not included in the list of cyclic pathways.

As discussed above, the IBP model for both objective (6.26) and (6.27) minimises the number of active reactions involved in the pathway. In cyclic pathways, the IBP model finds solutions whose length is far from the actual length of the experimentally determined pathway, as can be observed in Figure 6.7 and 6.8. For example, for pathway 8, Ψ_8=2, L_8=2 and L_8^{edp}=8. The biological significance of these solutions is however questionable. It is clear from the biochemistry literature (Nelson and Cox, 2005) that the purpose of cyclic pathways is to produce a key compound for the cell. Pathway 8 (TCA Cycle) produces one molecule of ATP and one molecule of NADH. ATP is the cell's energy currency, as described in Chapter 3. NADH is a key compound in the respiratory chain, which produces ATP via electron transfers (Nelson and Cox, 2005). In addition, NADH is generally considered a target compound to measure the performance of cellular metabolism (Mayevsky and Rogatsky, 2007). Pathway 38 (Glyoxylate Cycle) also produces one molecule of NADH. For that reason we found interesting to analyse the performance of the IBP model by including new constraints related to ATP and NADH in cyclic pathways.

Denoting ATP as compound 1 for simplicity, we can force ATP to be produced by including the following constraint:

$$\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r \geq 1, \text{ i.e. } e_1=1 \qquad \text{if } S=T \text{ and } Q_S=Q_T \qquad (6.28)$$

Denoting NADH as compound 2 for simplicity, we can force NADH to be produced by including the following constraint:

$$\sum_{r=1}^{R} p_{2r}t_r - \sum_{r=1}^{R} n_{2r}t_r \geq 1 \text{ , i.e. } e_2=1 \qquad \text{if } S=T \text{ and } Q_S=Q_T \qquad (6.29)$$

In addition, we might impose a constraint forcing that at least one of ATP and NADH must be produced:

$$e_1 + e_2 \geq 1 \qquad \text{if } S=T \text{ and } Q_S=Q_T \qquad (6.30)$$

Table 6.14 shows the performance of the IBP model with objective (6.26) model when these constraints are considered. We set K=1 for cyclic pathways (as opposed to K=2 for other pathways). The reason is that by definition a cyclic pathway is one from S back to T (=S) involving $Q_S=Q_T$ molecules. Hence constraining the solution to have just a single metabolic path from S to T (=S) seemed appropriate. Similar analysis can be done for objective (6.27). In order to recover pathway 8 four different strategies were applied. We found that the IBP model recovers this pathway when we force both ATP and NADH to be produced, i.e. including equation (6.28) and (6.29). In the case of pathway 36 we applied two strategies. The IBP model recovers pathway 36 either forcing NADH to be produced, equation (6.29), or forcing at least one of ATP and NADH to be produced, equation (6.30). These results show the strength of the IBP model to deal with cyclic pathways. They also reveal that biologically significant cyclic pathways might not be properly defined with the source (target) compound, i.e. they need extra knowledge related to the purpose of the pathway. Finally, (Q_S,Q_T) analysis has been carried out for the three 'yes' cases in Table 6.14, finding that we recovered the pair observed in the experimentally determined pathway.

Strategy	**Cyclic pathways Recovered?**	
	Pathway 8	Pathway 36
IBP model + (6.28)	no	-
IBP model + (6.29)	no	yes
IBP model + (6.28) + (6.29)	yes	-
IBP model + (6.30)	no	yes

Table 6.14: Performance of IBP model with additional cyclic pathway constraints

6.3.4.2 The Glycolysis pathway and other non-recovered non-cyclic pathways

Considering Table 6.13, the IBP model does not recover the Glycolysis pathway (pathway 3). The Glycolysis pathway represents one of the most popular and accepted pathways in biochemistry. This pathway converts one molecule of D-Glucose into two molecules of Pyruvate. In addition, it is well known for being a very active pathway for producing ATP. For this reason, we found of interest to examine the performance of the IBP model forcing ATP to be produced to excess, namely $\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r \geq 1$, i.e. $e_1=1$. When we applied the IBP model to this situation, we achieved structural recovery using objective (6.26) for K=1. We also recovered the pair associated with the pathway when (Q_S,Q_T) analysis was conducted (see Appendix C for details). This fact shows that the addition of biologically meaningful constraints to the IBP model clearly refine the performance of the model.

With respect to Pathway 7 (Salvage pathway deoxythymidine phosphate), we would like to note that, despite the fact that the IBP model obtains a solution different from the experimentally determined pathway presented in Appendix C, such solution is a known alternative pathway, as can be noted in http://biocyc.org/ECOLI/.

Finally, the IBP model does not recover Pathway 27 (Anaerobic Respiration). The reason for this flaw is that Pathway 27 does not satisfy equation (6.23) relating to cofactor pairs. This basically implies that Pathway 27 is not contained in the set of feasible solutions, i.e. Pathway 27 will never be recovered unless this constraint is excluded from the IBP model**.** Since this flaw uniquely occurs in this pathway, we think that Pathway 27 might be ill-defined in EcoCyc (http://biocyc.org/ECOLI/), or possibly the cofactor constraint is inappropriate for this pathway. Note here that even if we exclude equation (6.23) however we still fail to recover the pathway.

6.4 Conclusions

We have presented the Improved Beasley-Planes (IBP) model for recovering experimentally determined metabolic pathways. The IBP model showed an acceptable performance, namely recovering the structure in 33 of our 40 experimentally determined pathways and the (Q_S,Q_T) pair in 30 cases. When further constraints related to cyclic pathways and glycolysis are included, the IBP model recovers the structure and the (Q_S,Q_T) pair in 36 and 33 pathways, respectively. Though this result is somewhat less accurate than the BP model, the IBP model appropriately solves the issue of low presence unbalanced compounds described in Chapter 3. In addition, the IBP model presents a significant advance from the modelling standpoint, since elements of stoichiometric approaches and path finding approaches are incorporated. In the light of these modelling advances, and the results obtained, we think the IBP model is more general and applicable than previous approaches (including the BP model) for determining biologically meaningful metabolic pathways.

Chapter 7

Conclusions

7.1 Summary of the thesis

A central issue in understanding cellular metabolism relates to the identification and regulation of the specific metabolic pathways that operate inside the cell. Thanks to advances in the field of genomics, a computational/mathematical analysis of metabolic pathways at the genome-scale can now be carried out. In this thesis we have considered computational (mathematical) methods for determining biologically significant metabolic pathways from a given metabolic network.

In Chapter 2, we introduced and discussed existing approaches to metabolic pathway analysis, namely stoichiometric approaches and path finding approaches. Although both types of approach present clear limitations, we consider that path finding approaches present higher potential for determining biologically meaningful metabolic pathways.

In Chapter 3, we presented our initial approach, named the Beasley-Planes (BP) model, so as to refine the search for biologically meaningful metabolic pathways. The BP model can be considered a stoichiometric approach, since the pseudo steady state condition is applied to a subset of biochemical compounds called low presence compounds. In contrast to classic stoichiometric approaches, the BP model uses integer linear programming to obtain a single (optimal) metabolic pathway that converts a source compound into a target compound. This perspective is novel from the modelling point of view.

We tested the performance of the BP model in 40 experimentally determined pathways, achieving recovery in 37 out of 40 experimentally determined pathways. However we found that the pseudo steady state condition does not apply to every experimentally determined pathway, i.e. low presence unbalanced compounds appear in most of our 40 experimentally determined pathways. Hence for these pathways we did not force these compounds to be balanced. Subsequently, the BP model needs to know beforehand the low presence unbalanced compounds contained in the experimentally determined pathways. This clearly constitutes a limitation for predicting novel (unknown) metabolic pathways. Despite this issue, the accurate results obtained here indicate that there is reason to believe there is a general mathematical model underlying the many different experimentally determined pathways seen.

In Chapter 4 we proposed a novel pathway-oriented approach to investigate reaction knockout that is built upon the BP model. This approach was applied for optimally disrupting metabolic pathways. We distinguished two cases: the disruption of a single metabolic pathway, namely the glycolysis pathway; and the disruption of two (related) metabolic pathways: the glycolysis pathway and the TCA Cycle. The results obtained (in terms of the enzyme/reaction to target to best disrupt the pathway) accord with work done from a non-mathematical (biochemical/medical) perspective. Despite the issue of low presence unbalanced compounds described above, we show that the BP model can be successfully applied to a given unknown situation.

In order to find alternative ways to solve the issue of low presence unbalanced compounds, we carried out a detailed analysis of path finding approaches in Chapter 5. In particular, we developed a path finding approach (via integer linear programming) with a similar distance metric to one given previously in the literature. We tested our approach in the first ten experimentally determined pathways studied in Chapter 3. Although the results presented are clearly worse than the BP model, our path finding approach served as a building-block for a more refined approach (the IBP model described in Chapter 6), which overcomes the issue of low presence unbalanced compounds.

Chapter 6 describes our final approach: the Improved Beasley-Planes (IBP) model. As noted above, the IBP model is essentially a path finding approach. However it also includes stoichiometric constraints. This fact constitutes a clear advance from the modelling point of view, since, to the best of our knowledge, no approach in the literature has to date combined both types of approaches. In terms of results, the IBP model is somewhat less accurate than the BP model, namely recovering 33 of our 40 experimentally determined pathways. However, as noted above, the IBP model successfully overcomes the issue of low presence unbalanced compounds. This makes the IBP model a more suitable approach than the BP model for predicting (unknown) novel metabolic pathways. Overall, we think that the IBP model is more powerful than any other mathematical/computational approach in the literature (including the BP model) for determining biologically meaningful metabolic pathways.

7.2 Extensions and future directions

It is clear that, at least in our view, the IBP model is more general than the BP model. Naturally we are not claiming that the IBP model represents a final mathematical model that fully explains all metabolic pathways. Though unrecovered pathways might reflect that our *E. Coli* metabolic network is flawed, this obviously reflect that the IBP model is less than perfect. However we believe that we have made significant steps along the path to a more complete (and applicable) mathematical model for metabolic pathways. More can be done as to constraints, and much effort is needed to more fully determine (by computational experimentation) an appropriate objective function. In our work, for example, we found that the number of reactions and excess ATP were appropriate factors to include in the objective for the BP model. By contrast, in the IBP model we had different factors, the specificity and the number of unbalanced main compounds. A more general model may involve some/all of these factors in its objective, or there may be other factors that have eluded us in the work that we have undertaken. We believe however that this thesis clearly demonstrates that there is a significant chance of discovering/developing a mathematical optimisation model that underlies many/all experimentally determined metabolic pathways.

A natural extension to this thesis would be to refine the IBP model so as to achieve recovery in non-recovered pathways. In order to help achieve this task, bioenergetics constraints as described in Chapter 3 could be incorporated into the model, since full Gibbs Energies values are now available for *E.Coli*. Note that the publication of Gibbs Energy values for *E.Coli* is posterior to the development of both the BP model and the IBP model. As also noted in Chapter 3, the knowledge of cellular concentrations of biochemical compounds would help to improve the estimation of Gibbs Energies values. However, we are still far from having the concentration for every biochemical compound in the metabolic network.

It may be valuable to include data from gene expression profiles. Gene expression profiles provide a measure as to the activity of genes under very different conditions. Gene expression profiles are nowadays available at the genome scale. In relation to metabolic pathways analysis, gene expression might determine the appearance of a certain set of biochemical reactions in a metabolic pathway. The fundamental reason for this is that some (metabolic) genes code for enzymes and enzymes catalyse biochemical reactions. This fact provides a clear link between the genome and metabolism. In terms of mathematical modelling, gene expression profile might introduce additional constraints. For example, experiments in different organisms indicate that the expression of the genes involved in an experimentally determined pathway is correlated. This clearly may help refine the search for biologically meaningful metabolic pathways. Thus, a possible future research direction will be to find properties of metabolic pathways from gene expression profiles.

Recently, an updated and expanded metabolic network for *E.Coli* has been released. This refined network approximately doubles the number of reactions for *E.Coli* as compared with the metabolic network used in this thesis. A clear extension here would be to test the IBP model with this updated metabolic network. In addition, it would be of interest to apply the IBP model to the metabolic network of different organisms, such as Saccharomyces Cerevisiae (baker's yeast) or Homo Sapiens (humans). An analysis as to how metabolic pathways are evolved in different organisms constitutes a possible future direction.

With respect to our reaction knockout approach presented in Chapter 4 (which used the BP model), an obvious task to carry out in the future is to build such an approach using the IBP model. In Chapter 4 this approach was used to investigate the optimal disruption of a (number of) metabolic pathway(s). A further future direction here might be to understand how metabolic pathways reorder after a reaction/gene knockouts. This analysis can obviously be expanded to reaction (or gene) additions. Such questions directly relate to the regulation of cellular metabolism. Clearly, once we have a reliable approach for determining biologically significant metabolic pathways, the next scientific step is to elucidate how these metabolic pathways are regulated to produce coherent and coordinated global behaviour under different conditions. This question is however outside the boundaries of this thesis.

As a final note, we would like to emphasise that we believe that optimisation will play a very important and increasing role in modern cellular biology. What is not clear is that we can always find linear models. For that reason we think that research must be carried so as to find powerful algorithms to solve non-linear mathematical optimisation models.

Bibliography

1. Alper, H., Jin, Y.S., Moxley, J. F. and Stephanopoulos, G. (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metabolic Engineering,* 7, 155-164.

2. Aittokallio, T. and Schwikowski, B. (2006) Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics,* 7(3), 243-255.

3. Arita, M. (2000) Metabolic reconstruction using shortest paths. *Simulation Practice and Theory,* 8, 109-125.

4. Beasley, J.E. and Planes, F.J. (2007) Recovering metabolic pathways via optimization. *Bioinformatics,* 23(1), 92-98.

5. Bebek, G., Yang, J. (2007) PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*; 8, 335.

6. Becker, S, Price, N.D. and Palsson, B.O. (2006) Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics,* **7,**111.

7. Bell, S.L., Palsson, B.O. (2005) Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics*, 21(8), 1739-1740.

8. Borodina, I., Krabben, P. and Nielsen, J. (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Research,* 15, 820-829.

9. Bui,T. and Thompson,C.B. (2006) Cancer's sweet tooth*.* *Cancer Cell,* 9, 419-420.

10. Burgard, A. P. and Maranas, C. D. (2001) Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions. *Biotechnology and Bioengineering*, 74(5), 364-375.

11. Burgard, A. P., Pharkya, P. and Maranas, C. D. (2003) OptKnock: a bilevel

programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6), 647–657.

12. Becker, S.A. and Palsson, B.Ø., (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, 5(1):8.

13. Bell, S.L. and Palsson, B.O. (2005) Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics*, 28(8), 1739-1740.

14. Çakir, T., Tacer, C.S. and Ülgen, K.Ö. (2004) Metabolic pathway analysis of enzyme-deficient human red blood cells. *BioSystems,* 78, 49-67.

15. Carlson, R. and Srienc, F. (2004) Fundamental Escherichia Coli biochemical pathways for biomass and energy production: identification of reactions. *Biotechnology and Bioengineering,* 85(1), 1-19.

16. Clarke, B.L. (1980) Stability of complex reaction networks. In *Advances in Chemical Physics* Vol. 43 (Prigogine, I. and Rice, S.A., eds.), pp. 1-215, John Wiley & Sons, New York.

17. Colom, J.M. and Silva, M. (1991) Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computational of minimal P-semiflows. *Lecture Notes in Computer Science*, 483, 79-112.

18. Covert, M.W. and Palsson, B.O. (2003) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of Theoretical Biology,* 221, 309-325.

19. Croes, D., Couche, F., Wodak, S.J. and van Helden, J. (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Research*, 33, W326-W330.

20. Croes, D., Couche,F., Wodak, S.J. and van Helden, J. (2006) Inferring meaningful pathways in weighted biochemical networks. *J. Mol. Biol.*, 356, 222-236.

21. Dandekar, T., Moldenhauer, F., Bulik, S., Bertram, H. and Schuster, S. (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems*, 70(3), 255-270.

22. Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemical Journal,* 343, 115-124.

23. Deutscher, D., Meilijson, I., Kupiec. M. and Ruppin, E. (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nature Genetics*, 38(9), 993-998.

24. Deville, Y., Gilbert, D., van Helden, J. and Wodak, S. J. (2003) An overview of data models for the analysis of biochemical pathways. *Briefings in Bioinformatics*, 4(3), 246-259.

25. Duarte, N. C., Herrgard, M. J. and Palsson, B. O. (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14, 1298–1309.

26. Dooms, G., Deville, Y. and Dupont, P. (2005) Constrained metabolic network analysis: discovering pathways using CP(Graph). http://www2.info.ucl.ac.be/people/YDE/Papers/wcb05.pdf

27. Ebenhöh, O., Heinrich, R. (2001) Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bulletin of Mathematical Biology*, 63(1), 21-55.

28. Ebenhöh,O. and Heinrich,R. (2003) Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. *Bulletin of Mathematical Biology*, 65, 323-357.

29. Edwards, J. S. and Palsson, B. O. (2000a) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10), 5528–5533.

30. Edwards, J. S. and Palsson, B. O. (2000b) Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*, 1(1).

31. Fantin,V. R., St-Pierre, J. and Leder, P. (June 2006) Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 9, 425-434.

32. Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V. and Palsson, B.Ø. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information *Molecular Systems Biology*, 3, 121.

33. Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D. and Palsson, B. O. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering*, 91(5), 643-648.

34. Förster, J., Famili, I., Fu, P., Palsson, B.O. and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2), 244-253.

35. Förster, J., Gombert, A.K. and Nielsen, J. (2002) A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology and Bioengineering,* 79, 703-712.

36. Gagneur, J. and Klamt, S. (2004) Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5, 175.

37. Garber, K. (2004) Energy boost: the Warburg effect returns in a new theory of cancer. *J. Nat. Cancer Inst.* 96, 1805-1806.

38. Gatenby,R.A. and Gillies,R.J. (2004) Why do cancers have high aerobic glycolysis? *Nature Reviews: Cancer*, 4, 891-899.

39. Guerriero,F., Musmanno.R., Lacagnina,V. and Pecorella,A. (2001) A class of label-correcting methods for the K shortest paths problem. *Operations Research*, 49, 423-429.

40. Happel, J. and Sellers, P.H. (1982) Multiple reaction mechanisms in catalysis. *Industrial & Engineering Chemistry Fundamentals,* 21, 67-76.

41. Happel, J. and Sellers, P.H. (1989) The characterization of complex systems of chemical reactions. *Chemical Engineering Communications,* 83, 221-240.

42. Hellerstein, M. K. (2007) A critique of the molecular target-based drug discovery paradigm based on principles of metabolic control: Advantages of pathway-based discovery. *Metabolic Engineering* to appear, doi:10.1016/j.ymben.2007.09.003

43. Heinrich,R., Schuster, S. and Holzhütter, H-G. (1991) Mathematical analysis of enzymic reaction systems using optimization principles. *European Journal of Biochemistry*, 201, 1-21.

44. Heinrich,R., Montero, F., Klipp, E., Waddell, T.G. and Meléndez-Hevia, E. (1997) Theoretical approaches to the evolutionary optimization of glycolysis. Thermodynamic and kinetic constraints. *Euopean Journal of Biochemistry*, 243, 191-201.

45. Heinrich,R. and Ebenhöh,O. (2001) Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bulletin of Mathematical Biology*, 63, 21-55.

46. Henry, C.S., Broadbelt, L.J. and Hatzimanikatis, V. (2007) Thermodynamic-based metabolic flux analysis. *Biophysical Journal,* 92, 1792-1805.

47. Horne,A.B., Hodgman, T. C., Spence, H. D. and Dalby, A.R. (2004) Constructing an enzyme-centric view of metabolism. *Bioinformatics,* 20, 2050-2055.

48. Hüffner, F., Wernicke, S. and Zichner, T. (2007) FASPAD: fast signalling pathway detection. *Bioinformatics*, 23(3), 1708-1709.

49. Ihmels, J., Levy R. and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature Biotechnology*, 22(1):86-92.

50. ILOG CPLEX, (2005) http://www.ilog.com/products/cplex/news/whatsnew.cfm#cplex90

51. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabási, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, 407, 651-654.

52. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., De Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research,* 33, D428-D432.

53. Kamp, A.V. and Schuster, S. (2006) Metatool 5.0: fast and flexible elementary mode analysis. *Bioinformatics,* 22(15), 1930-1931.

54. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Research,* 26(1), 43-45.

55. Karp, P.D., Krummenacker, M., Paley, S.M. and Wagg, J. (1999) Integrated pathway/genome databases and their role in drug discovery. *Trends in Biotechnology,* 17(7), 275-281.

56. Karp, P., Riley, M., Paley, S. and Pellegrini-Toole, A. (2002a). The MetaCyc database. *Nucleic Acids Research,* 30(1), 59-61.

57. Karp, P., Riley, M., Saier, M., Paulsen, I., Paley, S. and Pellegrini-Toole, A. (2002b). The EcoCyc database. *Nucleic Acids Research,* 30(1), 56-58.

58. Kauffman, K. J., Prakash, P. and Edwards, J. S. (2003) Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5), 491-496.

59. Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R. and Ideker,T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences*, 100(20):11394-11399.

60. Klamt, S. and Stelling, J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports,* 29(1-2), 233-236.

61. Klamt, S. and Stelling, J. (2003) Two approaches for metabolic pathway analysis? *Trends in Biotechnology,* 21, 64-69.

62. Koch, I., Junker, B.H. and Heiner, M. (2005) Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics,* 21, 1219-1226.

63. Küffner, R., Zimmer,R. and Lengauer,T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16, 825-836.

64. Lee, J. M., Gianchandani, E. P. and Papin, J. A. (2006) Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 7(2), 140-150.

65. Liao, J.C, Hou, S. and Chao, Y. (1996) Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology and Bioengineering,* 52, 129-140.

66. Ma, H. and Zeng, A.P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19, 270-277.

67. Mathupala, S.P., Ko, Y.H. and Pedersen, P.L. (2006) Hexokinase II: cancer's double-edged sword acting as both facilitator and gatekeeper of malignancy when bound to mitochondria. *Oncogene* 25, 4777-4786.

68. Mavrovouniotis, M. L. (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical-compounds in aqueous-solution. *Biotechnol. Bioeng.* 36,1070–1082.

69. Mavrovouniotis, M. L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.* 266, 440–14445.

70. Mavrovouniotis, M.L. (1992a) Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 1. A synthesis approach in the context of simple examples. *Industrial & Engineering Chemistry Research,* 31, 1625-1637.

71. Mavrovouniotis, M.L. (1992b) Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 2. Formalization and analysis of the synthesis algorithm. *Industrial & Engineering Chemistry Research,* 31, 1637-1653.

72. Mavrovouniotis, M.L. (1993) Identification of qualitatively feasible metabolic pathways. In L. Hunter (Editor) *Artificial Intelligence and Molecular Biology*, 325-364, AAAI Press/MIT Press.

73. McShan,D.C., Rao,S. and Shah,I. (2003) PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*,19,1692-1698.

74. Meléndez-Hevia,E. and Isidoro,A. (1985) The game of the pentose phosphate cycle. *J. Theor. Biol.* 117, 251-263.

75. Meléndez-Hevia,E. and Torres,N.V. (1988) Economy of design in metabolic pathways - further remarks on the game of the pentose phosphate cycle. *J. Theor. Biol.* 132, 97-111.

76. Meléndez-Hevia,E. (1990) The game of the pentose phosphate cycle - a mathematical approach to study the optimization in design of metabolic pathways during evolution. *Biomedica Biochimica Acta* 49, 903-916.

77. Meléndez-Hevia, E., Waddell, T. G. and Montero, F. (1994) Optimization of metabolism: The evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *J. Theor. Biol.* 166, 201-220.

78. Meléndez-Hevia, E., Waddell, T.G and Cascante, M. (1996) The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *J. Mol. Evol.,* 43, 293-303.

79. Meléndez-Hevia, E., Waddell, T.G., Heinrich, R. and Montero, F. (1997) Theoretical approaches to the evolutionary optimization of glycolysis. Chemical analysis. *Eur. J. Biochem.* 244, 527-543.

80. Milner, P.C. (1964) The possible mechanisms of complex reactions involving consecutive steps. *Journal of Electrochemical Society*, 111, 228-232.

81. Mittenthal,J.E, Yuan A., Clarke B. and Scheeline A. (1998) Designing metabolism: Alternative connectivities for the pentose phosphate pathway. *Bull. Math. Bio.* 60, 815-856.

82. Nelson, D.L. and Cox, M.M. (2005) Lehninger Principles of Biochemistry. Fourth edition, Worth Publishers, New York.

83. Overbeek, R., Larsen, N., Pusch, G.D., Souza, M.D., Selkov Jr, E., Kyrpides, N., Fonstein, M., Maltsev, N. and Selkov, E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research,* 28(1), 123-125.

84. Palsson, B.O., Price, N.D., Papin, J.A. (2003) Development of network-based pathway definitions: the need to analyze real metabolic networks. *Trends in Biotechnology*, 21(5), 195-198.

85. Palsson, B.O. (2006) Systems Biology: Properties of Reconstructed Networks. *Cambridge University Press*.

86. Papin, J.A, Price, N.D. and Palsson, B.O. (2002) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Research,* 12, 1889-1900.

87. Papin J.A**.,** Price N.D**.,** Wiback S.J**.,** Fell D.A**.** and Palsson B.O. (2003) Metabolic pathways in the post-genome era. *Trends in Biochem. Sci.* 28, 250-258.

88. Papin, J.A., Stelling, J., Price, N.D., Klamt, S., Schuster, S. and Palsson, B. O. (2004) Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8):400-405.

89. Pharkya, P., Burgard, A. P. and Maranas, C. D. (2004) OptStrain: a computational framework for redesign of microbial production networks. *Genome Research*, 14(11), 2367–2376.

90. Pharkya, P. and Maranas, C. D. (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Met. Eng.*, 8, 1-13.

91. Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J.C., Montero, F. and Schuster, S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics,* 15, 251-257.

92. Planes, F.J. and Beasley, J.E. (2007) Path finding approaches and metabolic pathways. To appear in *Discrete Applied Maths*.

93. Poolman, M.G., Fell, D.A. and Raines, C.A. (2003) Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *European Journal of Biochem*istry, 270, 430-439.

94. Price, N.D., Papin, J.A, Edwards, J.S. and Palsson, B.O. (2002) Determination of redundancy and systems properties of the metabolic network of helicobacter pylori using a genome-scale extreme pathway analysis. *Genome Research,* 12, 760-769.

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

96. Rahman,S.A., Advani,P., Schunk,R., Schrader,R. and Schomburg,D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), *Bioinformatics*, 21, 1189-1193.

97. Rahman,S.A. and Schomburg,D. (2006) Observing local and global properties of metabolic pathways: ‘load points’ and ‘choke points’ in the metabolic networks. *Bioinformatics*, 22, 1767-1774.

98. Reed, J.L., Vo, T.D., Schilling, C.H. and Palsson, B.O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology,* 4(9), R54.

99. Schilling,C.H.*,* Schuster, S., Palsson, B.O. and Heinrich, R. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* 15, 296-303.

100. Schilling, C.H., Letscher, D. and Palsson, B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology,* 203, 229-248.

101. Schilling, C.H. and Palsson, B.O. (2000) Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. *Journal of Theoretical Biology,* 203, 249-283.

102. Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S and Palsson, B.O. (2002) Genome-scale metabolic model of Helicobacter pylori 26695. *Journal of Bacteriology,* 184(16), 4582-4593.

103. Schomburg I., Chang A. and Schomburg D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Research*, 30(1), 47-49.

104. Schuster, S. and Hilgetag, C. (1994) On elementary flux modes in biochemical reaction system at steady state. *Journal of Biological Systems* 2, 165-182.

105. Schuster, S., Fell, D.A. and Dandekar, D. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology,* 18, 326-332.

106. Schuster, S., Hilgetag, C., Woods, J.H. and Fell, D.A. (2002a) Reaction routes in biochemical reactions systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology,* 45, 153-181.

107. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T. (2002b) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics,* 18(2), 351-361.

108. Schwarz, R., Musch, P., von Kamp, A., Engels, B., Schirmer, H., Schuster, S. and Dandekar, T. (2005) YANA – a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6, 135.

109. Scott, J., Ideker, T., Karp, R.M. and Rhoden, S. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2):133-144.

110. Segrè, D., Vitkup, D. and Church, G. M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *P. Natl. Acad. Sci. USA* 99(23), 15112-15117.

111. Selkov, E.J., Grechkin, Y., Mikhailova, N. and Selkov, E. (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Research,* 24(1), 26-28.

112. Selkov, E.J., Grechkin, Y., Mikhailova, N. and Selkov, E. (1998) MPW: the Metabolic Pathways Database. *Nucleic Acids Research,* 26(1), 43-45.

113. Seressiotis, A. and Bailey, J.E. (1986) MPS: An algorithm and data base for metabolic pathway synthesis. *Biotechnology Letters,* 8, 837-842.

114. Seressiotis, A. and Bailey, J.E. (1988) MPS - An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnology and Bioengineering,* 31, 587-602.

115. Shlomi, T., Berkman, O. and Ruppin, E. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *P. Natl. Acad. Sci. USA*, 102(21), 7695-7700.

116. Shlomi, T., Segal, D., Ruppin, E. and Sharan, R.(2006) QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199.

117. Steffen, M., Petti, A., Aach, J., D'haeseleer, P. and Church, G. (2002) Automated

modelling of signal transduction networks. *BMC Bioinformatics*, 3:34.

118. Stephani,A. and Heinrich,R. (1998) Kinetic and thermodynamic principles determining the structural design of ATP-producing systems. *Bull. Math. Bio.*, 60, 505-543.

119. Stephani, A., Nuno, J.C. and Heinrich, R. (1999) Optimal stoichiometric designs of ATP-producing systems as determined by an evolutionary algorithm. *Journal of Theoretical Biology*, 199(1), 45-61.

120. Schwarz, R., Musch, P., Kamp, A.J., Engels, B., Schirmer, H., Schuster, S. and Dandekar, T. (2005) YANA – a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6:135.

121. Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W.M., Siezen, R.J. and Smid, E.J. (2006) Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *Journal of Biological Chemistry,* 281(52), 40041-40048.

122. Thiele, I., Vo, T. D., Price, N. D. and Palsson, B. O. (2005) Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an *in silico* genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.,* 187(16), 5818–5830.

123. Urbanczik, R. and Wagner, C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics,* 21, 1203-1210.

124. Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. Ser. , B* 268, 1803-1810.

125. Wilback, S.J. and Palsson, B.O. (2002) Extreme pathway analysis of human red blood cell metabolism. *Biophysical Journal,* 83, 808-818.

126. Xu, R., Pelicano, H., Zhou, Y., Carew, J.S., Feng, L., Bhalla, K.N., Keating, M.J. and Huang, P. (2005) Inhibition of glycolysis in cancer cells: a novel strategy to overcome drug resistance associated with mitochondrial respiratory defect and hypoxia. *Cancer Res.,* 65, 613-621.

127. Yeung, M., Thiele, I. and Palsson, B.O. (2007) Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics*, 8:363.

128. Zevedei-Oancea, I. and Schuster, S. (2003) Topological analysis of metabolic networks based on Petri net theory. *In Silico Biology,* 3, 0029 http://www.bioinfo.de/isb/2003/03/0029/main.html