Optimised Analysis and Visualisation of Metabolic Data Using Graph Theoretical Approaches

John M. Easton

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Electronic, Electrical and Computer Engineering
College of Engineering and Physical Sciences
The University of Birmingham
27th August 2009

ABSTRACT

Since the completion of the Human Genome Project in 2003, it has become increasingly apparent that while genomics has a major role to play in the understanding of human biology, information from other disciplines is necessary to explain the web of interacting signals that allow our bodies to function on a day to day basis and respond to rapid changes in our local environment. One such field, that of metabolomics, focuses on the study of the set of low molecular weight compounds (metabolites) involved in metabolism. Metabolomic studies aim to quantify the concentrations of each of these compounds within a subject under particular conditions, resulting in either information on the physiological effects of a disease or environmental factor (such as a toxin) on the organism, or the identification of metabolites or groups of metabolites that serve as biochemical markers for a state or illness.

Whilst metabolite concentrations alone can give great insight into a chosen state, additional information can be obtained by considering the ways in which metabolites interact with each other as parts of a larger system. One method of tackling this problem, metabolic networks, is gaining popularity within the community as it offers a complementary approach to the traditional biological method for studying metabolism, the metabolic pathway. Construction methods are varied; ranging from the mapping of experimental data onto pathway diagrams, through the use of correlation-based techniques, to the analysis of time-series data of metabolic fluxes. However, while many attempts have been made to capture and visualise the complex web of reactions within an organism, few have yet succeeded in showing how they can be used to help identify the metabolites that are most significantly involved in the differences between groups of biological samples.

This thesis discusses ways in which graphs may be used to aid researchers in both the visualisation and interpretation of metabolomic datasets, and provide a platform for more automated analysis techniques. To that end, it first presents the background to the relevant topics, metabolomics and graph theory, before moving on to show how metabolic correlation networks can be used to identify and visualise differences in metabolism between groups of subjects. It then introduces Linked Metabolites, a software package that has been developed to help researchers explain differences in metabolism by highlighting relationships between metabolites within the metabolic pathways, and to compile those relationships into directed metabolic graphs suitable for analysis using metrics from graph theory. Finally, the thesis explains how the directed metabolic graphs produced by Linked Metabolites could potentially be used to integrate data gathered from the same sample using different experimental techniques, refining the areas of the underlying biochemistry needing further investigation.

ACKNOWLEDGEMENTS

I'd like to thank my supervisors, Theo Arvanitis, Andrew Peet and Mark Viant for their assistance and patience during what seemed, at times, to have been a very long project.

I'd also like to thank Alex Gibb for proof-reading my (badly typed) drafts, Greg Reynolds and Robert Foster for providing the excellent eethesis LaTeX template, my family for putting up with my near-constant grumbling about work over the last five years, and Lisa Harris for providing the motivation for the creation of the Linked Metabolites software.

I'd like to thank the EPSRC for funding the first three years of this work, along with Prolego Technologies Ltd and the European Union InLOT and HealthAgents projects for providing me with employment after that.

Finally, I'd like to thank all those who have collected the data used in this thesis, in particular Andrew Southam from the School of Biosciences.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Metabolomics, the study of the set of low molecular weight compounds involved in metabolism, is an ideal tool for the monitoring the physiological state of a cell or organism. In particular, its ability to be used non-invasively either *in vitro* on samples of biofluids (for example blood, urine, or breath condensates) or *in vivo* by Magnetic Resonance Spectroscopy (MRS), mean that it has strong potential in fields such as medical screening, toxicology and environmental monitoring.

In order to separate groups of subjects within a study (for example diseased and healthy) many metabolomics experiments aim to identify a list of key compounds, known as biomarkers, whose concentrations differ substantially between the groups. The identification of biomarkers is often performed using statistical tools such as Principal Components Analysis (PCA) and can result in a long list of potential biomarkers, which then have to be checked by hand in order to remove "the usual suspects"[1], compounds that are involved in processes such as cell death and will appear to alter in many different diseases.

One way in which biomarkers are validated is through identifying their roles in important metabolic processes, such as the breaking down of sugars releasing energy (glycolysis). Each process, known as a metabolic pathway, has been identified through experimentation over a number of years and is commonly presented as a metabolic pathway diagram[2]. Unfortunately, while individual metabolic pathways can be interpreted reasonably easily by eye, the complete metabolic network in a cell or organism is made up of over one hundred pathways, all of which are interconnected. As a result of this, the validation of biomarkers is a difficult and time consuming task.

1.1 Aim and objectives

This thesis will investigate the relationship between biomarkers and the underlying biochemical processes as represented by the metabolic pathways. As part of this we will explore how experimental data can be visualised in ways that enable the relationship between metabolites to be easily determined, how automated searching of the metabolic pathways might be implemented and used to reduce the complexity of biomarker validation, and how, in the future, data from other -omics approaches, such as genomics or transcriptomics, might be combined and used to further refine the search space. All of this work is presented as a step towards the long-term goal of providing a platform for the application of techniques from graph theory to search and pattern-matching problems in biological networks.

1.2 Thesis organisation and contributions

Chapter 2 provides the background to the area, beginning with a brief discussion of metabolomics along with some examples of its use and moving on to discuss the challenges facing the field over the next few years. It then provides an introduction to graph theory and a selection of different graph structures & metrics, before finally discussing the metabolic pathways and methods for constructing different types of biological networks.

Chapter 3 explains how the background provided in the previous chapter relates to the topic of the thesis, and formulates the problem it will address.

Chapter 4 gives an example of how correlation networks, built from experimental data, were used to investigate the metabolic changes taking place in liver tumours in dab, a common flatfish. It shows how the networks allow the identification of possible relationships between compounds and how they also allow the researcher to see where changes in those relationships are taking place between sample groups. It also serves as an example of how metabolomics practitioners then go on to relate compounds that may be potential markers to the metabolic pathways, in an attempt to explain the biochemical reasons for their differing behaviour between groups.

Chapter 5 introduces Linked Metabolites, a software package that allows the user to combine metabolic pathways and then search them for paths between two groups of compounds, forming what we call directed metabolic graphs. Linked Metabolites is the major contribution of this work to the field, and is intended to assist researchers with the most time-consuming section of the biomarker validation process; explaining how compounds seen to be acting as potential biomarkers are related in the underlying biochemistry. The directed metabolic graphs it produces also provide a basis for the extension of this work with metrics from graph theory. The chapter then goes on to discuss different ways in which compounds might be grouped based on the experimental data, and compares the performance of Linked Metabolites against that of a similar tool, KEGG PathComp. Finally, it provides an example of the use of Linked Metabolites in the investigation of a set a ratios that can be used to differentiate between three types of childhood cerebellar tumour.

Chapter 6 discusses how a graph-based approach to metabolomics might fit into a larger, integrated genomics study. It gives some examples of how directed metabolic graphs can be built from gene expression (microarray) data, which could then, in combination with suitable graph alignment techniques, be used to refine the graphs produced by Linked Metabolites.

Chapter 7 then summarises the work, discusses how it addresses the original problem, and speculates on how it could be extended in the future.

CHAPTER 2

BACKGROUND

This chapter presents the background for the thesis. It begins with a discussion of metabolomics, including examples of its applications in both bioscience and medicine. It then goes on to introduce graph theory, discussing a number of different graph topologies and the metrics that can be used to classify them. Finally, it brings the two topics together, describing the ways in which metabolomics is suited to a graph theory-based analysis approach; this is done with reference to the traditional method of presenting metabolic knowledge, the metabolic pathway diagrams, and a method of exploring metabolomic datasets, metabolic correlation networks.

2.1 Metabolomics

In 2005, Hatzimanikatis *et al.* began a paper by stating that “Metabolism, the network of chemical reactions that make life possible, is one of the most complex processes in nature.”[3] It is unlikely that there are many who would dispute this statement; estimates suggest for example, that as many as 200,000 different metabolites (the low molecular weight compounds that act as substrates or products in metabolic reactions) may exist within the plant kingdom[4], although this figure is much lower for animals. The complete set of metabolites found within an organism is referred to as the metabolome, although it has been argued that since not all cells within an organism contain the same set of metabolites the term should actually refer to the set of metabolites encountered within a cell type[5]. Regardless of whether you subscribe to the local or global view of the metabolome, metabolomics (as with other -omics technologies such as transcriptomics for the transcriptome or proteomics for the proteome) is simply the measurement of the concentrations of its constituent compounds with the aim of determining the physiological, developmental or pathological state of a cell, tissue, organ or organism[6].

A typical metabolomics experiment involves the following steps:

1. Samples, either tissues, biofluids (blood, urine) or gas condensates (often breath samples collected as the subject exhales) are gathered and prepared as appropriate for the technique to be used.

2. High-throughput technologies such as Nuclear Magnetic Resonance (NMR) or Mass Spectrometry (MS), which may be used in conjunction with separation techniques such as liquid or gas chromatography to improve their resolution, are then used to record spectra for the samples.

3. Peaks in the recorded spectra may be identified and assigned to particular metabolites. Whether or not this step takes place is determined at least in part by the aim of the study; be it a metabolomics experiment in the canonical sense, where changes in the metabolome are quantified and interpreted in terms of the underlying biochemical processes, or a metabonomics experiment, in which traditionally the goal was the identification of a fingerprint for a particular condition or state[7].

4. Multivariate statistical analysis techniques such as Principal Components Analysis (PCA) are applied to the data to determine whether groups of samples within the dataset (for example diseased and healthy) can be distinguished from one another and which metabolites

are primarily responsible for the separation.

In recent years, despite the availability of other, more established -omics approaches such as transcriptomics (the measurement of gene expression via mRNA) and proteomics (the levels of proteins within a sample), there has been a steady increase in the use of metabolomics within the biological and medical communities. There are several reasons for this growth. Firstly, it should be noted that far from being in competition with each other, -omics technologies are complementary. Cancer, for example, is often ultimately the result of a genetic problem and a genomics-based technique such as a microarray analysis would help to establish this; however it would not show us how the genetic change was reflected in the function of the sufferer's cells, which is important when considering potential treatments or monitoring their effects. Secondly, after the initial purchase of a spectrometer a metabolomics analysis can cost as little as US$1 per sample, this compares very favourably with the cost of transcriptional or proteomic approaches[8]. Finally, metabolomics is a reasonably portable technique, since while gene and protein sequences vary between species, the metabolites are largely the same. Thus, rather than trying to measure the amount of a particular enzyme, requiring the design of protein microarrays with the appropriate sequences for each organism being studied, it may be much more convenient to measure the effect on the concentrations of the substrates and products of reactions that the enzyme catalyses[9].

2.1.1 Applications of metabolomic datasets

While metabolomics has been applied to a range of different fields, two of the most interesting areas are environmental studies (such as the work on dab liver tumours, see chapter 4), which use metabolomics to investigate the response of an organism to natural or anthropogenic stresses under real or simulated environmental conditions[10], and medical applications such as drug toxicity, disease diagnosis and the monitoring of response to treatments.

Much of the early metabolomics work focused on plants. In 2000, Fiehn *et al.*[4] published a paper that demonstrated the use of metabolic profiling in functional genomics. In their study, which involved four different strains of Arabidopsis (two different 'parent' ecotypes and a mutated version of each), they showed that by performing PCA on the relative concentrations of 326 metabolites it was possible to separate the samples into their four groupings (although the separation was much better between the parent strains than between the parents and their children).

In 2003, Viant *et al.*[11] published the first application of NMR-based metabolomics to the aquatic environment. Their study, which focused on Withering Syndrome, a fatal disease that had already decimated black abalone (a type of edible sea snail) populations in California, showed that it was possible to distinguish between healthy and diseased red abalone, as well as those with stunted growth, using metabolomic data from samples of the animal's foot muscle, digestive gland and hemolymph tissue. The study also identified several potential biomarkers for the condition, including decreased levels of phenylalanine, tryptophan, tyrosine & glycine, and increased levels of formate and homarine.

In 2004, Bundy *et al.*[12] performed a study to determine if an NMR metabolomic analysis of relative histidine concentrations, which had previously been identified as a potential biomarker for copper contamination in semi-field-scale trials, could be effectively used in the field as an indicator of metal contamination. The study involved three species of earthworm (as does much of the literature in the area owing to their relative immobility and close contact with their surround environment[13]), two native and one imported, spread across six sites with varying degrees of contamination. No species of worm was at all six sites. While the experiments did show that separations between groups of worms from the different sites could be created based on PCA for all three species, the degree of contamination at which the separation took place differed for each species. The analysis also identified maltose as a potential biomarker for metal contamination, although while one of the native species of worm showed an increasing relative concentration of maltose in the more highly contaminated sites, the other species showed a decreasing level of maltose. This is a good illustration of the need to apply biomarkers carefully due to differences in metabolism between species.

Metabolomics has a role to play in a whole range of healthcare-related fields; in disease prevention, through studies of how diet and nutrition impact on various forms of illness[14], in disease diagnosis, where biomarkers for particular conditions can help with their rapid diagnosis, and in treatment, both through drug development and selection and through the monitoring of responses to therapy.

While the primary aim of any preclinical toxicology study has to be to ensure the safety of any recipients of a new treatment, an important secondary consideration is that "the later that a molecule or molecular class is lost from the drug development pipeline, the higher the financial cost"[15]. As such, preclinical toxicology studies require a careful balance to be struck between assuring the rapid loss of dangerous drugs and permitting those that have clinical benefits to pass through the development process. Toxicity studies typically generate clearly definable endpoints such as clinical signs or histopathology and can be arrived at reasonably quickly due to high dosage; however, these studies tend to take place after the *in vivo* testing of the effectiveness of the compound and those effectiveness tests seldom have as many robust endpoints that could be used as early warnings of potential toxic effects and frequently take several weeks to complete[1]. Metabolomics can therefore, play an important role in the drug development process by allowing the early screening of biofluid samples from the efficacy test animals for biomarkers of toxic effects. While these markers may not themselves be sufficient evidence to stop the development of a drug, they can be used to suggest target organs for later full toxicology studies. Metabolomics also has a role to play in the later, testing phases of the drug development process. Here, biofluid samples from the test subjects can be used to ensure the absence of toxic effects, helping to avoid situations such as that seen in 2006 when six volunteers suffered multiple organ failures after taking part in the trial of a new anti-inflammatory drug[16].

Although the non- or minimally-intrusive nature of metabolomics of biofluids lends itself to use in clinical practice, a great deal of preclinical work has also been done to try and identify biomarkers and understand the metabolic processes going on in diseases. Metabolomics has for example been used to distinguish between different types of tumour cell lines for a number of years[8]. The question remains however as to how reflective of a tissue an extracted sample for metabolomic analysis can be when it has potentially undergone any number of non-biological reactions associated with the extraction procedure[1]. One variant of NMR metabolomics, Magic Angle Spinning (MAS) avoids this problem by allowing the experiment to be performed on a small sample of intact tissue, rather than an extract. This has the advantage that all the metabolites are still in-situ within the cellular structures that normally contain them and the risk of them undergoing additional reactions is much reduced. They can also be related to the histology[15]. Magic Angle Spinning also allows water and lipid soluble metabolites to be observed simultaneously, a process that would otherwise require separate extraction procedures[8].

Rapid and unobtrusive disease diagnosis and screening is one of the most attractive features of metabolomics in a clinical setting and much work has been published on the topic. In 2002 for example, Brindle *et al.*[17] showed that it was possible to diagnose and determine the severity of coronary heart disease in patients using NMR of serum samples. In 2005, Kenny *et al.*[18] identified a pair of rules involving three key metabolite concentrations that could be used to diagnose pre-eclampisa from the blood plasma of pregnant women; early diagnosis is of great importance in such cases as the potential therapies are most likely to be effective before the clinical presentation of the disease. Coen *et al.*[19] have shown that it is possible to distinguish patients with bacterial meningitis from those with viral meningitis based on NMR of cerebrospinal fluid, a step that can be key in ensuring the patient receives the correct treatment and therefore has the best chance of survival. Finally, Carraro *et al.*[20] have shown that exhaled breath condensate can be used to distinguish children with asthma from those without.

Abnormal brain masses, both malignant and benign, pose difficult diagnostic problems and require critical therapeutic decisions that can involve risk and possible damage to the patient[21]. This is a particularly important issue in children, where brain tumours are a leading cause of cancer-related death and the tumours seen differ from those in adults[22]. While Magnetic Resonance Imaging (MRI) is widely used for determining the extent of a mass and to provide an initial diagnosis, surgical biopsies are still the gold standard for diagnosis[23]. However, not all masses are cancerous[21] and many that are display a high degree of heterogeneity meaning that a histopathological diagnosis from a single site may not fully characterise the bulk of the tumour[23]. As such, there is a need within oncology for a non-invasive technique that can serve as a guide for biopsy procedures or provide indications as to whether abnormal masses are likely to be cancerous to assist with treatment planning and reduce the need for unnecessary surgery.

Magnetic Resonance Spectroscopy (MRS) allows a low-resolution NMR metabolomics experiment to be performed non-invasively on a patient's tissues using a conventional MRI scanner. While it had been shown as early as the mid-1990s that MRS could be used to distinguish between the four most common types of brain tumour (meningioma, astrocytic, oligodendroglioma & metastasis, and cysts) in adults using pattern recognition techniques[24, 21], its adoption into routine clinical usage has been hampered by a lack of available data and of simple rules for diagnosis that can be easily applied by radiologists. These issues are now beginning to be addressed, both through work such as that by Harris *et al.*[25] which uses ratios of the heights of key points in a spectrum to distinguish between different types of childhood cerebellar tumour with parameters that can be easily updated as more data becomes available (this work is the basis for the example usage of Linked Metabolites, see chapter 5), and through multi-centre databases such as that being produced by the eTumour project[26] or distributed database and classification systems such as HealthAgents[27]. Even so, despite its potential, data obtained from MRS can not yet be used independently because of significant overlap and non-specificity of the results[28].

2.1.2 Challenges facing metabolomics

The metabolism of an organism (as we have seen from the applications to environmental studies) is extremely sensitive to the environment around it (factors such as availability of food, temperature and local pollutants) as well as to its own biological characteristics (age, sex etc.). As such, many metabolomics studies suffer from noise due to biological variation and experiments must be carefully planned to try and ensure that all the subjects are as closely matched as possible in every respect save the factor under investigation. While such approaches can help to ensure that the biomarkers identified are indeed due to the factor under investigation, they also introduce the risk that the markers may only be applicable under a very narrow subset of conditions and in a particular location. It has been suggested that these difficulties can be overcome by defining a Normal Metabolic Operating Range (NMOR) for an organism at a particular site, that being an area of the multi-dimensional metabolic space that contains 95% of the individuals of a population at that site and then describing markers as a deviation relative to that space[10]. Alternatively, it has been proposed that composite biomarkers should be utilised, groups of changes that together can be used to characterise the change in metabolism, thus reducing the effect of noise on any one metabolite concentration due to external factors[29]. Even so, much work remains to be done in this area.

Metabolomics practitioners must also address the problem of “the usual suspects”[1]; compounds that frequently appear in the literature as biomarkers for a huge range of conditions. These compounds, which include molecules such as citrate, creatine, glucose and lactate are actually often related to other cellular processes, such as energy metabolism or apoptosis, and hence frequently appear altered in organisms experiencing some type of stress. Thus, while these compounds may indeed be altered in the conditions under investigation, they are an important reminder that metabolomics needs to focus on the most specific markers to a particular condition, and that the metabolites driving pattern separations in PCA are not necessarily the most interesting or appropriate for use as biomarkers[1]. The natural solution to this problem lies in relating the biomarkers to the underlying biology of the system (see chapter 5), both by placing them in the context of biological processes that are already known to be involved in the condition under investigation, and potentially by using the presence of particular markers to infer new knowledge about a particular condition based on biological processes that have been investigated elsewhere. Traditionally, these comparisons are performed using a set of graphs known as the metabolic pathway diagrams (for a description of the metabolic pathways, see section 2.2.6), as such the following sections will provide a brief introduction to graph theory before returning to the biology in order to discuss some of the ways in which experimental data can be presented as a graph.

2.2 Graphs as representations of real-world systems

2.2.1 What is a graph?

In mathematics a graph is a collection of points (commonly referred to as nodes or vertices) that are connected by a set of links (either called edges if the link is bidirectional or arcs if the link only applies in a single direction). More formally a graph $G = (V, E)$, is a mathematical structure consisting of two sets V and E[30]. The elements of the set V are the nodes (or vertices) of the graph, while the elements of the set E are pairs of nodes representing the start and end points of the edge. If the graph is directed, then the elements of E are an ordered pair; in this case the directed edge (or arc) runs from the first node in the pair (the source) to the second (the target).

In real-world contexts, graphs are often, although by no means exclusively, used to represent systems where some kind of transfer is involved; these include communications networks where nodes (senders or receivers) exchange information, power distribution networks, in which nodes commonly represent power stations, distribution units such as sub-stations or consumers and the edges represent power lines, and transportation networks, where nodes commonly represent stations, airports or cities and edges represent the tracks, flight paths or roads between them.

Graph-based models are commonly used to solve a wide range of problems, from network routing to epidemiology and the determination of the most efficient order for robots to place components onto printed circuit boards.

2.2.2 The bridges of Königsberg

Graphs have been used in the mathematical modeling of real-world systems since the early eighteenth century. The first recognised use of a graph as a modeling technique was in the

1730's when the Swiss-born mathematician Leonhard Euler used it as the basis for his solution to the Königsberg bridge problem. The problem is as follows: in the city of Königsberg, on the banks of the Pregel, there are seven bridges (Figure 2.1). Is it possible for a man to walk across all seven bridges and never cross the same bridge twice?

Euler's solution to the problem was beautifully simple, rather than taking the approach used in coffee shops throughout the city and attempting to find a path that crossed all seven bridges, he decided to try to prove whether or not such a path could ever be found. In order to achieve this he drew out the map of the city in a simplified form, as a graph (Figure 2.2), to which he applied the following reasoning:

1. Nodes in a graph that have an odd number of links must be either the starting point or the end point of a path.

2. A single, continuous path covering every edge in a graph can only have one starting point and one end point.

3. Therefore, no single, continuous path covering every edge in a graph can exist if that graph has more than two nodes with an odd number of links.

In the graph of the Königsberg bridges, Euler noted, there were four nodes with an odd number of edges and therefore there could be no route that crossed all of the seven bridges exactly once.

In the 140 years that followed the publication of Euler's solution to the Königsberg bridge problem, nobody ever found a route across the seven bridges. It wasn't until 1875, when the townsfolk built a new bridge between the banks of the river (thus making two of the four nodes even) that a walk fulfilling the criteria was finally found.

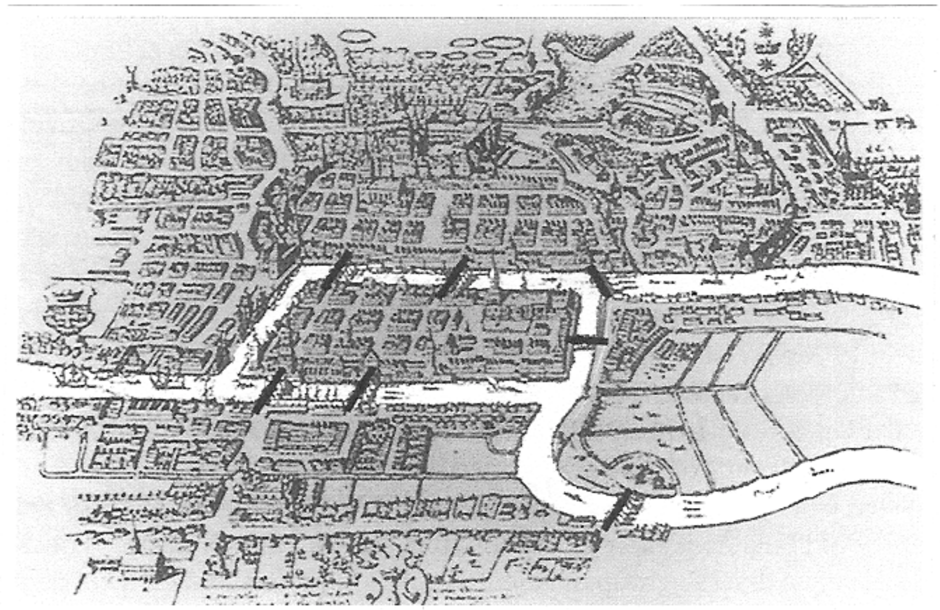


Figure 2.1: A map of the city of Königsberg in the eighteenth century. Reproduced from [31].

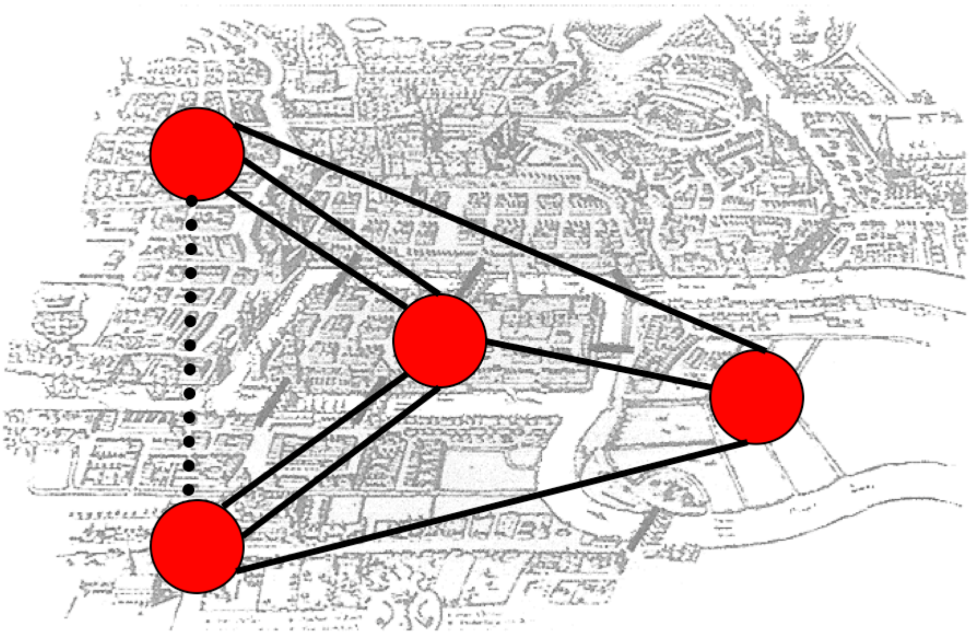


Figure 2.2: A graph representation of the bridges of Königsberg. The dotted edge represents the 1875 bridge.

2.2.3 Small worlds

By the mid 1960's the idea that how people were connected to each other on a local level had an important effect on the overall structure of the society as a whole was already well accepted by sociologists. The use of random graphs as models of large social systems was also beginning to become more popular (indeed sociology was one of very few areas outside of mathematics where graph theory was actively used as an analysis technique at the time). Initial work by Pool and Kochen[32], which was originally written in 1958 but didn't appear as a published work for another two decades, had suggested that in random graph models of human contact networks (where nodes represent individuals and links represent a relationship between them such as a friendship, business relationship or a familial tie) if each person knew, on average, 1000 other people then most people should be able to be linked to each other via a chain of at most two intermediaries. This result appeared to agree well with the reasonably common "small world phenomenon"; the experience of meeting a person who you believe to be a complete stranger and discovering that you have a common acquaintance. Such a discovery is often accompanied by the exclamation "It's a small world!"

In 1967, sociologist Stanley Milgram (who was by this point a rather controversial figure in the field following his research into the human response to authority figures) began an experiment designed to investigate the length of chains linking individuals in real social groupings. This was an experiment that he would repeat in greater detail and with a large sample two years later[33]. In his experiments Milgram asked groups of volunteers to attempt to pass a letter to a target person, a Boston stockbroker, identified by his name, address, occupation, place of work, college, year of graduation, military service dates, wife's maiden name and hometown. The only stipulation was that the letter should be passed only to a person known to the sender on a personal basis and whom the sender believed was more likely to know the target than the sender themselves. In addition to the passing of the letter, participants were asked to add their name to a list that accompanied the letter (to prevent the creation of loops) and also to send a card back to Milgram so that he could track chains that were incomplete as far as was possible.

The 296 initial volunteers for the 1969 study formed three groups, blue-chip stockholders from Nebraska, randomly selected individuals from Nebraska, and randomly selected individuals from the Boston area. Of these initial volunteers, approximately $\frac{2}{3}$ actually passed on the letter and therefore began a chain. Ultimately $\frac{1}{3}$ of the letters reached the target recipient.

The mean chain length across all three groups was 5.2 steps; while this is considerably larger than Pool and Kochen's figure of two, it is still small compared to the size of the population of the United States at the time (approximately 200 million). Perhaps more impressive is the fact that the figure didn't alter hugely with geographical area, with the Boston random group having a mean chain length of 4.4 compared to a mean chain length of 5.4 for the Nebraska stockholders and a mean chain length of 5.7 for the Nebraska random group. Milgram's results were the first study into what is now commonly referred to as the "six degrees of separation", although that term was never used by Milgram himself, it originated instead in John Guare's 1991 play of the same name[34].

In 1973, around five years after Milgram's experiments, Mark Granovetter published a paper based on research he had performed as part of his doctoral dissertation[35]. For his study, Granovetter questioned a number of professional, managerial and technical workers living in a Boston suburb; each of the subjects had recently changed job and Granovetter was interested in how they had come upon the information that led to their new positions. In those cases where the information had come through a contact, Granovetter asked the subjects how regularly they saw that contact based on the following scale: often = at least twice a week, occasionally = more than once a year but less than twice a week, rarely = once a year or less. Over half of the subjects reported that they only saw the contact occasionally and additionally, in many cases indicated that the source of the information was not someone in their current network of contacts. Instead the contact was commonly a former college friend or an ex business contact such as a former employer. When tracing the original source of the information back through the contact, Granovetter discovered that in the vast majority of cases (over 80%) the contact either knew the source of the information directly, or were within one intermediary of them. Granovetter's research showed two things; firstly that 'new' information within a social network is unlikely to come from those people that you interact with on a regular basis, those with whom you have a 'strong' tie, as it is likely that you have access to the same information sources. Secondly, that it is likely the person you do get such information from will be someone you only have a 'weak' tie to, an occasional contact who moves in another circle of people and who acts as a bridge between you and that group. It is that remote group who will have access to information that is different to your own. The presence of this type of bridge in social networks had two main implications; firstly, that the larger-scale, macroscopic structure of the system could be at least as important as the microscopic structure surrounding the subject, but also that the structure of some networks may differ in very significant ways from that of the random graphs that were being used to model them on a large scale.

MTV's "Jon Stewart Show" might seem an unlikely source for an academic idea to be brought squarely into the public eye; however it was on an edition of that programme in 1994 that three students of Albright College, Pennsylvania grabbed the attention of the audience with their ability to link the actor Kevin Bacon, who they claimed was the centre of the Hollywood universe*, to any other actor or actress suggested to them[34]. The "six degrees of Kevin Bacon" as the game became known is played as follows: first, select an actor or actress. The object of

*Kevin Bacon is not actually the best connected actor in Hollywood, that honour goes to Rod Steiger with Bacon ranking 1049th (based on the IMDB in June 2004).

the game is to link that person to Kevin Bacon via a chain of intermediaries they have worked with on the big screen. As an example, consider Star Trek's DeForest Kelly; Kelly worked with Christian Slater on "Star Trek VI" in 1991 and Slater would later work with Kevin Bacon on "Murder in the First" in 1995.

A few months after the programme aired, a web-based version of the game, "The Oracle of Bacon"[36] was created by two computer scientists from the University of Virginia. The website used information from the Internet Movie Database (IMDB)[37] to calculate the shortest path from an actor to Bacon. Borrowing from a similar game played by academics, in which people linked themselves to Paul Erdös via coauthorships and then assign themselves a number equivalent to the length of the shortest path[38], the site also came up with 'Bacon numbers' for each actor. In the earlier example, DeForest Kelly was linked to Kevin Bacon in two steps, hence he has a Bacon number of two. There are of course, plenty of other examples; Kyle MacLachlan for instance was in "Dune" with Max von Sydow, who was in "Minority Report" with Tom Cruise. Cruise acted alongside Kevin Bacon in "A Few Good Men" and so Kyle MacLachlan has a Bacon number of three. While such games are fun they are included here for a good reason; which is that just like Milgram's letters or Granovetter's contacts in the labour market, they display a surprisingly short average path length between individuals. According to "The Oracle of Bacon" the mean path length in the "six degrees of Kevin Bacon" is 2.946 (based on the IMDB in June 2004), while the mean path length for coauthorships with Paul Erdös is 4.65[39].

In 1998, Duncan Watts and Steven Strogatz proposed a new graph model, one which encapsulated both the highly clustered nature of regular lattices and the short average path length seen in random graphs[40]. The model was called the "small-world network", a reference to the small-world phenomenon that it displayed.

In order to generate small-world networks, Watts and Strogatz started with a standard ring lattice containing n nodes each with k edges. The networks were sparse, meaning that the number of nodes was far greater than the number of edges per node, however not so sparse that the graphs were in danger of becoming disconnected (having isolated fragments that are not connected to the rest of the structure). They then proceeded to rewire each edge at random, with a probability p. By altering the value of p, Watts and Strogatz were able to investigate how the degree of randomness they introduced into the graph affected two key parameters; the average path length, $L(p)$ and the clustering coefficient $C(p)$.

L is defined as the average shortest path length between all pairs of nodes in the graph. C is defined locally for a node v, with k_v neighbours (nodes connected to it by a single edge) and E_v edges between its neighbours by:

$$C = \frac{E_v}{\frac{1}{2}k_v(k_v - 1)} \tag{2.1}$$

This can then be averaged to give C for a graph containing n nodes:

$$C = \frac{\sum_{v=1}^{n} \frac{E_v}{\frac{1}{2}k_v(k_v-1)}}{n} \tag{2.2}$$

In regular lattices, there is no rewiring and therefore $p = 0$. In these cases the degree of clustering $C(0)$ is high, while the average path length $L(0)$ grows linearly with n. By comparison, in a random graph, where every edge is placed at random $p = 1$. The degree of clustering $C(1)$ for a random graph is low and the average path length $L(1)$ only grows logarithmically with n. While the logical assumption based on these boundary conditions might be that a large average path length always accompanies a large clustering coefficient and vice versa, Watts and Strogatz found that for a broad range of values of p, graphs were produced in which $L(p)$ was nearly as small as it is in random graphs and yet $C(p)$ was still considerably larger than in random graphs. The reason for this was that for small values of p, each rewiring, while only having a minor effect on the degree of clustering of the system as a whole, would reduce the distance not only between the source and the target nodes of the rewired edge but also between their neighbours, the neighbours of their neighbours, and so on. An important implication of this is that the change from the highly clustered lattice to a small-world would be almost imperceptible at a local level, since the degree of clustering remained largely unchanged.

Watts and Strogatz then checked their results by measuring L and C for three well-studied graphs; the network of collaborations between actors, the electrical power grid of the western United States, and the neural network of the nematode worm Caenorhabditis elegans. In all three systems they found the same characteristic low average path lengths combined with a high degree of clustering, demonstrating that far from being a fluke of social networks, the small-world phenomenon was probably common to many large, sparse networks in nature. The small-world property is displayed by the directed metabolic graphs in chapter 6.

2.2.4 Scale-free networks

In 1999, Réka Albert, Hawoong Jeong and Albert-László Barabási published the results of a study they had performed while investigating the topology of the World-Wide Web[41]. It should be noted at this point that the World-Wide Web, the network of interconnected documents that are commonly accessed through a browser such as Mozilla Firefox or Microsoft Internet Explorer, is distinct from the Internet, the hardware on which the documents are stored. In order to gather data on the topology of the web, Albert, Jeong and Barabási built a piece of software commonly known as a web spider; spiders gather information on web pages by recursively following URLs and adding the pages they discover to a database as they proceed. Using the topological information gathered by their spider (which they had set to explore the University of Notre Dame's domain, nd.edu), Albert, Jeong and Barabási began to investigate the in and out-degree distributions of the system (since a hyperlink is a directed concept, running from one page to another, each node will have a number of links that head into it, the in-degree of that node, and a number of links leaving, the out-degree). They discovered that the distributions of both the in and out degrees across the graph as a whole differed significantly from both those predicted by Erdös and Rényi for random graphs, which would be expected to form a Poisson distribution, and from that of the small-world networks of Watts and Strogatz. Instead, the distributions formed a power-law; meaning that for the vast majority of nodes the number of links was very small. As the number of links increased, the probability of a node having that number of links dropped logarithmically resulting in very few highly connected nodes, which dominated the topology of the graph.

On further investigation, Albert, Jeong and Barabási found that although the degree distribution of their data did not match with that of small-world networks, the data did form a small-world as it displayed the characteristic short average path length (11.2 steps for the nd.edu domain, which contained 325,729 nodes and 1,469,680 links). The reason for this was the few highly connected nodes (or hubs) in the system, which acted as short-cuts between otherwise distant parts of the graph. Albert, Jeong and Barabási's "scale-free" graphs, so called because of the several orders of magnitude over which the degree distributions displayed a power law tail, quickly became a topic of great research interest, with several groups worldwide reporting them in a diverse range of systems; including the routing topology of the Internet[42], metabolic networks[43], and social networks including the network of scientific collaborations[44], the network of human sexual contacts[45] and the terrorist cell responsible for the September 11^{th} attack on the World Trade Center in New York[46].

Scale-free networks appeared to be common in nature, but what caused them and their unusual degree distribution? As a follow-up to their work describing scale-free networks, Barabási and Albert published a paper arguing that a combination of network growth over time and preferential attachment could result in just such a graph[47]. Barabási and Albert noticed the following: in the graphs normally used to model large systems (either the random graphs of Erdös and Rényi or more recently the small-world networks of Watts and Strogatz) the size of the system is fixed at the time of its creation. A graph is generated, which contains N nodes, and these are then either connected with a given probability or formed into a lattice that is then rewired. In real-world systems this is seldom the case; the core of a network will form and it will then expand, with new nodes being added over time. They also noted that rather than being connected at random to other nodes in the graph, a new node, upon joining the network is more likely to be connected to some of the existing nodes than it is to others. Actors staring in their first movie for example, are likely to be working alongside more established figures rather than in a cast comprised entirely of newcomers. In the same way, a new website is far more likely to link to sites such as Google or Amazon than it is to a page describing sheep farming methods in the Ukraine, unless of course it is about a similar topic. In real-world systems it seemed, the "rich get richer".

Based on their observations, Barabási and Albert proposed the following growth model for scale-free networks (Figure 2.3 – Figure 2.6). Start with a small number of nodes, m_0. Now begin to add additional nodes to the graph one at a time. As each node is added create edges between it and a number of the existing nodes in the graph less than or equal to the number of nodes that

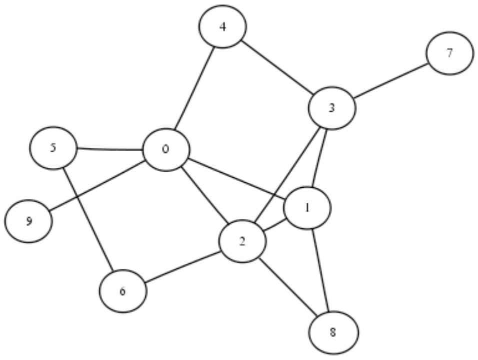


Figure 2.3: Early stage of a scale-free network. This network began with 2 nodes connected by a single edge. With each new node either 1 or 2 edges were added giving an effective m of 1.5.

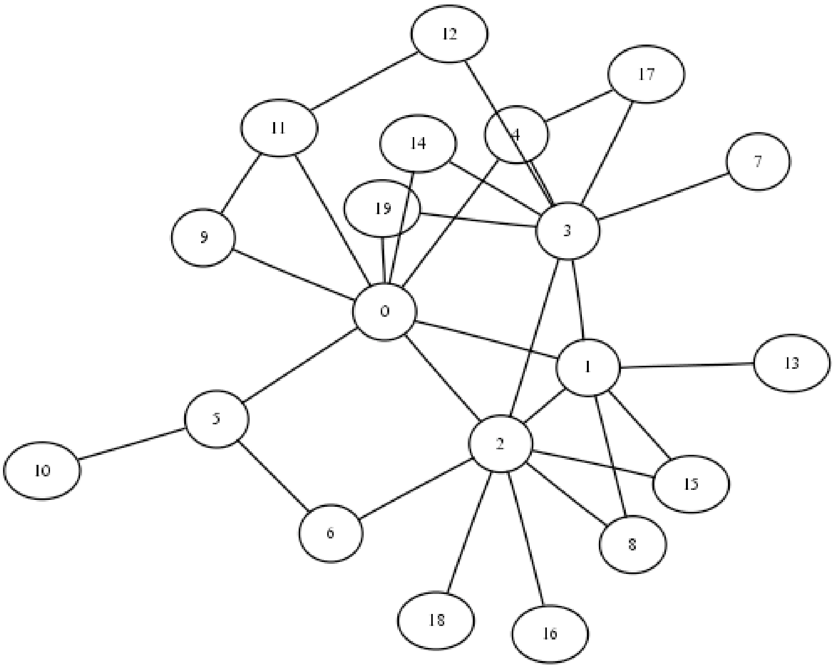


Figure 2.4: The network after 18 time-steps (now containing a total of 20 nodes). At this stage nodes 0, 1, 2 and 3 can be seen to be emerging as hubs.

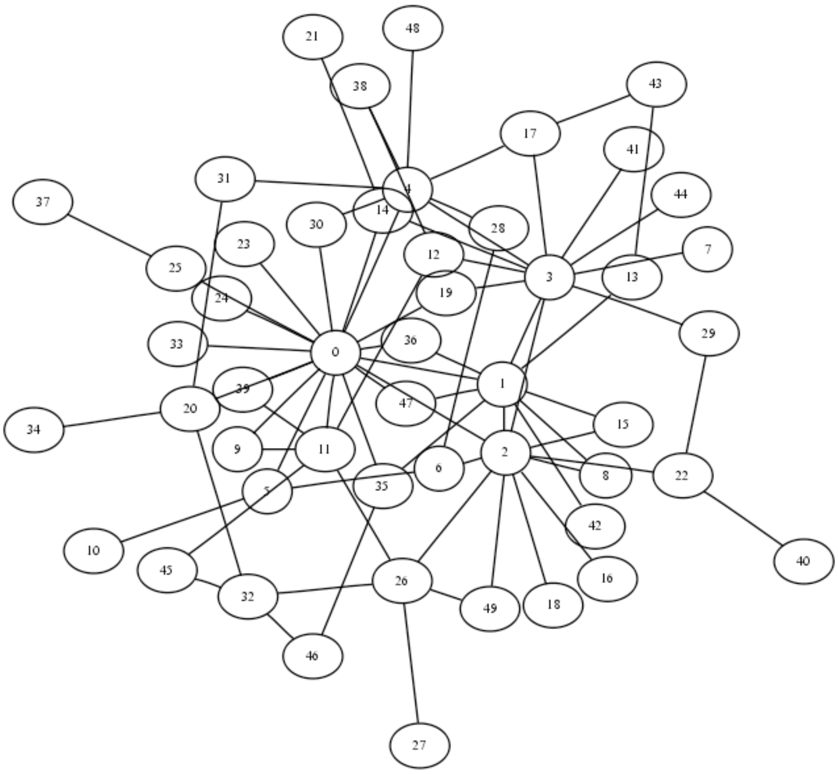


Figure 2.5: The network after 48 time-steps (a total of 50 nodes).

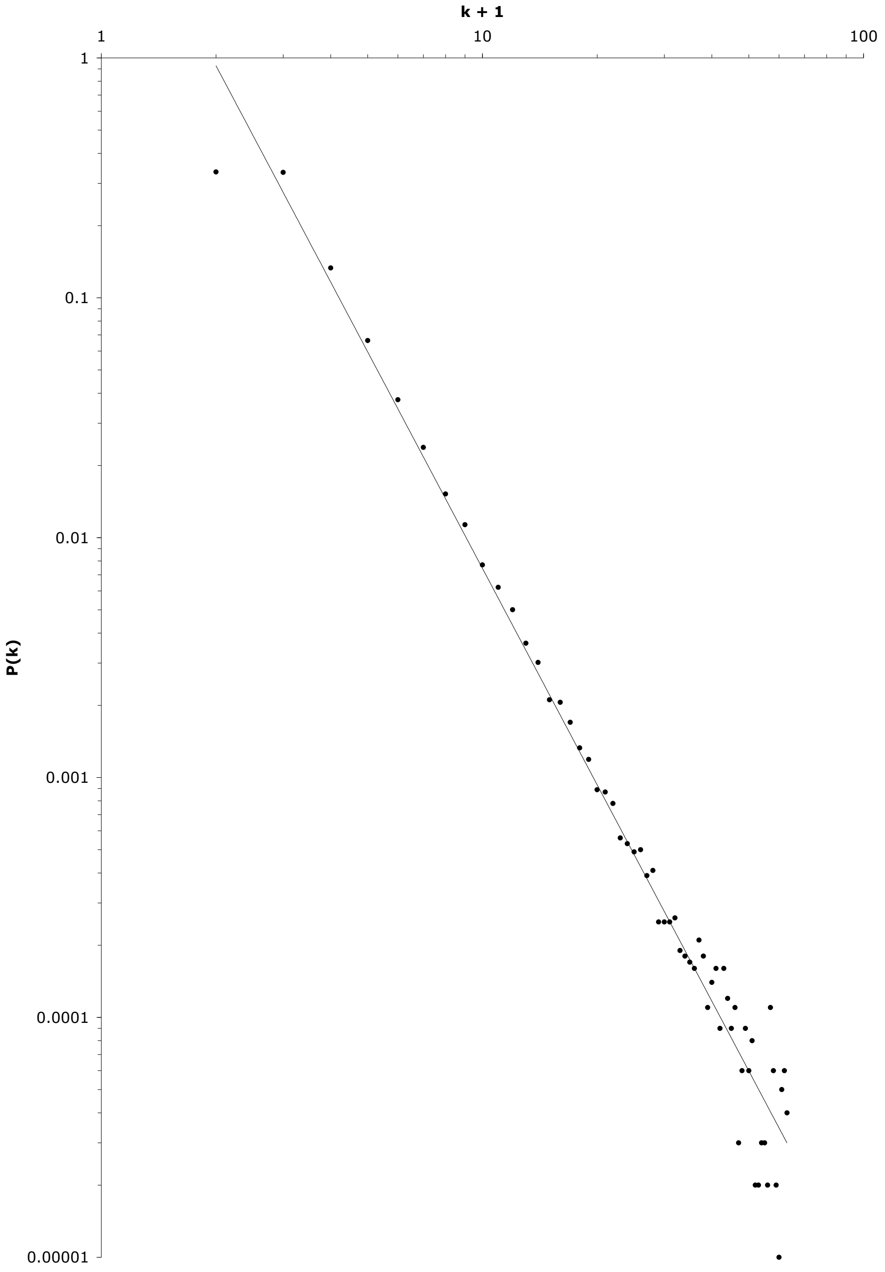


Figure 2.6: The probabilities of a node having degree k for $1 \leq k \leq 100$ as taken from the completed scale-free network of 100,000 nodes. The trendline is for a power law with the equation $P(k) = 7.4k^{-2.998}$

were initially placed. Select the nodes to which the edges will be attached based on a probability that is individually scaled, (for each node) according to its degree at that time. Specifically, at any given time t the graph contains n nodes and the probability of attachment to node i, which has degree k_i, is given by:

$$\prod(k_i) = \frac{k_i}{\sum_{j=1}^{n} k_j} \tag{2.3}$$

While Barabási and Albert's model does generate scale-free networks, it has problems. It does not state how to perform the preferential attachment in the case of the initial graph (where there are no edges) and it can only generate graphs where the exponent of the power law is 3 (in real-world systems this parameter can vary)[48]. A more mathematically rigorous process, based on Baraási and Albert's criteria, is the LCD model of Bollobás and Riordan[49].

The LCD growth model differs from Barabási and Albert's in several ways; firstly self-loops (edges that link a node to itself) are allowed, as are multiple edges between a pair of nodes. These features, which are not created by the Barabási-Albert model, can be seen in real-world networks in situations such as a webpage linking to itself (often found in contents tables for long documents) or as multiple links between two pages (for example references to another topic from different parts of an online encyclopedia article). In order to allow the creation of self-loops, the probability of attachment to each node in the graph has to take into account the contribution to the degree distribution made by the outgoing half of the edge being added. As such, the probability that a given node, $V(s)$ should be the target of one of the edges from the node being added in the current time-step, $V(t)$ is given by:

$$p(Edge(V(t), V(s))) = \begin{cases} \frac{d_{G_1^{t-1}} V(s)}{2t-1} & 1 \leq s \leq t-1 \\ \frac{1}{2t-1} & s = t \end{cases} \tag{2.4}$$

Where $d_{G_1^{t-1}}V(s)$ is the degree of node $V(s)$ in the graph at the previous time-step.

It should be noted that since the outgoing portion of each new edge being added to the graph counts towards the degree of the source node, $V(t)$ the probabilities for each node need to be recalculated as every edge is added and not just for each time-step as is the case in the Barabási-Albert model.

The second major difference between the two models is in the state of the initial graph; under the LCD model the initial graph is empty (it contains no nodes or edges) or it only contains a single node with a self-loop. This combined with the new formula for determining which node should be the target of the edge being added means that unlike in the Barabási-Albert model there is never any ambiguity over how the preferential attachment should take place (no nodes exist with degree zero).

Albert, Jeong and Barabási went on to study the vulnerability of the networks generated using the Barabási-Albert model to random failures and targeted attacks[42]. They found that in random graphs, the failure of a small fraction of the nodes caused a noticeable increase in the diameter of the graph (the length of the shortest path between the two most distant nodes in the graph) but that it made no difference to the rate of change of the diameter if the nodes that failed were selected at random or as the result of a targeted selection process. In scale-free networks by comparison, the effects of random failures were much less pronounced, however targeted removal of the most connected nodes caused a rapid increase in diameter. The effect of the removals on the structure also differed, with failures in the random graphs causing them to fragment at a given threshold, whereas failures in the scale-free networks initially only caused the size of the largest cluster to gradually decrease, with fragmentation only taking place when the largest cluster became very small. In a targeted attack scenario however, the effect of node removals on scale-free networks is similar to that observed in random graphs but it occurs at a much higher rate.

The reason for this difference in behaviour was the topology of the graphs. In random graphs, where the degree of each node is approximately the same, all nodes contribute equally to the network's connectivity. As such, the removal of a node will always have an effect on the diameter of the system and that effect will be similar regardless of the node selected. In scale-free networks by comparison, the vast majority of the nodes are of low degree and therefore play a very minor role in the connectedness of the graph. As such the removal of a node at random, which is highly likely to be a node of low degree, will have a negligible effect on the diameter of the graph. In a targeted attack however, a hostile agent can choose to remove one of the few, highly connected hubs; in this case a much larger than normal proportion of the paths through the graph would be lost, leading to a rapid rise in the diameter of the system.

Based on their studies of the World-Wide Web, Albert, Jeong and Barabási had shown over the course of two years that scale independence existed in real-world, large-scale systems. They has also demonstrated that such systems were resilient to the random failure of a significant fraction of their nodes, and developed a growth model that explained both their evolution over time. Additionally, they had offered an explanation as to why other traditional models had not shown similar behaviour. Now they moved away from the web and began working with a set of important biological graphs, metabolic networks[43] (for an introduction to metabolic networks see section 2.2.6). For this study Albert, Jeong and Barabási joined with two biologists from Northwestern University in Chicago, Bálint Tombor and Zoltán Oltvai. Using data from the WIT (now ERGO) database, they built graphs of the metabolic processes predicted to be taking place in 43 different organisms, based on their annotated genomes. Analysis of the graphs showed that they did indeed possess the characteristic power-law degree distribution of scale-free networks; furthermore, it showed that the substrates acting as hubs in the metabolic networks were practically identical across all the graphs, despite only 4% of the complete list of substrates being present in all 43 systems (the directed metabolic graphs presented in chapter 6 are also scale-free). The diameter of the networks showed almost no variation with the complexity of the organism, which given the average value of just over three steps made metabolism a very small-world indeed.

The scale-free network model seemed to fit with a wide range of metabolic systems, but was it solely responsible for their topology? Two years later, and now working with Ravasz, Somera and Mongru, Oltvai and Barabási published another paper about the structure of metabolic networks, this time discussing how they apparently differed from the standard scale-free model[50]. Looking at the same 43 systems, they showed that the clustering coefficients of the graphs were an order of magnitude higher than was predicted by the scale-free model. This suggested that the traditional, biologists' viewpoint of metabolism as a modular system, consisting of strongly linked groups of nodes with weaker ties between them (much like Granovetter's model of the labour market), may also be an important element in the topology of metabolic networks. To explain this phenomenon, Ravasz *et al.* proposed the 'hierarchical network', a scale-free system made up of repeating, nested groups of nodes rather than individuals thus allowing both a high clustering coefficient and a power-law degree distribution.

2.2.5 Metrics for the determination of network structure

The range of metrics that can be used in the determination of network structure is huge and a small section in a document such as this can not hope to give any more than a superficial coverage of the topic. With this in mind the following section attempts to briefly discuss some very basic metrics and their uses, and then moves on to cover two more complex metrics that are likely to be of use within the field of metabolomics. For a more complete review of metrics and how they relate to graph structure, the reader is directed to the papers by Newman[51], Albert[52] and Dorogovtsev[53].

A number of very simple metrics can give a surprising amount of important information about the structure of a graph. The numbers of nodes and edges alone can for example, be used to suggest how dense the graph is and whether it is likely to be a single, large structure or a collection of fragments. Knowing the degree of each node, the number of edges that are incident to it, is also important; the degree distribution for a graph, node degree plotted against the probability that a node in the graph has that degree, can be suggestive of many types of architecture. A Poisson distribution for example, might indicate a random structure, whereas a distribution that is sharply spiked at a particular value is more suggestive of a regular structure such as a lattice. It might also indicate a graph that has evolved from a regular structure such as the small-world networks of Watts and Strogatz. A degree distribution that follows a power-law is often considered indicative of a scale-free system, although caution is advised as many systems have been reported as scale-free based on questionable fits of power-laws to the data. In directed graphs, the degree of each node may be subdivided into an indegree, the number of arcs leading into the node, and an outdegree, the number of arcs leaving it. The indegree and outdegree distributions can further assist in the determination of structure by suggesting features such as layered structures, where one group of nodes within the graph feeds into another. Paths between nodes play an important part in many graph applications and the determination of structure is no exception to this; the network diameter, the length of the longest shortest path between two nodes within the graph, and the average path length between all node pairs in the graph, can be used to help separate small-world structures (both Watts/Strogatz and scale-free) from others. Another path metric, betweenness centrality, is the fraction of the shortest paths between all pairs of nodes within the graph that pass through the node or edge for which it is being calculated. High betweenness centrality values indicate that the node or edge may be a ‘bottleneck’ in the graph and that the structure may consist of groups of nodes connected by a few key elements (this is particularly interesting when considering how automated searches might identify chains of reactions responsible for observed biological effect, see chapter 5). The level of grouping of the nodes within a graph can also be investigated using clustering coefficients, a measure of the extent to which the neighbours of a node are also connected to each other. A low average clustering coefficient might indicate a tree structure or random graph, while higher values would suggest lattices or, as above, groups of nodes that are densely connected within themselves but only connected to each other by a few links. Figure 2.7 shows an example of how these simple metrics might be used to distinguish between some of the graph structures discussed.

While metrics such as the network diameter aim to describe the overall structure of the network under investigation, other metrics attempt to describe how the local environment around the nodes might appear. One such method, Triadic Census, was proposed by Holland and Leinhardt in 1970[54]. A triadic census is a survey of the graph for the sixteen unique configurations of three nodes and the arcs between them (Figure 2.8), and the comparison of the frequencies of occurrence for each of these structures against a set of ‘expected’ frequencies for a random graph of the same size. The sixteen triads used during the triadic census are not the only possible ways of linking three nodes with arcs; if you consider each of the nodes to have a label and therefore be different from the others, then there are 64 possible configurations. The triads are an example of the graph isomerism problem, the way in which two structures can appear to be different because of their layout or numbering but are identical from the structural point of view (Figure 2.9).

Motif search[55] uses the same ideas as triadic census but applies them to subgraphs of size n; effectively replacing the dictionary of sixteen, three node structures with that of the n node

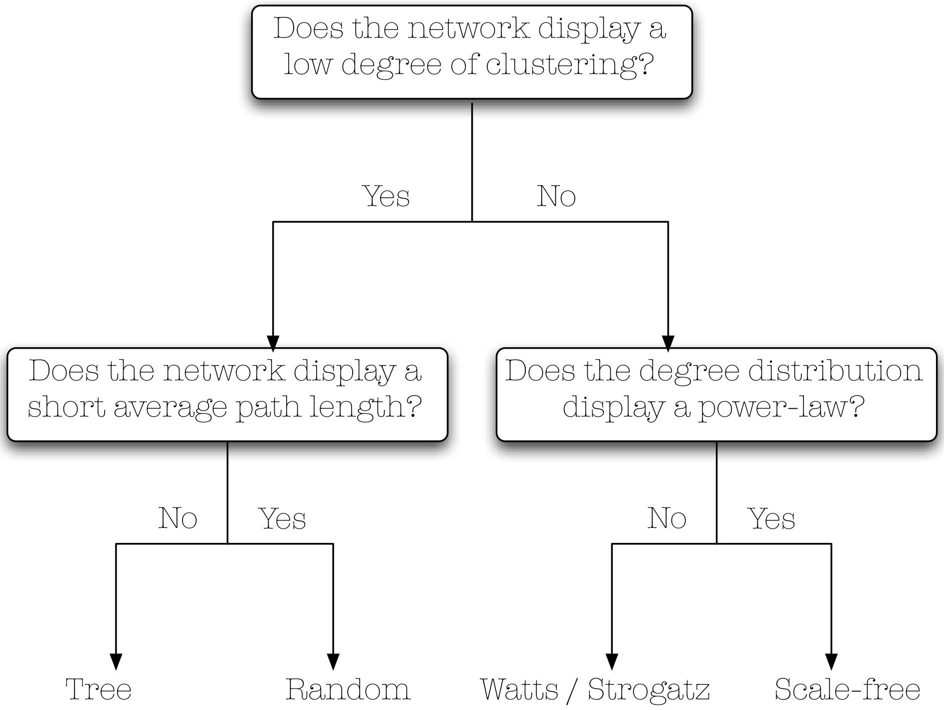

Does the network display a low degree of clustering?
Yes
No
Does the network display a short average path length?
Does the degree distribution display a power-law?
No
Yes
No
Yes
Tree
Random
Watts / Strogatz
Scale-free

Figure 2.7: An example of how a collection of simple metrics can be used to guide the determination of network structure.

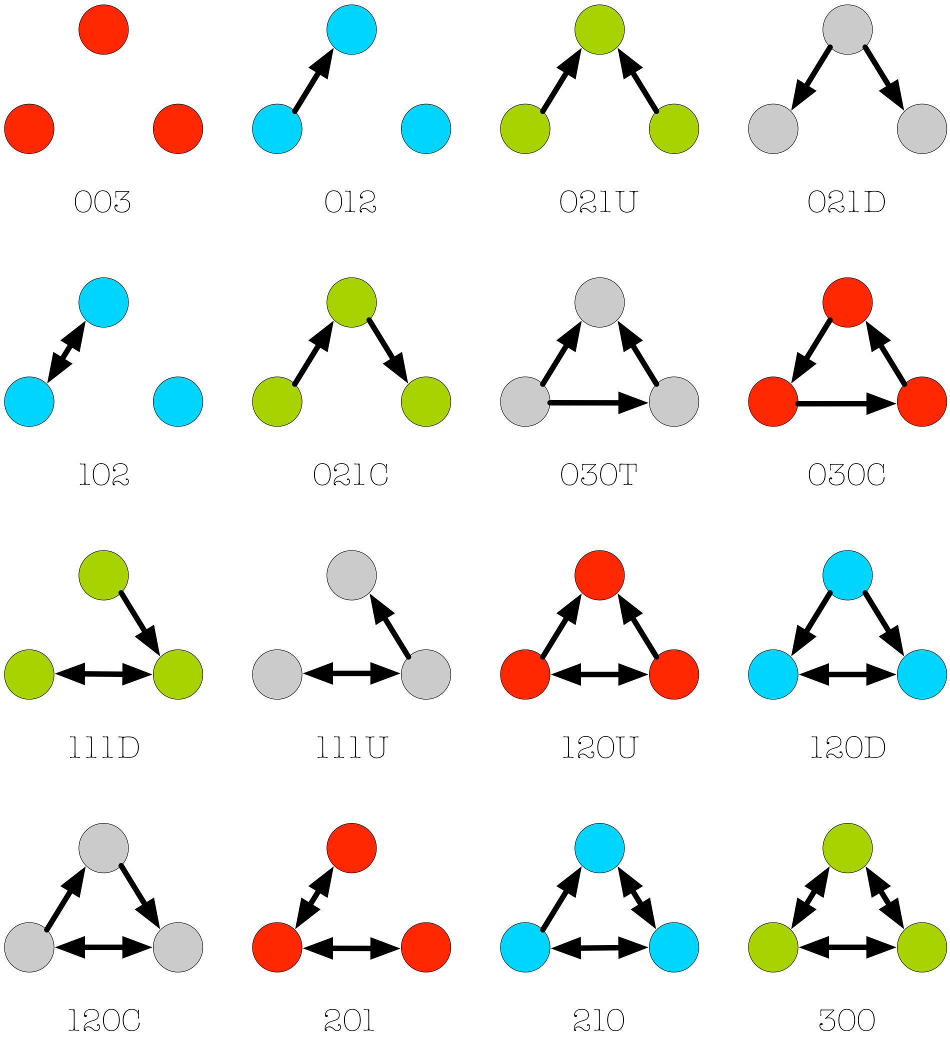


Figure 2.8: The complete set of triads used during triadic census. The numbering system xyz describes the triads based on x, the number of pairs of nodes linked by a bidirectional arc, y, the number of pairs of nodes linked by a unidirectional arc, and z, the number of unlinked pairs. The letter is used to further distinguish between the shapes, up, down, transitive or cyclic.

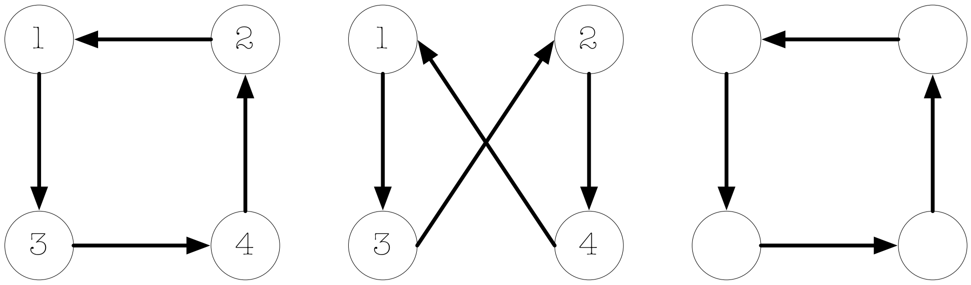


Figure 2.9: An example of the graph isomerism problem. The graphs on the left and in the centre of the figure appear to be different, but if the numbering is ignored and the nodes on the right-hand side of the graph in the centre (2, 4) are swapped we can see that it forms the graph on the right. Thus the graphs are the same from the structural point of view.

structures. Triadic census and motif search are both of potentially great value when dealing with very large graphs, since a knowledge of the types of structures that exist within a graph can give us some idea of how it might behave on a larger scale. This was perhaps best illustrated by Milo and Shen-Orr in 2002[56], who having applied motif search to the transcriptional regulatory network of E. coli showed that three motifs they had identified were of known biological significance. Later in the same year they published a second paper, which went on to show how the technique could also be applied to a range of other systems including food webs, the connections between neurons in C. elegans and electronic circuits.

Since Milo and Shen-Orr's papers in 2002, a great deal of additional work has been done in order to speed up the search for motifs and make the technique more accessible to a range of users. Of particular interest is the work by Berg and Lässig[57], which extends the idea of a motif by introducing scoring to allow partial but not necessarily identical structures to also be counted as part of the motif search. This could be used to take into account incomplete datasets (often the case in large networks) and in the case of biological systems in particular, the effects of evolution on the structure of the motif. A range of software tools for motif search exist, most notably Mfinder[58], MAVisto[59] and FANMOD[60].

2.2.6 Metabolic pathways

Metabolic pathways are chains of reactions which when combined show how a particular function such as the extraction of energy from food (glycolysis) or the biosynthesis of important molecules is performed by a cell. Metabolic pathways are usually presented as graphs in which nodes represent compounds and edges represent reactions, often labeled with the enzymes that catalyse them. The complete set of metabolic pathways for an organism is known as its metabolic network. A number of online resources that include information on metabolic pathways exist, of which arguably the most well-known is the Kyoto Encyclopedia of Genes and Genomes (KEGG)[61, 62]; a suite of databases that aim to assist scientists integrating genomic and metabolic data. The KEGG PATHWAY database (the component of KEGG that deals with metabolic pathway information and the current basis for the Linked Metabolites software presented in chapter 5) is a collection of static “reference pathways”, that is to say that each of the pathways contains experimentally determined reactions contributing to the process it describes but taken from a variety of organisms and then combined to form a single graph. As many key metabolic processes are well preserved amongst most organisms from mammals to bacteria[63] it is possible for KEGG and other similar resources to computationally generate organism specific pathways from the reference pathways by highlighting enzymes (and the reactions they catalyse) in the reference pathways that are known to be present in the organism based on its annotated genome (Figure 2.10).

MetaCyc[64, 65] is a database of reference pathways used by the BioCyc project as a base from which to generate organism specific pathways. Unlike in KEGG, where the generated pathways are presented as overlays to the reference pathway diagrams, the BioCyc project stores the extracted pathways as separate databases, effectively creating whole, organism specific pathway databases that can then be queried, edited and updated independently. When first created, the computationally generated pathway databases have not been subject to any human curation and may include errors or omissions; as such BioCyc has an evidence coding system to indicate the original source of the information, be it from primary literature, computationally predicted, predicted and then subject to curation etc. As of October 2008, over 370 computationally-derived pathway databases were available as part of the BioCyc project, of which 20 had undergone a limited curation effort. These included HumanCyc[66], an organism specific pathway database for Homo sapiens.

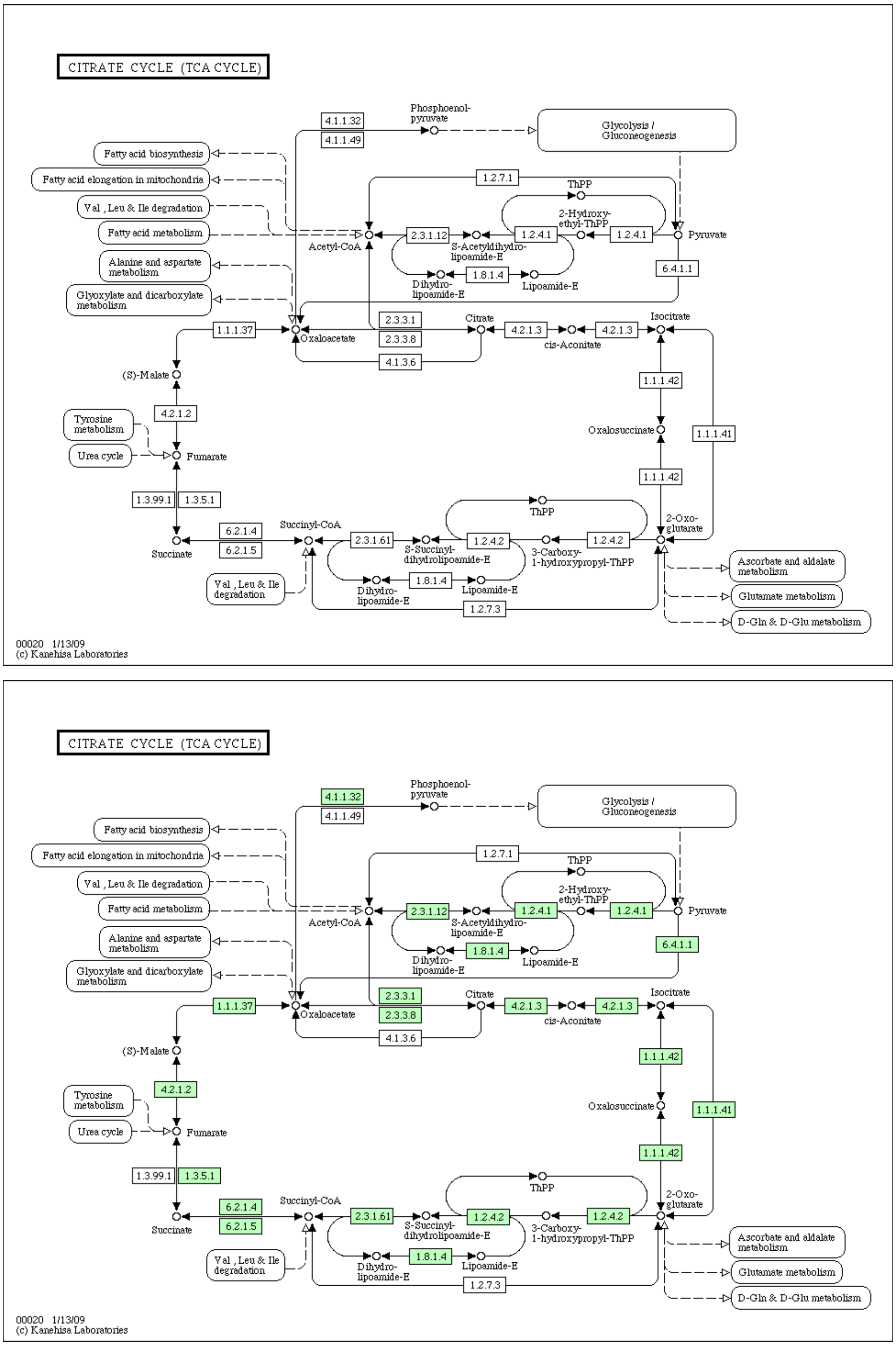


Figure 2.10: KEGG pathway diagrams for the TCA cycle. The top diagram is the reference pathway and below is the same diagram highlighted based on the annotated genome for Homo sapiens.

While the reference pathways in databases such as KEGG and MetaCyc represent specific biological processes, the point at which a process becomes a pathway can be somewhat arbitrary, with a single pathway in one database being represented by two or three in another. Methods exist that aim to allow the creation of organism specific metabolic pathways based on a definition of the concept of a pathway rather than from experimentally determined reference pathways. In 2003, Masanori Arita[67] published a method that generated possible metabolic pathways from a graph in which edges represented a transfer of atoms between the source and target molecules. Since creating pathways based on this type of transfer could in theory lead to almost any molecule transforming into any other, an important concept in Arita's work was that there should be some degree of conservation of moiety (a structural fragment of a molecule) along the entire length of any valid pathway, that is to say at least one atom from the source molecule should have been passed along the pathway to the target. This is analogous to the idea that a brush that has had two new handles and three new heads over a period of years is not the same brush. Arita's method contained three steps; in the first stage, just under 2,800 reference metabolic reactions were taken from the Enzyme Nomenclature database, manually checked against the literature for correctness (in some cases the reactions were either rearranged or balanced) and then passed to a software package that mapped carbon, nitrogen and sulphur atoms in the substrates of each of them to the equivalent atoms in their products. As was the case with the reactions from which they were generated, the mappings were then checked by hand and corrected where appropriate. Finally, the mappings were used to construct reference graphs for carbon metabolism, nitrogen metabolism and sulphur metabolism, in which the nodes represented metabolites and edges represented the transfer of molecular fragments between them. The second stage of the method involved the search for potential metabolic pathways between source and target compounds, however for this to take place the actual metabolism of the organism of interest (in Arita's case E. coli) needed to be selected in the reference graphs. As is the case in other methods using reference pathways, Arita's method achieved this by selecting reactions based on enzymes referenced in the annotated genome, which could be downloaded from the KEGG database. Rather than excluding reactions that were not reported as present in E. coli, Arita instead assigned them a much higher weighting[†] than those reported as present. As a result of this, the searching of the graph for potential pathways, which was performed using a k-shortest paths algorithm, could still report pathways that included those reactions even though it would preferentially select those pathways that only included reported reactions. This was important, since the annotated genome may have been either incomplete or contain errors and thus relying on it alone could cause valid pathways to be overlooked. Any potential pathways that looped back on themselves were rejected at this point, since while there are a small number of genuine metabolic pathways that form loops, for example the TCA cycle (also known as the citric acid cycle or the Krebs cycle), the vast majority of those results would be artifacts due to the reversible nature of chemical reactions. The third and final stage of the method involved the filtering of the potential pathways against the conservation of moiety criteria discussed earlier; this involved the checking of the entire pathway against the individual mappings in order to ensure that at least one atom from the source molecule passed along the entire length of the path to become part of the target molecule, thus making it a valid biochemical pathway. Ultimately, pathways derived in this way may prove a much more useful basis for automated search techniques than those arranged to be easily viewed by humans and currently available in KEGG (see chapters 5 and 6).

The computational derivation of organism specific metabolic pathways from reference pathways has one important advantage of methods such as Arita's; it avoids the problems associated with compartmentalisation, the idea that only a subset of the complete genome is expressed

[†] When searches are performed using a shortest paths algorithm, as was the case in Arita's work, edges with high weights are less likely to be selected than those with low weights as they increase the length of the path. Hence, in these cases a high edge weight is a penalty.

in a particular tissue or cellular structure, by virtue of the experimentally determined nature of the reference pathways. Despite this, computationally-derived databases are of limited use because of the lack of curation (in order to remove or correct erroneous pathways) and lack of references to evidence describing the pathway in the literature. An alternative group of metabolic pathway databases exist that address this question by focusing on a single organism and only containing curated data. The huge effort associated with the curation of such databases means that they can only be created for important model organisms. A good example of this type of database (and the database that provided much of the initial pathway information for MetaCyc) is EcoCyc[68], a metabolic pathway database for E. coli. More recently another curated database, Reactome[69, 70], was launched that focuses on human metabolism and then computationally extends that effort to a small number of other important organisms. Along with the human-specific pathway diagrams, Reactome contains citations to the literature, links to information in external resources such as UniProt and GO, and the ability to layer experimental results onto the diagrams. Critically, unlike other pathway databases, Reactome also includes information on sub-cellular compartmentalisation and citations to experimental evidence for each of its reactions; information that is vital if the database is to be used as the foundation for *in silico* simulation of the pathways.

The computer simulation of metabolic pathways is an important area of research, enabling the testing of theoretical models of metabolism against experimental results. Yet despite kinetic modeling and simulation packages such as Gepasi[71]. E-Cell[72] and Virtual Cell[73] having been available (in some cases) for two decades, it is only recently that the community has really begun to embrace the field. While a component of this new-found interest is almost certainly due to the increasing availability of low-cost, high-performance computing facilities, a large part of the credit must also go to efforts designed to facilitate interoperability between the myriad compound and pathway databases, the exchange of models between simulation packages and the reuse of existing models as components of larger systems. Amongst these two stand out, BioPAX[74] and the Systems Biology Markup Language (SBML)[75].

BioPAX is a data exchange format for biological pathway information and is of great importance since its ontology allows the unambiguous description of pathway objects, thus avoiding the issue of different naming conventions across databases. BioPAX data is stored using OWL/RDF and is therefore suitable for automated querying and combination of data from multiple sources by web services or software agents. Most critically for any data exchange format, BioPAX has good community support, which currently includes the BioCyc databases and Reactome.

SBML is an xml-based format for the representation of models of biochemical reaction networks, allowing the specification of information such as reactant species, cellular compartments and mathematical descriptions of the interactions between objects. Its aim is to provide a common format for the exchange of such models between simulation packages thus removing the risk of models becoming unusable if the package they were originally written in ceases to be supported, while also making it easier for researchers to check and reuse published models that previously may have needed to be manually recoded in the proprietary formats of multiple simulation packages.

While metabolic pathway information in BioPAX format is available in great volume and with unambiguously named elements allowing the easy combination of smaller pathways to form larger ones, it does not encode all the information required for the simulation of the pathways. SBML by comparison does contain the necessary equations and rate parameters but does not require the unambiguous naming of elements outside the scope of an individual model. While recent versions of SBML partially address this by allowing models to be annotated with references to external information, the annotations are optional and SBML does not impose any restrictions on their content. Thus, while SBML does allow models to be used in multiple simulation tools, it does not necessarily solve the issue of reuse of models as parts of more complex systems. Projects such as the BioPAX modeling framework[76] are currently attempting to address this problem by allowing fully annotated SBML models to be constructed from BioPAX data.

The methods chosen for the visualisation and exploration of metabolic pathways can have a great impact on their usefulness, particularly if the features of interest are spread across multiple pathways. Traditionally, metabolic pathways are visualised using semi-static views like those in KEGG, diagrams created and laid out in advance by a curator. Typically such diagrams cannot be edited by the end user, although they may have the facility for additional information such as experimental results to be overlaid on them, and can only be navigated by the use of hyperlinks, with the diagram of the linked pathway replacing that of the pathway previously being examined. While this method can be appropriate for the visualisation of well-defined reference pathways (particularly when there is only one pathway being investigated) it is not at all suited to the visualisation of large systems containing multiple pathways or to the potentially huge numbers of computationally-derived pathways. In response to this need, packages such as KGML-ED[77] have been created, which based on xml versions of the metabolic pathways allow users to create customised, dynamic views of the pathways. These often include the ability to customise the pathways by adding or removing elements, include multiple pathways in a single diagram, hide information by collapsing pathways into single nodes that can be expanded as required, and change the layout either manually or based on standard layout algorithms from graph theory. Many also allow the customised pathways to be exported in formats such as SBML for visualisation elsewhere or as the basis for simulations. Packages such as Cytoscape[78] and VisANT[79] take this a stage further by allowing the user to combine data on metabolic pathways with gene regulation networks or by facilitating the colouring of the on-screen network with experimental data emulating functionality previously available as part of semi-static visualisations through tools such as those based on the KEGG EXPRESSION database[80], which enables the intensity ratios of spots on gene expression arrays to be mapped onto the corresponding objects in appropriate pathway diagrams. Metabolite data has also been visualised in-context as an overlay to the pathway diagrams, with work by Dwyer *et al.*[81] that used columns of coloured blocks to represent time series concentration data being a particularly good example. Finally, it would be inappropriate to talk about the visualisation of experimental data in the context of metabolic pathways without mentioning VANTED[82]; a recent software package that includes many of the techniques mentioned (combination of data from multiple fields into pathway diagrams, flexible layouts) and combines them with analysis techniques such as the ability to build correlation networks from experimental data.

2.2.7 Metabolic correlation networks

One of the major challenges in metabolomics is the inherent biological variability between samples, both across groups of subjects and between samples taken from the same subject at different points in the day. Factors such as diet, the age of the subject and environmental stresses can all lead to large variations in the metabolic profile of a sample, making direct comparisons difficult. Fortunately, this variability also has its uses; pairwise comparisons of the levels of metabolites across whole groups can lead to the identification of correlations within the data, indicating a possible relationship between the metabolites in question. While heatmaps offer one method of visualising this type of data, an alternative option is to generate a metabolic correlation network[83], in which metabolites that are correlated more strongly than a given threshold are linked together to produce a graphical representation of the important relationships within the system (Figure 2.11).

The first stage in the creation of a metabolic correlation network is the assignment of correlation values to each pair of metabolites being studied. Correlations may be calculated in a number

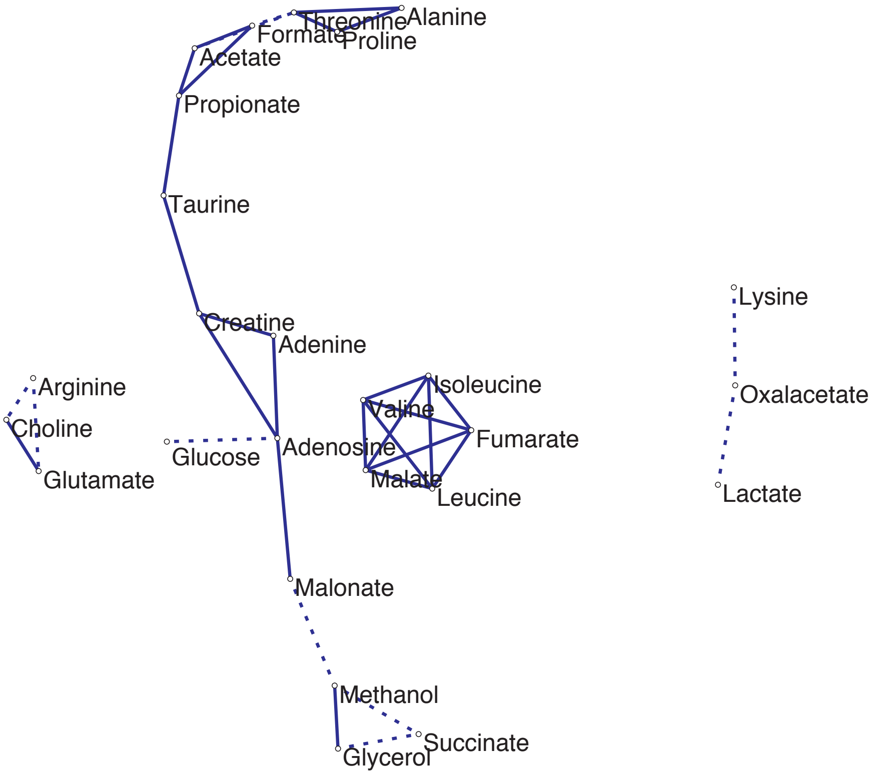


Figure 2.11: Example of a metabolic correlation network.

of ways, but the most common are the Pearson Correlation Coefficient (Equation 2.5) and the Spearman Rank Correlation Coefficient (Equation 2.6).

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sigma_X \sigma_Y} \tag{2.5}$$

$$r = 1 - \frac{6 \cdot \sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2_{-1})} \tag{2.6}$$

The major difference between the Pearson and Spearman correlation coefficients lies in the fact that the Pearson coefficient is parametric (there is an assumption that the data is normally distributed) while the Spearman coefficient is not. If the data can be assumed to be normally distributed, the Pearson correlation coefficient should be used in preference to the Spearman correlation coefficient. They also differ due to the ranking approach used by the Spearman correlation coefficient. In the Pearson coefficient, the correlations are calculated based on the distance between equivalent points in the data series being used. The Spearman correlation coefficient however, uses the distance between the ranked order of the equivalent points, meaning that data series that are identical in shape but differing in average amplitude will appear more highly correlated than they would using the Pearson correlation coefficient. The Spearman correlation coefficient will also return a more favourable result in situations where the data contains a relatively small number of outliers, as the difference in the rankings will be small compared to the actual difference between the values of those points.

The strength of a correlation is not necessarily proof of its significance, which is affected by factors such as the size of the underlying dataset. As such before a metabolic correlation network can be constructed a p value (measure of significance) must be calculated for each r value (strength of correlation). Correlations that result in a p value of less than 0.05 are normally considered to be significant, although this value should really be adjusted to reflect the number of correlations being calculated through the use of Bonferroni corrections, false discovery rates[84] or similar. The conversion of r to p values can be performed either by reference to standard tables, or through the use of t-tests. Since t-tests require the assumption that the samples within the dataset are normally distributed, metabolic correlation networks should not be constructed from datasets with fewer than 20-30 samples.

Once p values have been determined for all pairs of metabolites under investigation, the construction of the network becomes a simple task. Every metabolite is represented in the network by a node. Next, the correlation value for each pair of metabolites is considered; if the significance (p value) for that correlation is below the threshold value determined previously then an edge is added to the network between the correlated metabolites. Commonly, the weighting of the new edge is then set to the strength (r value) of the correlation it represents.

For an example of the use of metabolic correlation networks in the investigation of tumours in dab liver samples, see chapter 4.

2.2.8 Time-series methods for biochemical network discovery

While metabolic correlation networks can be a useful tool for the exploration of experimental data, the difficulties associated with relating interesting features in the networks to the underlying biochemistry mean that the results of such analyses can at best only be considered indicative. There are however, a set of methods that allow the direct reconstruction of networks of biochemical reactions from time-series data of simple systems. Although the type of data required for this process[‡] means that such approaches are currently of limited applicability to

[‡] When not being driven by external factors such as changes in temperature or the lack of availability of required ‘input’ compounds, systems of chemical reactions will tend to settle into a steady state where each of the reactions proceeds at a constant rate and the concentrations of the compounds involved are effectively fixed. The network reconstruction methods discussed here require the perturbation of the system away from this state and as such can only really be used on data gathered *in vitro*.

metabolomics, a summary of them is included in appendix A as an interesting future alternative to the use of metabolic correlation networks and reference pathways.

2.3 Summary

This chapter has introduced the key concepts that will be used in this thesis. It began with a description of metabolomics, the study of the low molecular weight compounds involved in metabolic processes, and went on to give examples of its applications in environmental toxicology and healthcare. The following sections discussed graph theory, introducing a number of different types of graph and metrics that could be used to distinguish between them, before showing how the structure of a graph can give insights about the system that it is modeling. Finally, it brought the two topics together, showing how biologists have chosen to represent their current knowledge of metabolism in a series of graphs known as the metabolic pathway diagrams, how they are attempting to explore their experimental datasets using metabolic correlation networks, and mentioning how they might reconstruct networks of biochemical reactions from experimental datasets in the future. The next chapter will go on to discuss some of the limitations of the current approaches, and ask what might be done to address them.

CHAPTER 3

LIMITATIONS OF CURRENT APPROACHES

3.1 Metabolomics and the identification of biomarkers

Although a comparatively recent addition to the -omics family, metabolomics with its low per-sample cost and good portability across species (when compared to other -omics technologies), has huge potential both as a standalone analytical technique and as a component of larger integrated genomics studies. In particular, two key features of the technology, the fact that it focuses on the low-level molecular processes of the cell and its ability to be used non-invasively through MRS or on biofluids such as urine, make it an ideal tool for screening work in areas such as medicine or environmental toxicology.

Unfortunately, many of the factors that contribute to the strengths of metabolomics within these fields are also its greatest weaknesses. The biological variation between subjects due to age and sex, local environmental conditions and dietary differences are all reflected in the results of a metabolomics experiment and often make the data noisy and difficult to use. While it is sometimes possible to reduce the impact of these effects on the data, through the use of careful experimental design, this is often undesirable as the control measures themselves may mask interesting features in the data or add to the risk of any discoveries that are made only being applicable in a narrow range of experimental conditions, greatly limiting their more general usefulness. The experimental difficulties will however be reduced as the technology matures, and ultimately it will be the ease of interpretation of the data that determines how widely and in which fields metabolomics is adopted.

Since it is highly unlikely that the concentration of a single compound alone will uniquely identify and determine the extent of a particular condition, current methods for the identification of biomarkers centre on multivariate statistical tools such as Principal Components Analysis (PCA). These techniques, when used in conjunction with significance tests, will often identify an extensive list of compounds that in some way contribute to the differences between the samples in a study. Many of these will be "usual suspects"; compounds that are involved in processes related to, but not the root cause of, the condition being studied such as apoptosis (cell death) or anaerobic respiration. The inclusion of such compounds in a list of biomarkers would lead to highly complex identification rules for the condition - rules that would be cumbersome to implement, difficult to extend as further data became available, and most importantly considering the essentially safety-critical nature of many of the applications*, impossible to explain and validate.

Thus, within every metabolomics study a careful balancing act must be performed. The set of biomarkers for a condition needs to be complex and resilient enough to specifically identify that condition despite experimental noise and biological variation, simple enough that it can be easily applied by practitioners in the field, and only contain compounds that have a clear and understandable biological rationale for their inclusion. At the heart of this problem is the ability to relate trends observed in the experimental data, to the well-characterised biochemical

*In this case, the safety-critical component is due not to a risk to the subject posed by the use of metabolomics itself (which owing to its potential for non-invasive use actually carries a lower risk of damage to the subject than current methods such as biopsy), but to the potential for problems that would result in the case of false-negative results, where diseases or toxic effects were allowed to continue untreated.

processes that make up the metabolic pathways.

3.2 Relating experimental data to biological processes

The generation of a mapping between experimental data and the metabolic pathways is far from being a trivial task. It could be argued that to date the most effective approaches have also been the most straightforward; those that layer experimental data onto pathway diagrams. These techniques, initially driven by work such as that of Dwyer[81] and available as part of tools like VANTED[82], prove most useful when the set of metabolites being studied is taken from either a single or a small number of pathways, allowing the researcher to judge the relationship between the metabolites by eye. They are also particularly effective when used with time-series data, allowing changes in concentrations to be traced through the pathways. However, studies focused on a single pathway are far from the norm and the large number of pathways stored in databases like KEGG can make comparison by eye alone a slow and difficult task. This problem is compounded by the relatively low-resolution of the technology as it currently stands; a difficulty that is particularly evident in the case of MRS where the number of metabolites that can be observed is presently only a small fraction of that for *in-vitro* methods. This low coverage of the metabolome means that it is highly unlikely that the set of metabolites being observed in an experiment will naturally fall in close proximity to one another within a single metabolic pathway diagram. Instead, they are likely to be spread over several different pathways, making it difficult to form coherent chains of reactions that link the compounds together.

Correlation analyses can be very helpful when interpreting metabolomic data, as they can be used to indicate how compounds might be related before the time-consuming process of searching the known metabolic pathways takes place. However, difficulties do exist with this approach; firstly the existence of a correlation between two metabolites is not in any way proof that it is significant.

While a number of methods for determining the significance of a correlation exist, including both Bonferroni and false discovery rate-corrected p values from t tests, the choice of method and of threshold value still has a very noticeable effect on the selection of correlations, an effect that is undesirable when trying to establish a common basis for further research. Secondly, we have the issue of “long-range” effects; just because a correlation exists between two metabolites there is not necessarily an obvious link between them in the underlying biochemistry, as Steuer, Kurths, Fiehn and Weckwerth remarked “...there is no straightforward connection between the observed correlations and the underlying reaction network. We observe strong correlations between seemingly distant metabolites, whereas metabolites sharing a common reaction are not necessarily correlated.”[85]. Still, while the combination of correlation analyses and the metabolic pathway diagrams may be a far from perfect way of determining how experimental results can be explained in terms of the underlying biochemistry, they are considerably more practical (given current experimental limitations) than more advanced methods such as those based on the reconstruction of the metabolic network from time-series data.

Since a correlation analysis of even a small system produces a reasonably large number of results (a correlation analysis of time-series data of 4 metabolites results in 6 correlations assuming you ignore correlations of a compound to itself), the method chosen for the visualisation is important. Correlations are normally viewed as heatmaps, however an interesting alternative is the metabolic correlation network (Section 2.2.7) as proposed by Steuer *et al.*[85]. Metabolic correlation networks offer possibilities to further automate the process of relating experimental data to the metabolic pathways since they are essentially in the same form, graphs.

The categorisation of graph structure has been developed greatly over the past 50 years, mainly due to the interest of sociologists, and a wide range of metrics have been devised to determine the types of properties a graph may have based on its structure. Of these, motif search (Section 2.2.5) stands out as a particularly good example, showing how repeated structural features within the network may be related to important functions within the system it represents. The expected types of structure within the network being studied do of course impact heavily on the metrics that should be used. In the case of locating a set of reactions that correspond to a correlation for example, one sensible approach might be to study the betweenness centrality values of metabolic pathways that link the compounds involved in the correlation. High betweenness centrality would suggest that a large proportion of the routes involve that step and therefore that it is likely to be involved in the observed correlation.

To assume that such simple rules of thumb could be made to solve all the issues involved in this problem would however be a gross under-estimation, as the way a correlation appears can depend on a large number of factors. As an example, consider a short section of the glycolysis pathway (Figure 3.1(a)) in which circles represent metabolites, arrows represent reactions, and the numbered boxes represent enzymes. To begin with, assume that there is a metabolomics dataset that has concentrations for glycerate 3-phosphate (glycerate-3P) and phosphoenolpyruvate. In glycolysis glycerate-3P is turned into phosphoenolpyruvate so we'd expect that these two compounds would be correlated (Figure 3.1(b)). Now assume that our dataset has been expanded to include glycerate 2-phosphate (glycerate-2P), since glycerate-3P forms glycerate-2P and glycerate-2P then goes on to form phosphoenolpyruvate we might expect the correlation network to look much like the pathway (Figure 3.1(c)); however, correlation networks are not like pathway diagrams, glycerate-3P and phosphoenolpyruvate will still be correlated, despite not being directly connected in the pathway (Figure 3.1(d)). This is an example of how the experimental 'coverage' of the compounds can drastically alter the appearance of the correlation network compared to the metabolic pathway it represents; a small number of relatively distant compounds may look quite similar to the pathway (although there is of course a risk that some other factor might mean they are no longer correlated), whereas compounds that are close to-

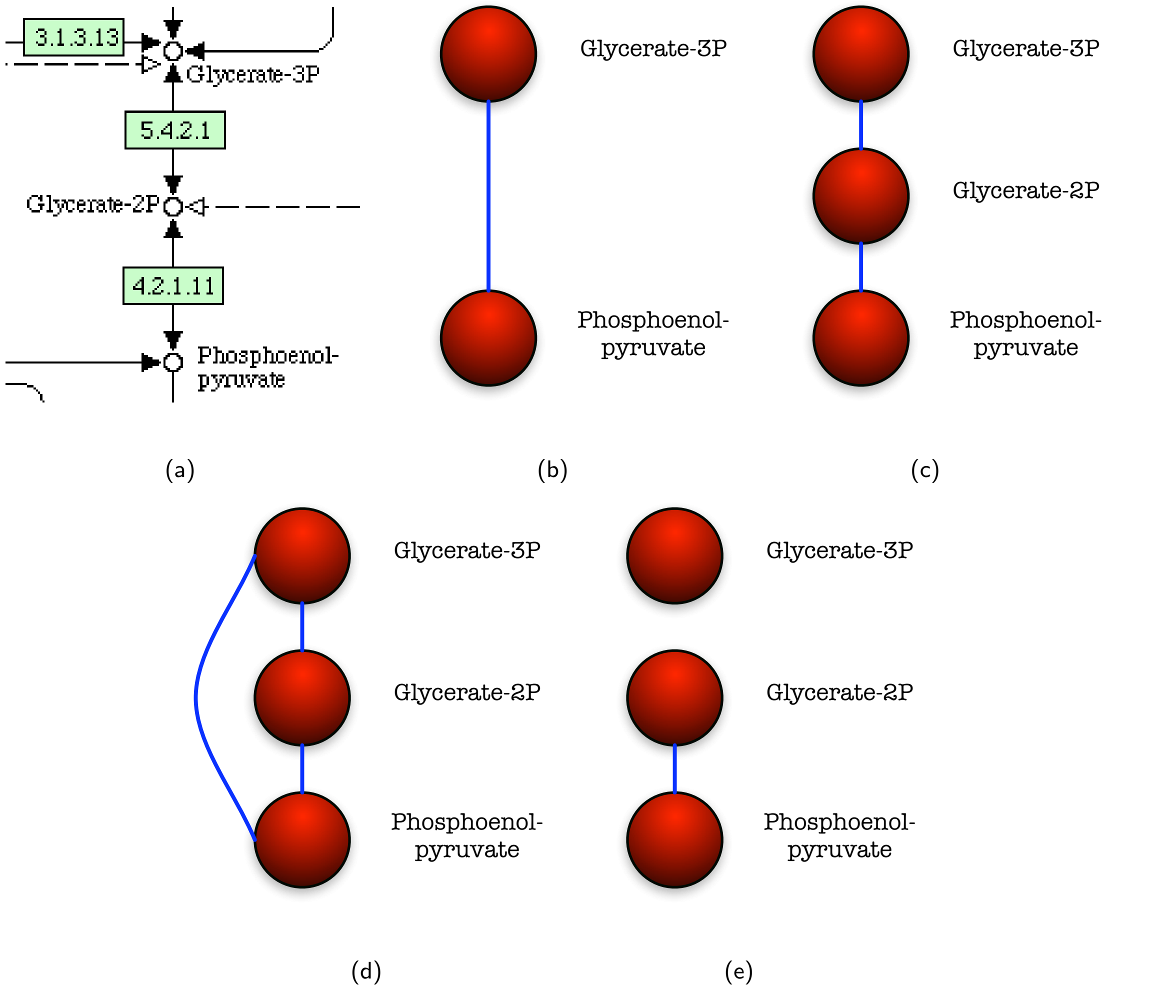


Figure 3.1: A small section of the glycolysis pathway shown with several possible metabolic correlation networks.

gether will tend to form cross-connected groups. To further complicate this situation, while the metabolic pathways are drawn to represent particular processes in an easily interpretable form, many pathways are at work within a single cell at any one time forming a much larger, interconnected system. The dotted arrow coming into the pathway (Figure 3.1(a)) from the right hand side is KEGG's way of showing that glycerate-2P can also be formed in another pathway (in this case the pentose phosphate pathway) and feed into glycolysis at that point. If in the other pathway glycerate-2P is formed from a compound other than glycerate-3P, and the rate of flow of glycerate-2P into glycolysis from the other pathway is faster than the rate of formation from glycerate-3P, then the correlation between glycerate-3P and glycerate-2P may not be significant (Figure 3.1(e)). The correlation network does not of course, contain any information on our artificial partitioning of the metabolic system; however, when looking at it you could be forgiven for thinking that the glycolysis pathway is not involved in the process the network represents, when actually it may be important but just flowing at a low rate.

Thus, while correlation networks can be used to represent experimental data in a form that is comparable to the metabolic pathways, issues of "coverage" of the pathways by the experimental data, the completeness in terms of the underlying metabolism of the metabolic pathways themselves, and the current dynamic state of the metabolic networks within the subjects (the rates of flow through various parts of the pathways as determined by enzyme activity, temperature etc), mean that there is currently no straightforward way of automatically aligning the two. Indeed, the question of network dynamics is as of yet largely untouched and while likely form the major area of research for the graph theory community over the next decade.

Given the lack of a suitable framework for the development of a completely automated solution to this problem, we are left with attempting to facilitate current manual efforts. Many of the current approaches involve selecting compounds from the lists generated by statistical tests that all feature in the same pathways, and then investigating the links between them by eye. Alternatively tools such as KEGG PathComp might be used to link them automatically, but this is usually limited to either a single-source to multiple-targets, or multiple-sources to single-target search.

Here, we will investigate how metabolic correlation networks can be used to identify sets of related metabolites which seem to contribute to the differences between the sample groups. We will then look at how an all simple paths search algorithm can be used to perform multiple-source, multiple-target searches, creating a directed metabolic graph, which can be used to visualise the relationship between the compounds groups in terms of the biochemistry.

CHAPTER 4

METABOLIC CORRELATION NETWORKS OF LIVER TUMOURS IN DAB

4.1 Background

One of the greatest strengths of metabolomics as an analysis technique lies in the ability of the metabolome to quickly reflect the effects of changes in the environment. Unfortunately this means that the data gathered is often very noisy, due not only to biological variation but also factors such as diet, age and time of day, along with those being studied. In situations where sufficiently large numbers of subjects are available, using correlations between metabolite concentrations can help overcome this problem. Unfortunately, even a relatively small dataset of 30 metabolites will result in over 400 separate correlations[*], leaving the researcher facing what is still a highly complex analysis task. This chapter aims to show how using metabolic correlation networks (Section 2.2.7) to present the data in a manner that is easily interpretable by eye can help researchers identify features in their data, which they can then use as the basis for further research such as searches of the metabolic pathway diagrams.

[*] Assuming the correlation between A & B is the same as the correlation between B & A, the number of unique correlations is given by $\frac{n^2-n}{2}$ where n is the number of metabolites.

4.2 Initial investigation (2005 analysis)

4.2.1 Method

Metabolic correlation networks were constructed from NMR metabolomics data of 20 samples taken from dab, a common flatfish. Flatfish are frequently used in environmental monitoring work, since they live in close proximity to the ocean floor and are therefore exposed to the chemicals that accumulate there. Of the 20 samples, which were taken from a total of 10 fish, 10 were samples of tumour tissue (one per fish) and the remaining 10 were matched samples of normal liver tissue taken from the same 10 fish. Analysis of the NMR data led to the identification of 33 separate metabolites, each of which was then quantified. All data collection and metabolite quantification work was performed by metabolomics practitioners within the School of Biosciences. The samples were then analysed as three groups; normal tissue, tumour tissue, and a third group in which both the normal and tumour tissue samples were treated as a single, combined dataset. In each case, a series of networks was produced with cut-off values for the inclusion of a correlation into the network ranging from ± 0.1 to ± 0.9. The Pearson correlation coefficient was used to generate correlation values for all pairs of metabolites within the groups. Positive correlations were represented in the networks by solid lines, while dashed lines were used to represent negative correlations. The calculations were performed using R[86], a free software environment for statistical computing that is widely used within the biological community.

While the metabolic correlation networks do summarise a large amount of information effectively, they can still be difficult to compare by eye. In order to highlight the differences between the groups and thereby ease the task of analysis, the intersection between the normal and tumour tissue networks was taken for each threshold value and subtracted from both of the networks.

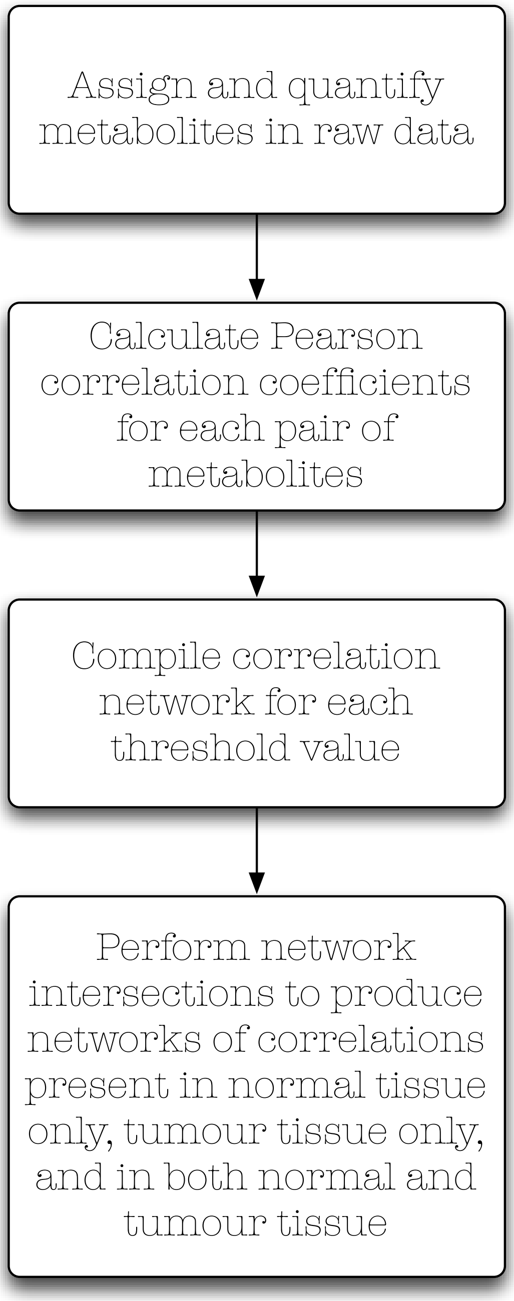

Assign and quantify metabolites in raw data
Calculate Pearson correlation coefficients for each pair of metabolites
Compile correlation network for each threshold value
Perform network intersections to produce networks of correlations present in normal tissue only, tumour tissue only, and in both normal and tumour tissue

Figure 4.1: Process used for the generation of the metabolic correlation networks in the 2005 analysis.

This had two effects, firstly two new series of metabolic correlation networks were created that contained only those correlations which appeared in either the normal tissue network or the tumour tissue network at that threshold value. Secondly, the intersections themselves could be used as a third new series of networks, which showed those correlations common to the normal and tumour tissue networks. The key steps in the method are summarised in figure 4.1. The networks were manipulated and visualised using Pajek[87], a program originally designed for the analysis of social networks (Figures 4.2 to 4.11).

4.2.2 Results

The networks showing the correlations present in the normal tissue samples but absent in the tumours (Figures 4.5, 4.6 and 4.7) indicate that there are strong positive correlations between fumarate and leucine, and fumarate and isoleucine (Figure 4.6). A positive correlation usually indicates that one of the metabolites in involved either directly or indirectly in the formation of the other. In this case the valine, leucine and isoleucine degradation pathway feeds into the TCA cycle just upstream of fumarate (although both leucine and isoleucine can also feed in at an alternative point further round the cycle). The presence of these correlations in normal tissue but not in the tumour tissue suggests therefore that there may be an alteration to the activity of that section of the TCA cycle within the tumour tissue. The presence of additional, weaker positive correlations between isoleucine and malate, fumarate and valine, and malate and valine (Figure 4.7), when malate is also involved in the TCA cycle immediately downstream of fumarate also appear to support this hypothesis. In contrast to this, the correlation between fumarate and malate appears in both normal and tumour tissue (although at slightly higher strength in the tumour tissue) suggesting that this step remains largely unaltered, possibly due to one of the alternative pathways that can feed fumarate into the system (Figure 4.12).

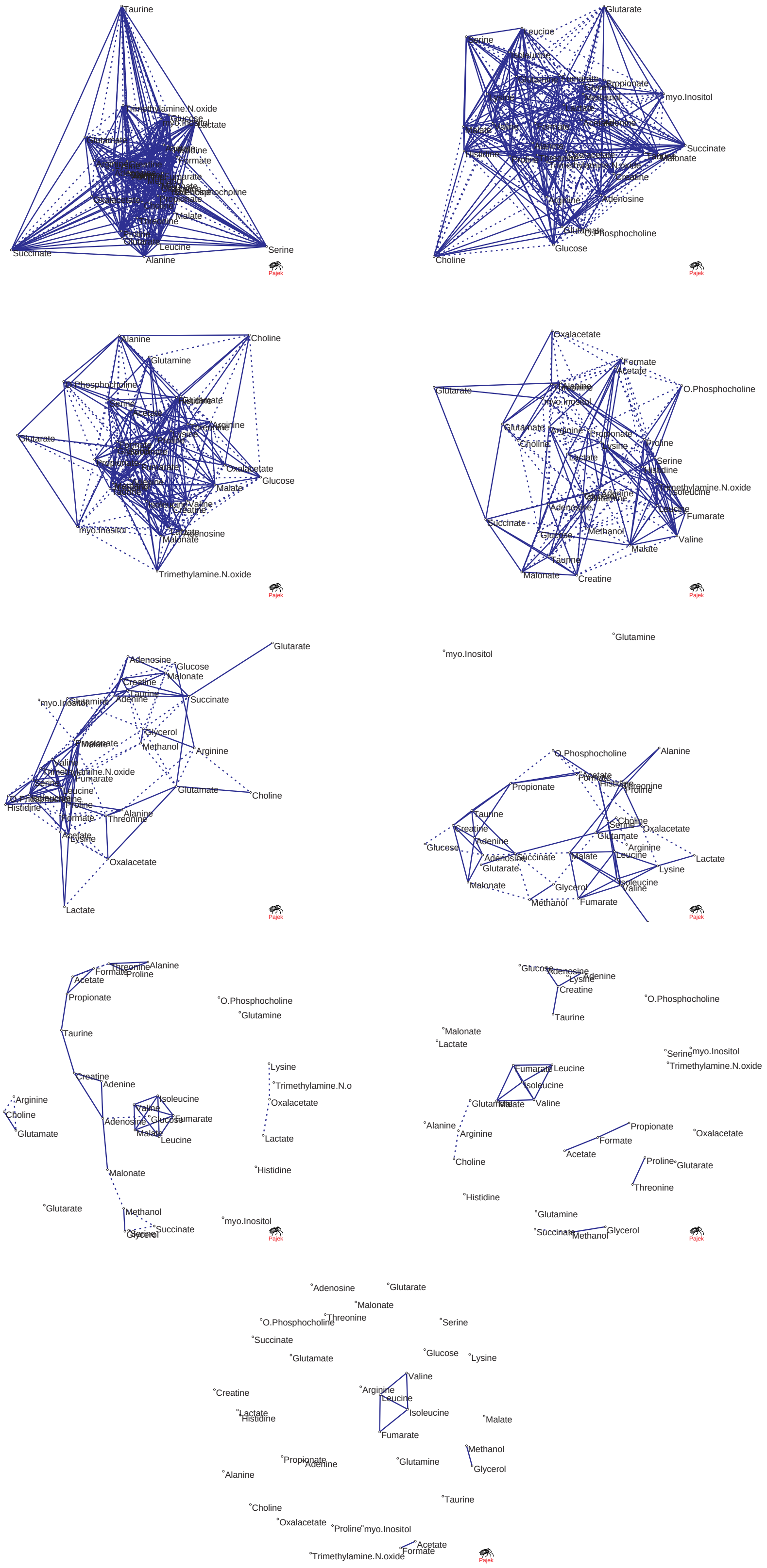


Figure 4.2: Metabolic correlation networks from dab liver samples. Normal tissue group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

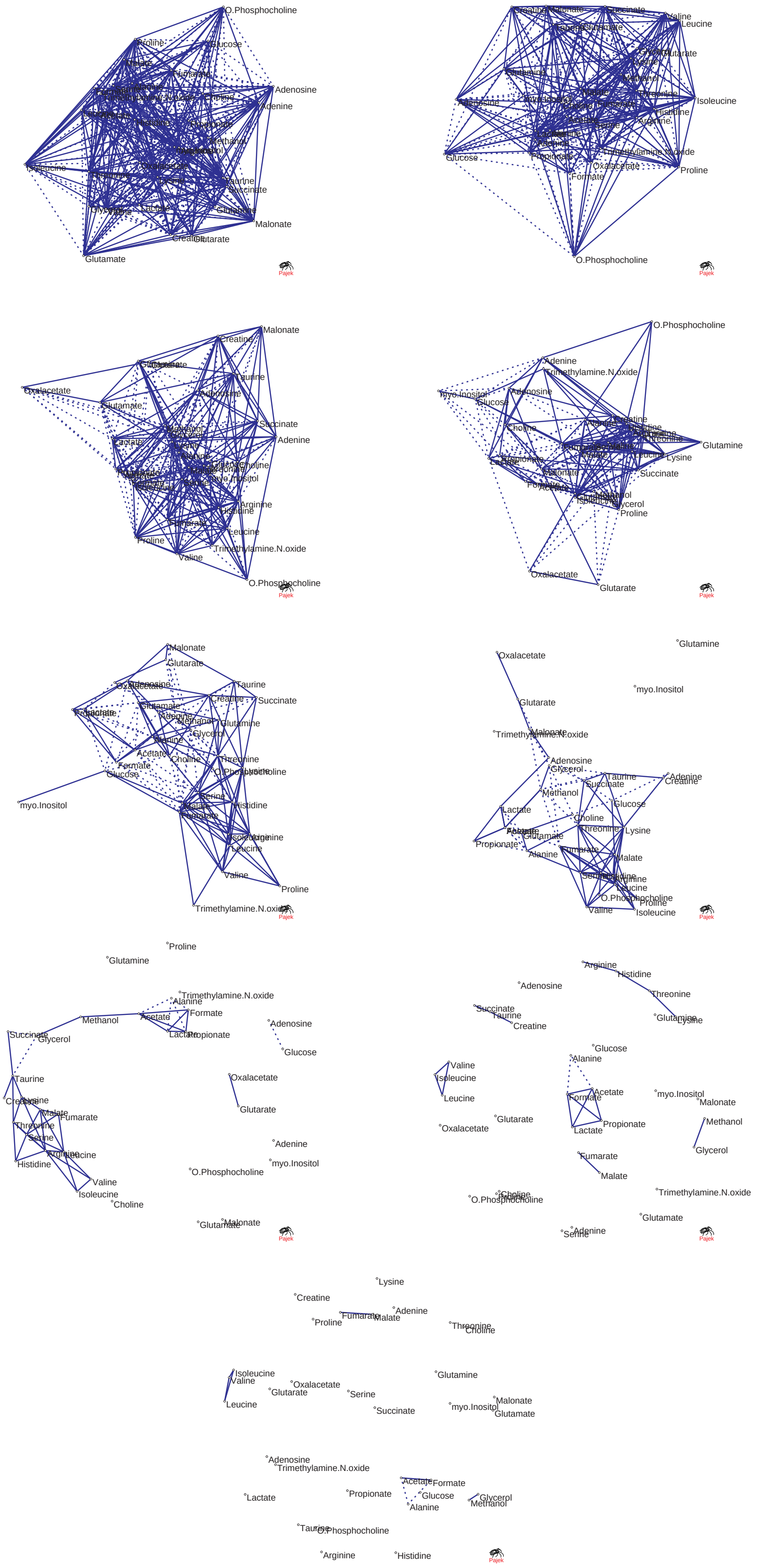


Figure 4.3: Metabolic correlation networks from dab liver samples. Tumour tissue group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

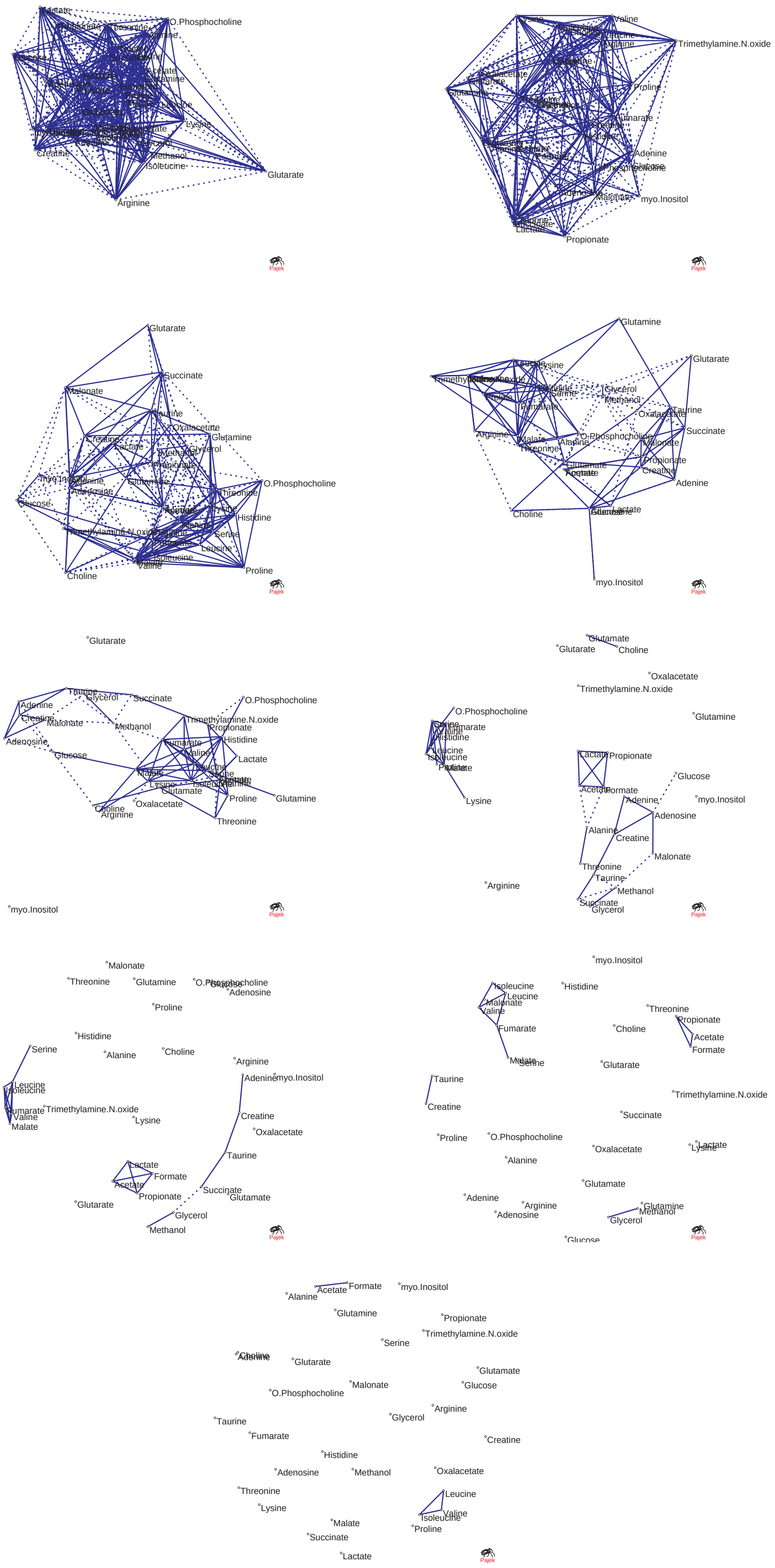


Figure 4.4: Metabolic correlation networks from dab liver samples. Normal and tumour tissue combined group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

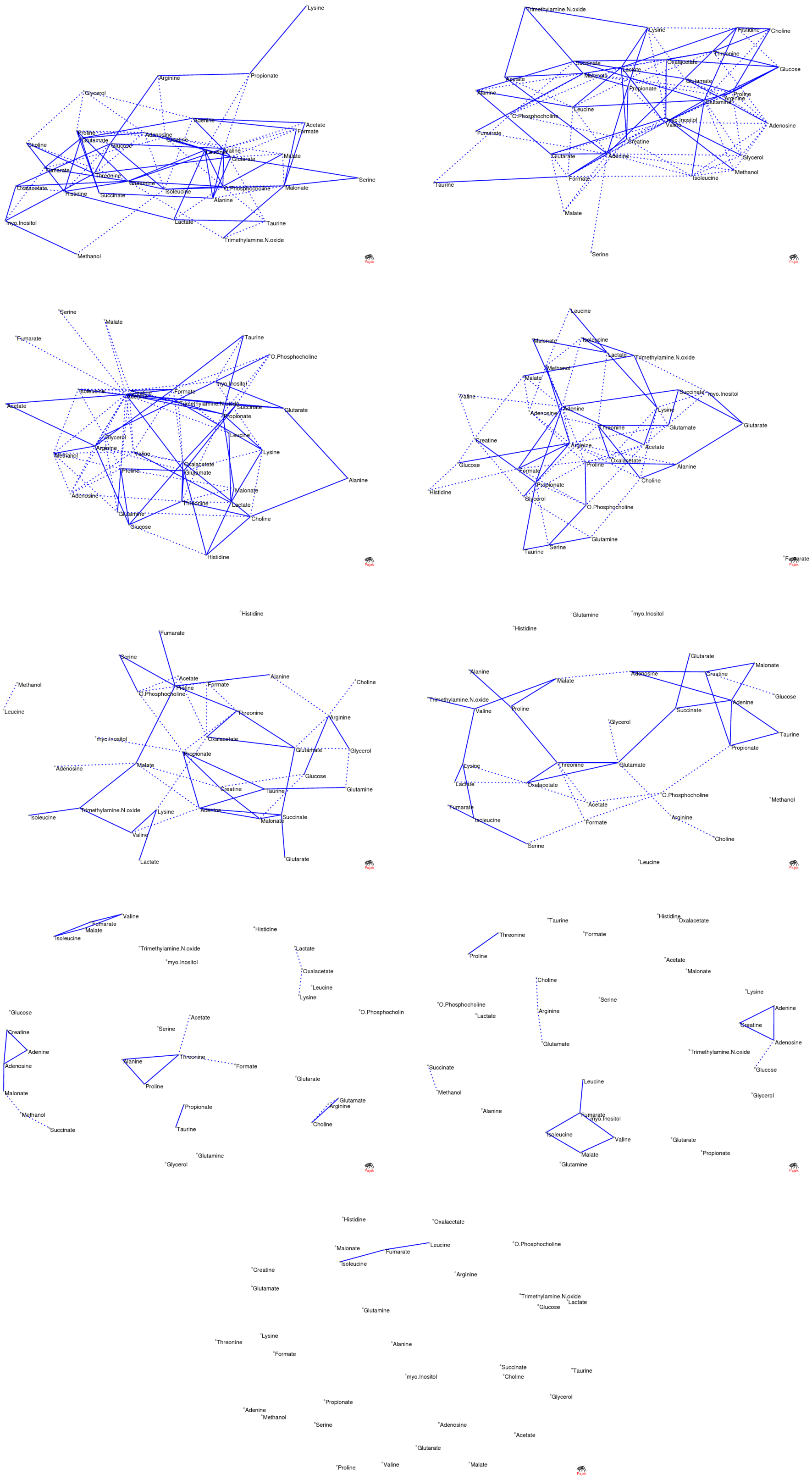


Figure 4.5: Metabolic correlation networks from dab liver samples. Correlations present only in normal tissue group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

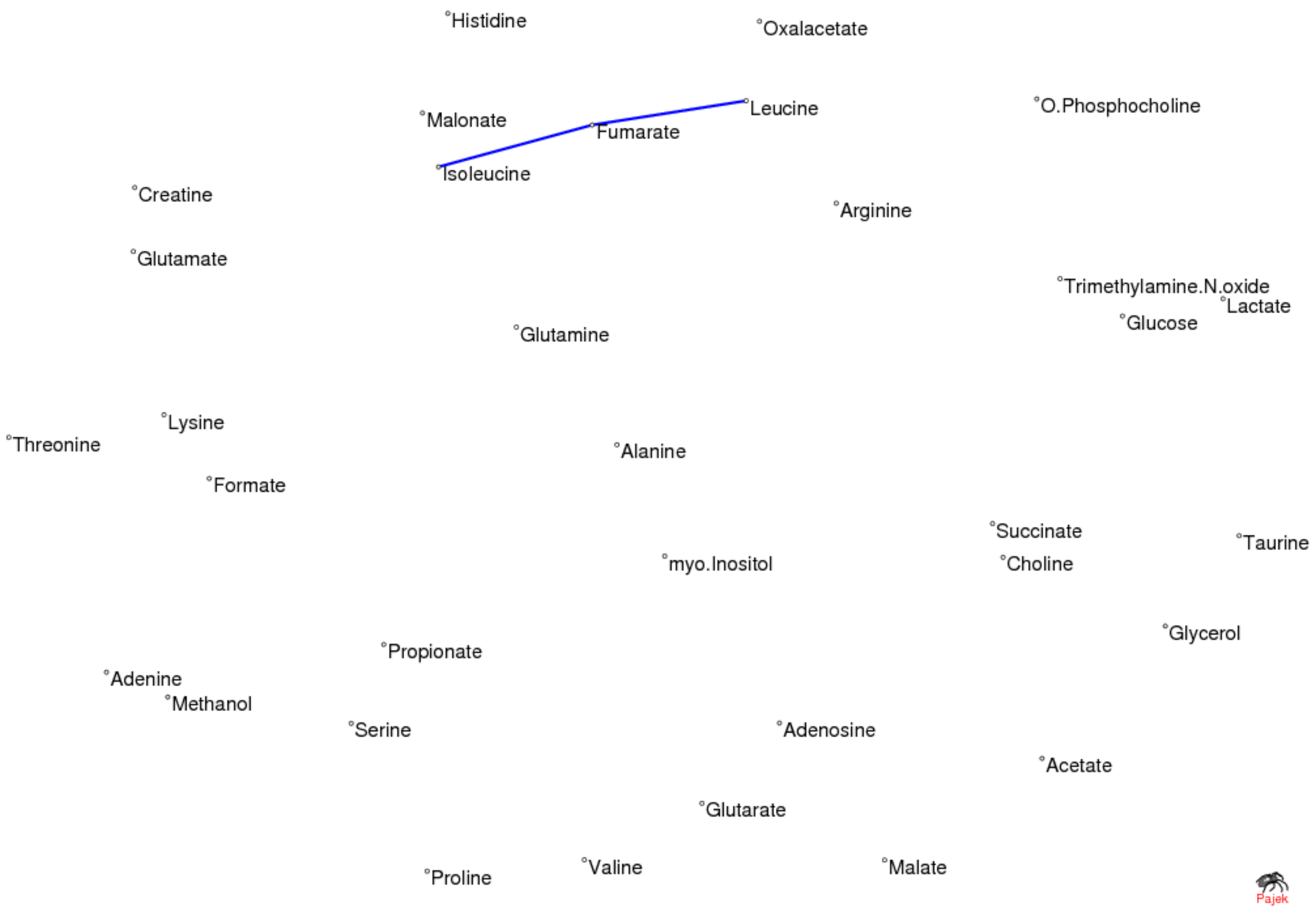


Figure 4.6: Metabolic correlation networks from dab liver samples. Correlations present only in normal tissue group. Threshold of ±0.9. Data taken from the 2005 analysis.

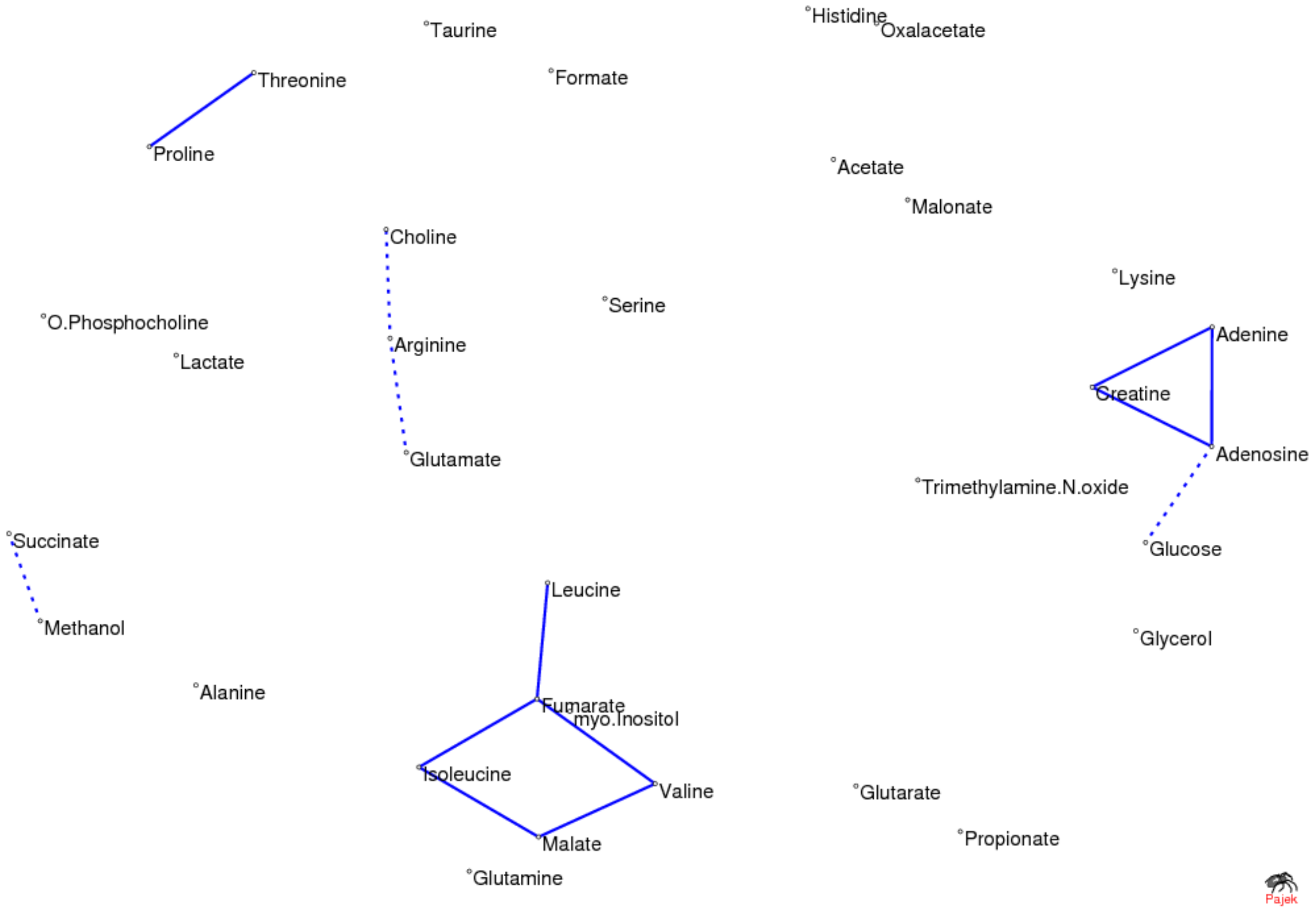


Figure 4.7: Metabolic correlation networks from dab liver samples. Correlations present only in normal tissue group. Threshold of ±0.8. Data taken from the 2005 analysis.

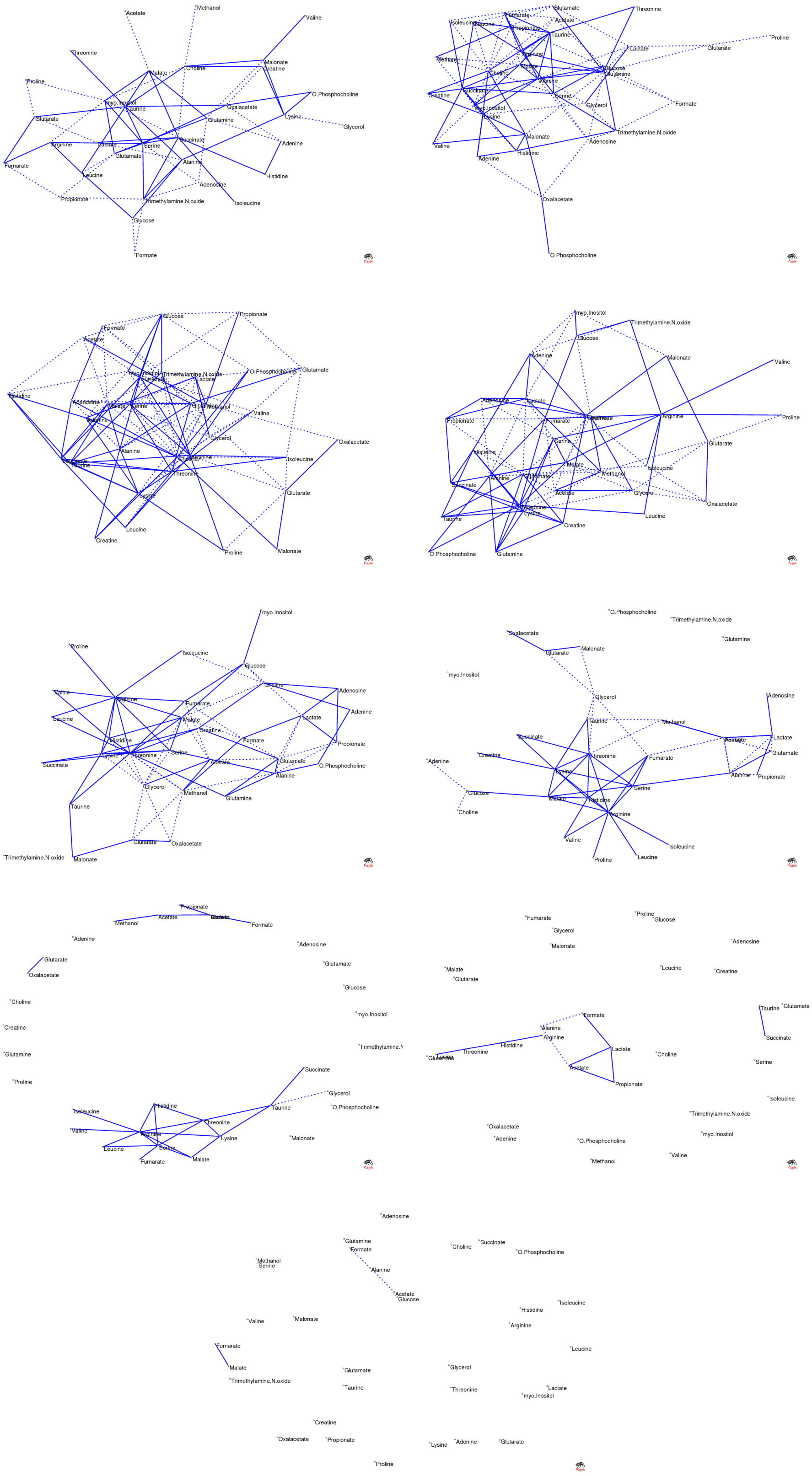


Figure 4.8: Metabolic correlation networks from dab liver samples. Correlations present only in tumour tissue group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

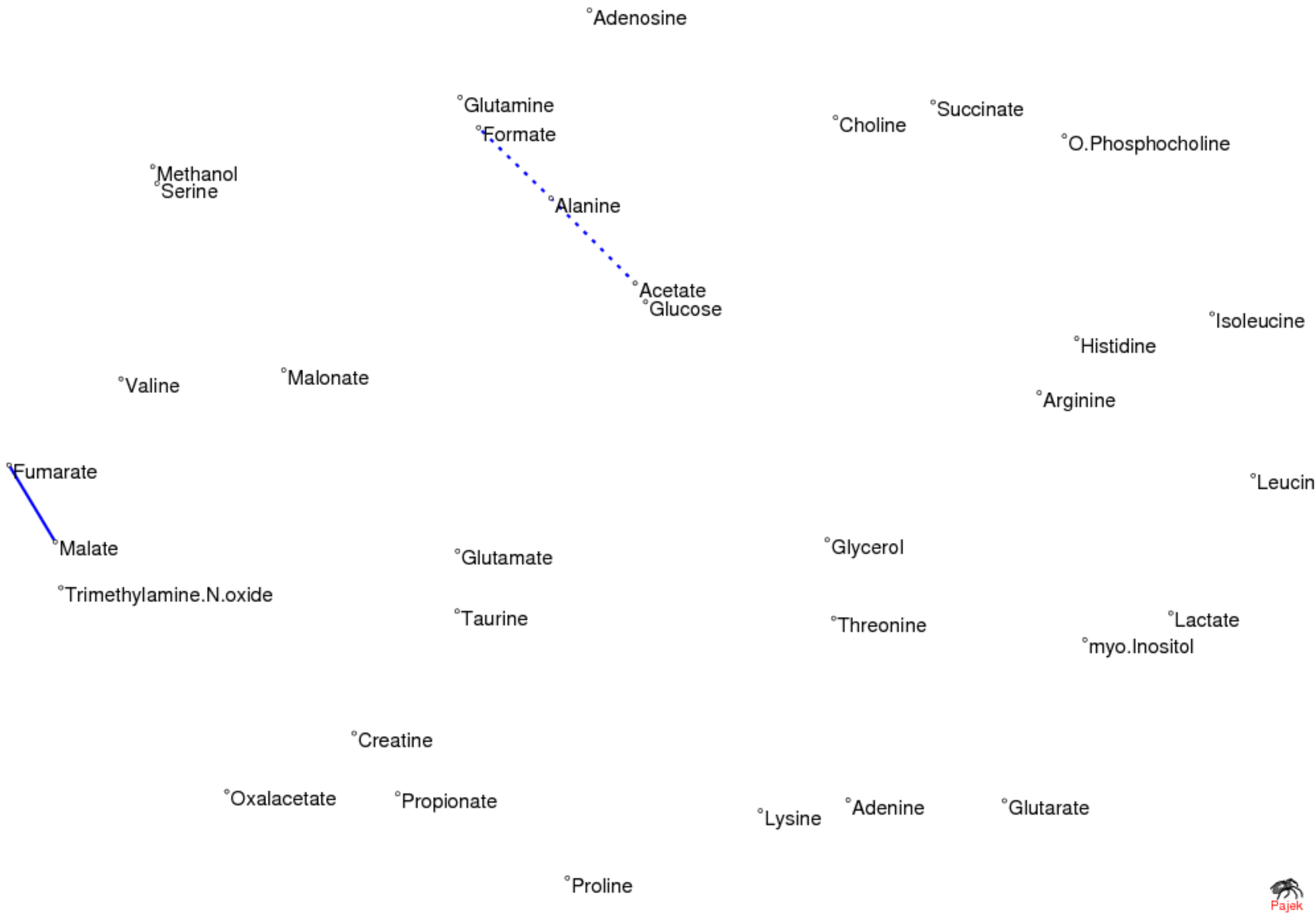


Figure 4.9: Metabolic correlation networks from dab liver samples. Correlations present only in tumour tissue group. Threshold of ±0.9. Data taken from the 2005 analysis.

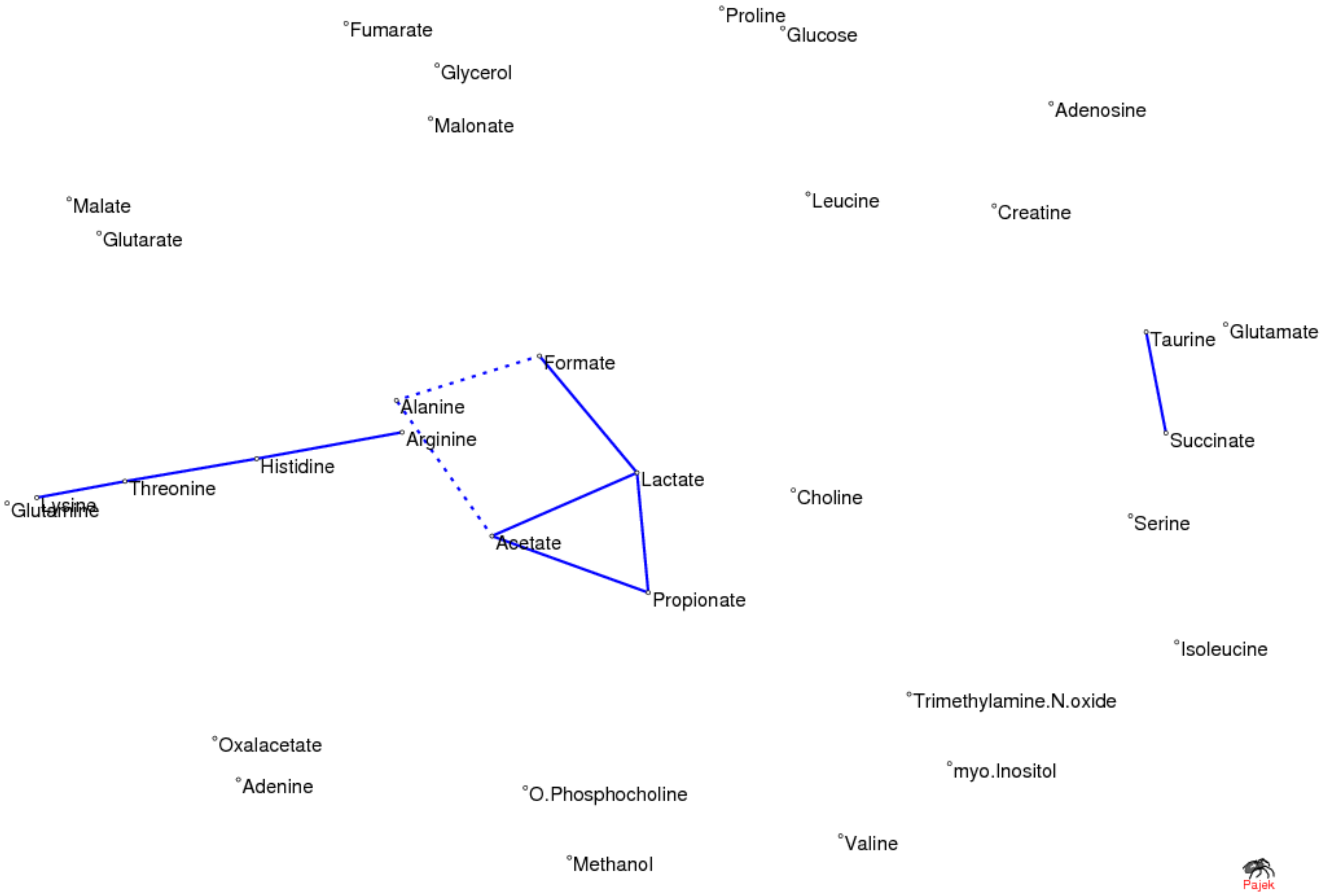


Figure 4.10: Metabolic correlation networks from dab liver samples. Correlations present only in tumour tissue group. Threshold of ±0.8. Data taken from the 2005 analysis.

The networks showing the correlations present in the tumour tissue samples but absent in the normals (Figures 4.8, 4.9 and 4.10) show a high negative correlation between alanine and formate, and alanine and acetate(Figure 4.9). This is interesting as the US National Cancer Institute say that a high level of the enzyme alanine transferase may be a sign of liver damage, cancer and other diseases [88]. It is also known that under fasting conditions, alanine, derived from protein breakdown, can be converted into pyruvate and used to synthesise glucose in the liver via the gluconeogenic pathway[89]. Since it is common for tumours to rapidly outgrow the sources of oxygen and nutrients available to them (leading to necrosis in the centre of the tumour) it would not be surprising for tumour tissue to behave in a way which resembled fasting, hence altering alanine metabolism within the sample. An apparently strong correlation between fumarate and malate can also be seen in the tumour tissue at the ± 0.9 threshold(Figure 4.9), although looking back at the ± 0.8 threshold level it no longer appears (Figure 4.10). This suggests that the correlation is not a result of a major difference in metabolism within the tumour tissue, but is instead due to a slight alteration in the strength of the correlation between the normal and tumour tissue sample groups that occurs near the threshold.

4.3 Detailed investigation (2008 analysis)

4.3.1 Method

The data from the dab samples was recently reanalysed prior to publication in the Journal of Proteome Research[90], resulting in the identification of two additional metabolites. Pearson, Spearman and Kendall correlation coefficients were all calculated for the normal and tumour tissue groups, although only the results of the Pearson correlations were used in the paper. The thresholding approach used in the previous analysis was rejected in favour of a more formal method, which used p values and a false discovery rate[84] less than 10% for the calculation

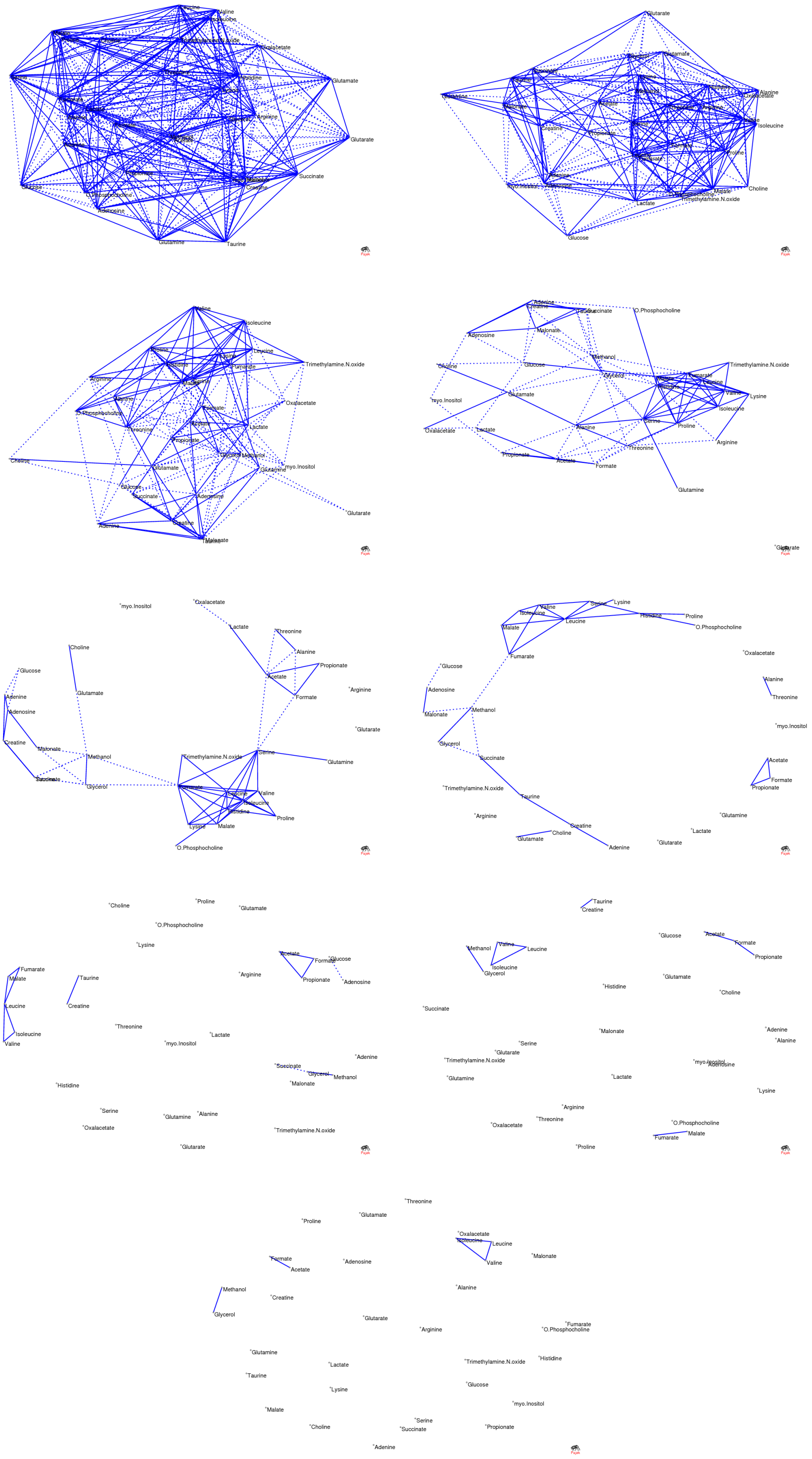


Figure 4.11: Metabolic correlation networks from dab liver samples. Correlations present in both the normal tissue group and the tumour tissue group. Thresholds from ±0.1 (top left) to ±0.9 (bottom). Data taken from the 2005 analysis.

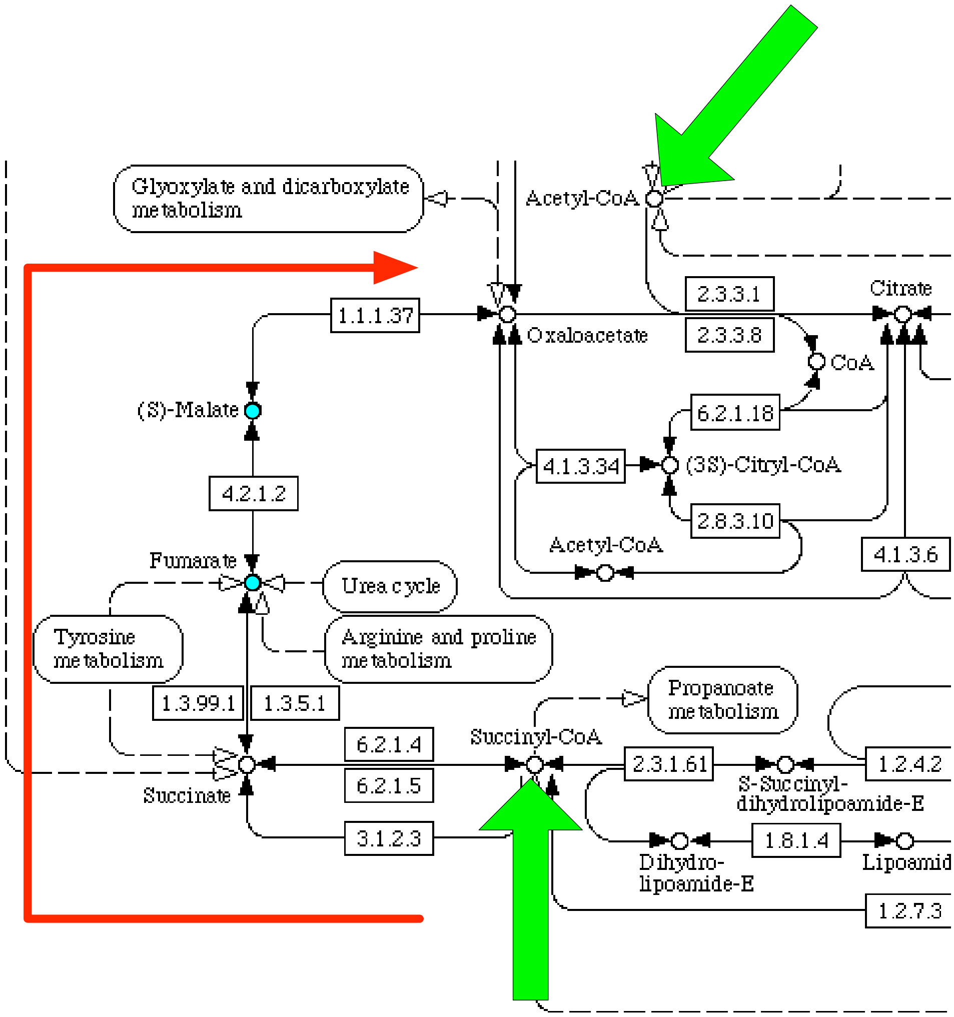


Figure 4.12: A section of the KEGG metabolic pathway diagram for the TCA cycle. The green arrows indicate the points at which the valine, leucine and isoleucine degradation pathway feeds into the cycle. The red arrow shows the direction of the dominant flow around the cycle. fumarate and malate are shown in blue.

of significance and hence inclusion in the networks. Two metabolic correlation networks, one for normal and one for tumour tissue were produced and these were manually inspected for differences. The key steps in the method are summarised in figure 4.13.

4.3.2 Results

The 2008 analysis was performed by an expert biologist and the results are included here to allow a comparison to be made with the results of the 2005 analysis. A total of 33 significant correlations were noticed in two groups. Of these, fumarate was involved in seven correlations and six of those were present only in the healthy tissue. Several significant correlations involving acetate were seen in the tumour tissue samples, as were correlations involving alanine, succinate, and NAD^+. All of these were interpreted as being indicative of a switch from aerobic to anaerobic metabolism in the tumour tissue due, at least initially, to the tumour outgrowing its oxygen supply; it is also thought however, that many of the metabolic processes involved in anaerobic metabolism may be beneficial to the growth of tumour cells. One result of a switch to anaerobic metabolism would be a reduction in the amount of NAD^+, normally produced by oxidative phosphorylation in the mitochondria. NAD^+ is required for the generation of ATP by glycolysis and is hence vital to the function of cells. The production of NAD^+ normally takes place as part of the TCA cycle, and a slowing of its production and feeding into the electron transport chain would also impact on the formation of $FADH_2$ during the succinate to fumarate step of the cycle. This would be in line with the number of altered correlations in that region of the cycle noted in the previous analysis. An alternative mechanism for the production of NAD^+ involves the conversion of alanine to pyruvate and then to lactate, producing NAD^+ as a byproduct. This would explain the correlations involving alanine seen in the tumour tissue, and again, is consistent with the findings of the previous analysis. A slowing of the TCA cycle would also cause a buildup of compounds that feed into it such as acetyl CoA, this could be broken down by an

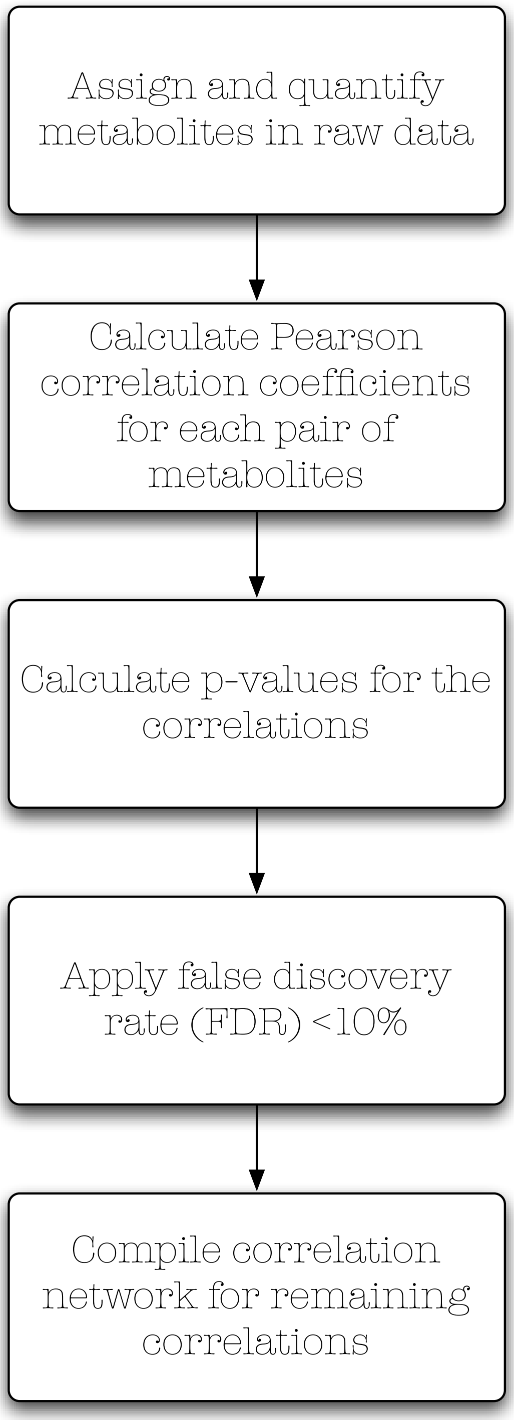

Assign and quantify metabolites in raw data
Calculate Pearson correlation coefficients for each pair of metabolites
Calculate p-values for the correlations
Apply false discovery rate (FDR) <10%
Compile correlation network for remaining correlations

Figure 4.13: Process used for the generation of the metabolic correlation networks in the 2008 analysis.

alternative mechanism forming acetate, explaining the correlations involving it and compounds such as alanine that can form pyruvate and then acetyl CoA.

A set of findings of the 2008 analysis that were not noticed in the earlier analysis involve choline metabolism. The alteration of choline metabolism, a process that can create a compound (S-adenosyl-L-methionine) which regulates gene expression, might lead to the expression of unwanted genes including those involved in the development of cancers. While these correlations were present in the networks built during the previous analysis, the difficulties associated with deciding which thresholds to use meant they were missed. This is a strong endorsement of the more formal thresholding approach used in the more recent analysis.

4.4 Conclusions

While metabolic correlation networks offer little additional information than the correlations themselves they are certainly easier to interpret by eye, highlighting features and interrelationships that might otherwise be missed. As the issue of choline metabolism in this work has illustrated, the selection of thresholds is vital to the construction of a meaningful network; too low a threshold value and non-significant results are included making it difficult to see important features, too high and information is lost. As such, selecting an appropriate thresholding approach such as p values combined with Bonferroni corrections or a false discovery rate is of great importance to the success of a metabolic correlation network based analysis.

CHAPTER 5

THE DISCOVERY OF LINKS BETWEEN GROUPS OF COMPOUNDS WITHIN THE KEGG PATHWAY DIAGRAMS

5.1 Background

Multi-dimensional analysis techniques such as PCA will often, when combined with statistical significance tests and applied to complex datasets, identify multiple compounds that may be used to differentiate between groups of samples within the data. However, identifying these compounds is frequently only part of the battle as the much more difficult task involves establishing why they seem to contribute to the difference between the samples, placing that information in a biological context, and determining what, if anything, their ability to distinguish between the groups can tell us about the diseases or conditions that define them.

One way in which this problem can be tackled is through the identification of routes (chains of reactions and intermediary compounds) linking the compounds of interest within the metabolic pathways. Since a large number of routes can exist between two compounds, particularly if the search is allowed to span multiple pathways, many of the major metabolic pathway databases provide tools that allow this process to be performed quickly and efficiently. By carefully examining the routes produced it is possible to identify reactions, enzymes and compounds that may also be involved in the process under investigation; these can then be used either in the validation of the compounds as biomarkers for the disease or condition being studied (by the prediction of their behaviour and comparison of this to experimental results) or as the basis for further research. By combining the routes between a pair of compounds (Figure 5.1) it is possible to produce a network that in this document will be referred to as a directed metabolic graph; this is in order to draw a distinction between it and other types of metabolic networks such as the metabolic pathways or metabolic correlation networks. The combination of the individual routes into a directed metabolic graph has the advantage of allowing graph theory metrics for the determination of key structural features, such as betweenness centrality or motif search, to be applied to the data. These can be used to efficiently identify features such as reactions that are common to a number of routes, which may not be immediately apparent from the routes when considered in isolation (Figure 5.2). By extending the initial route search to allow either a group of source compounds or a group of target compounds, several of the available database search tools increase the amount of information in the directed metabolic graph; despite this the full potential of the technique is not realised, as no tool currently provides multiple source to multiple target searches and therefore the complete set of routes between all the compounds is not available in a single graph.

Thus, there is a need within the community for a tool that can take two groups of compounds, find routes between them through the metabolic pathways, and present the results both in the context of the pathway diagrams and as directed metabolic graphs suitable for further analysis using graph theoretical techniques.

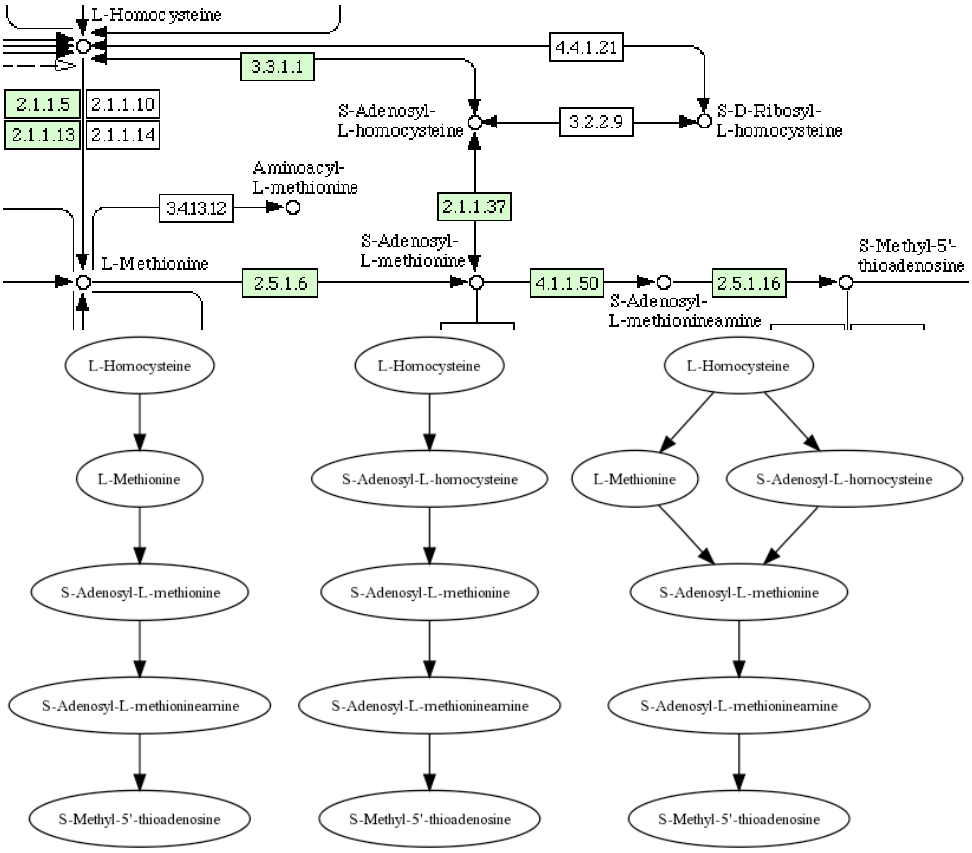


Figure 5.1: A section of the KEGG pathway diagram for methionine metabolism (top). Two routes exist between L-homocysteine and S-methyl-5'-thioadenosine (bottom left, bottom centre), which can be combined to form a directed metabolic graph (bottom right). Since in this example the routes are taken from a single pathway, the directed metabolic graph looks very similar to the metabolic pathway from which it was derived. However, in situations where additional pathways or compounds are included the graph will become too complex for easy manual interpretation and metrics that can summarise the structure will be required for analysis.

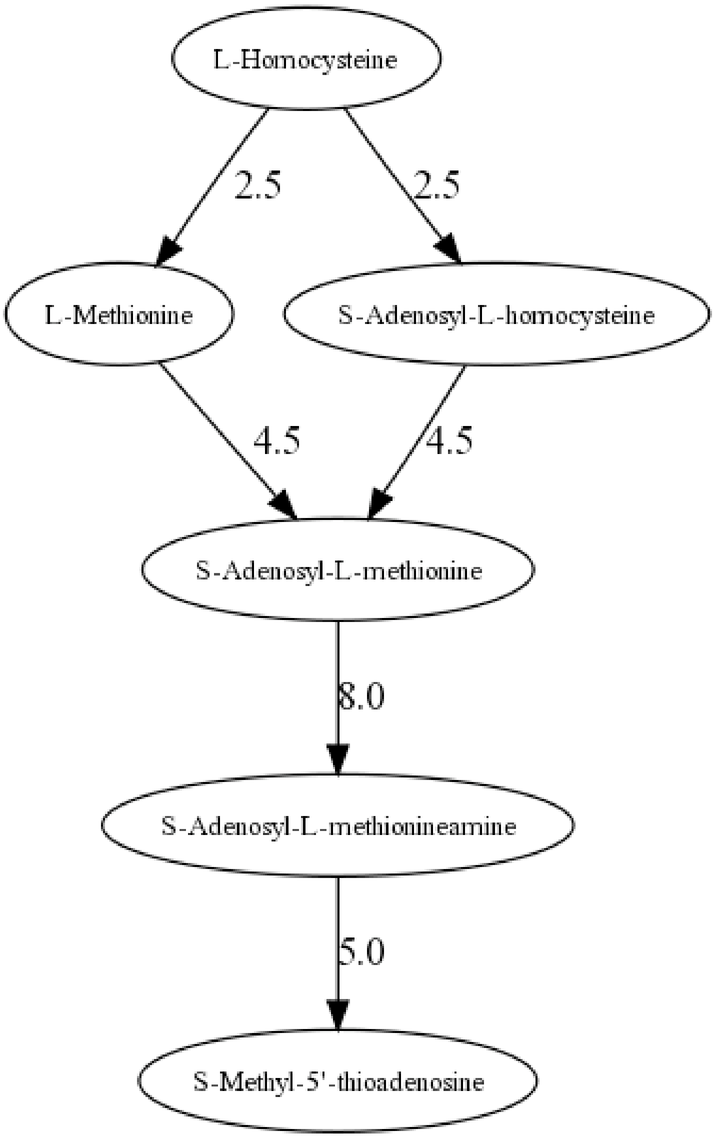

L-Homocysteine
2.5
2.5
L-Methionine
S-Adenosyl-L-homocysteine
4.5
4.5
S-Adenosyl-L-methionine
8.0
S-Adenosyl-L-methioninearnine
5.0
S-Methyl-5'-thioadenosine

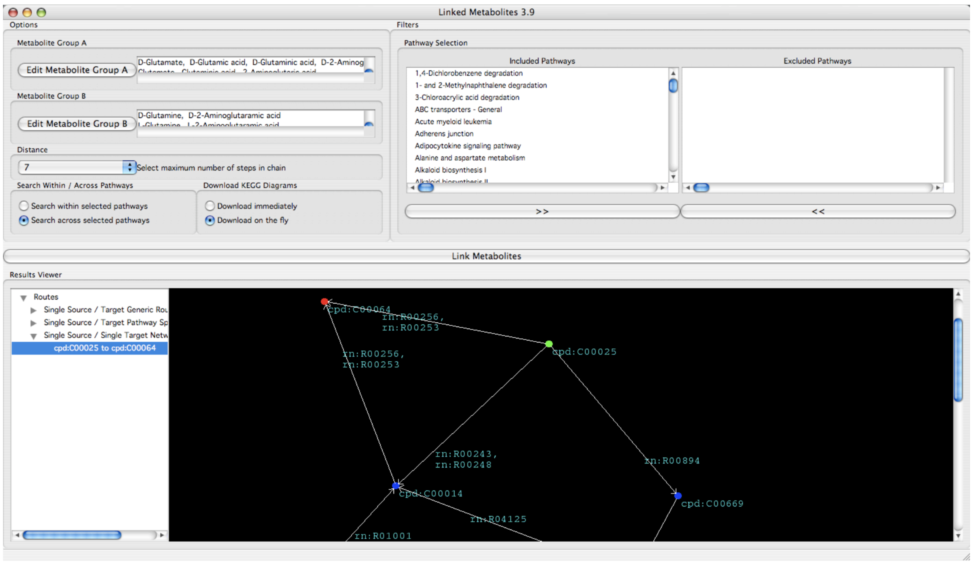
Figure 5.2: A directed metabolic graph with betweenness centrality values giving an indication of how the shortest paths through the graph are distributed

5.2 Method

Linked Metabolites, a graphical tool (Figure 5.3) written in Python was produced that creates directed metabolic graphs for two groups of compounds based on the KEGG pathway database.

Python, which is an interpreted rather than a compiled language and therefore not known for its speed, may not seem the obvious choice for this type of application. There were, however, several good reasons for its selection. Firstly, Python is cross-platform making it easy to circulate the code. This was important as although the application was not originally intended for release to the community, different members of the group work with a range of operating systems and supporting them all in a compiled language such as C would have been time consuming. Python also offered significant benefits when used for rapid prototyping; its ease of use and clear syntax, which are comparable to those of a scripting language, coupled with its support for object-oriented programing, allowed quite major changes to be made to the code in a relatively short time. Most critically, Python is based on C. Thus while Python itself is not particularly fast, natively compiled, third-party libraries written in C/C++ could be used for many of the important processing steps, greatly reducing the runtime. Of these, the most important was the Boost Graph Library[91], which contains fast implementations of a number of key graph algorithms including shortest paths, depth- and breadth-first searches, and checks for connected components.

The KEGG database was chosen for use in the tool as its diagrams of reference metabolic pathways allowed for the possibility that multi-species support may have needed to be added to the code at a later date. Unfortunately, it also meant that several compromises had to be struck. The KEGG database contains far more than just information on the metabolic pathways, including information on compounds, enzymes, and a number of annotated genomes. As a result, it would be impractical to expect every end-user to install complete, local copies of KEGG for



Metabolite Group A

D-Glutamate, D-Glutamic acid, D-Glutaminic acid, D-2-Aminog

Metabolite Group B

D-Glutamine, D-2-Aminoglutaramic acid

Distance

7 Select maximum number of steps in chain

Search Within / Across Pathways

○ Search within selected pathways

● Search across selected pathways

Download KEGG Diagrams

○ Download immediately

● Download on the fly

Pathway Selection

Included Pathways

1,4-Dichlorobenzene degradation
1- and 2-Methylnaphthalene degradation
3-Chloroacrylic acid degradation
ABC transporters - General
Acute myeloid leukemia
Adherens junction
Adipocytokine signaling pathway
Alanine and aspartate metabolism
Alkaloid biosynthesis I

Excluded Pathways

Routes
- Single Source / Target Generic Rou
- Single Source / Target Pathway Sp
- Single Source / Single Target Netw
  - cpd:C00025 to cpd:C00064

Figure 5.3: The graphical user interface of Linked Metabolites displaying a directed metabolic graph.

use with the tool. Fortunately, a SOAP web service is provided to allow developers to get around this problem, and is used by Linked Metabolites to obtain pathway diagrams, but the volume of data needed to perform the searches themselves was far too great to allow this type of access. As a result, Linked Metabolites relies on both the KEGG's web service and a locally-stored copy of the xml files for the pathways, plus definitions for the appropriate compounds and reactions. While this solution provides an acceptable balance between installed size and speed/runtime for the searches, it also generates several problems; Linked Metabolites must have a connection to the internet in order to function, the locally-stored information can quite rapidly fall behind the current version of KEGG being used by the web service forcing time-consuming updates, and debugging becomes very difficult as each user has a subtlety different version of the database (dependent on the time they last updated the stored files). Implementing the tool as a web service would have removed many of these issues as there would have been only one installation to update, however the processor and memory intensive nature of the searches and a lack of resources made this impractical at the time of coding.

Linked Metabolites constructs directed metabolic graphs in three stages. In the pre-processing stage potentially relevant information is extracted from the XML representations of the metabolic pathways for Homo sapiens and compiled into a series of temporary graphs, one for each pathway. The second stage of the algorithm involves the searching of those graphs for routes between pairs of compounds as selected from the source and target groups specified by the user. Finally, the post-processing stage combines the routes identified into directed metabolic graphs of varying forms; from a single source to a single target, from a single source to multiple targets, from multiple sources to a single target, and from multiple sources to multiple targets.

5.2.1 Pre-processing

The pre-processing stage of the algorithm (Figure 5.4) is concerned with the construction of the directed graphs on which the route searches will be performed from a user-defined subset of the XML representations of the KEGG pathway diagrams. Within a graph each compound is featured as a single node, regardless of the number of occurrences of that compound within the pathway diagrams themselves. Arcs linking the nodes are created based on substrate to product relationships between compounds in relevant reactions. Parallel arcs between compounds are permitted and occur if two compounds share a substrate to product relationship in multiple reactions. If the route search is to be performed within individual pathways a separate graph is created for each pathway, whereas if the search is to be performed across a set of pathways a single, combined graph is used. The graphs produced by this stage of the algorithm will henceforth be referred to as the "base graphs".

5.2.2 Route search

Once the base graphs have been compiled, the algorithm enters its second stage (Figure 5.5). A pre-search is performed on each base graph to ensure that at least one member of each compound group features within it. Pairwise searches are then performed using an algorithm devised by Rubin[92] to identify all the simple paths (where a simple path is a route in which each node can feature only once) between members of the two compound groups. Since the algorithm is $\mathrm{O}\big(V^3\big)$ it is important to reduce the size of the graph on which the search is performed as far as is possible, as such a cut-down graph for each of the pairwise searches is constructed from the appropriate base graph. The cut-down graph features only those nodes from the base graph for which the sum of the shortest path from the source node to the node of interest (calculated as a single step for each cut-down graph using the Bellman-Ford all shortest paths algorithm[93, 94])

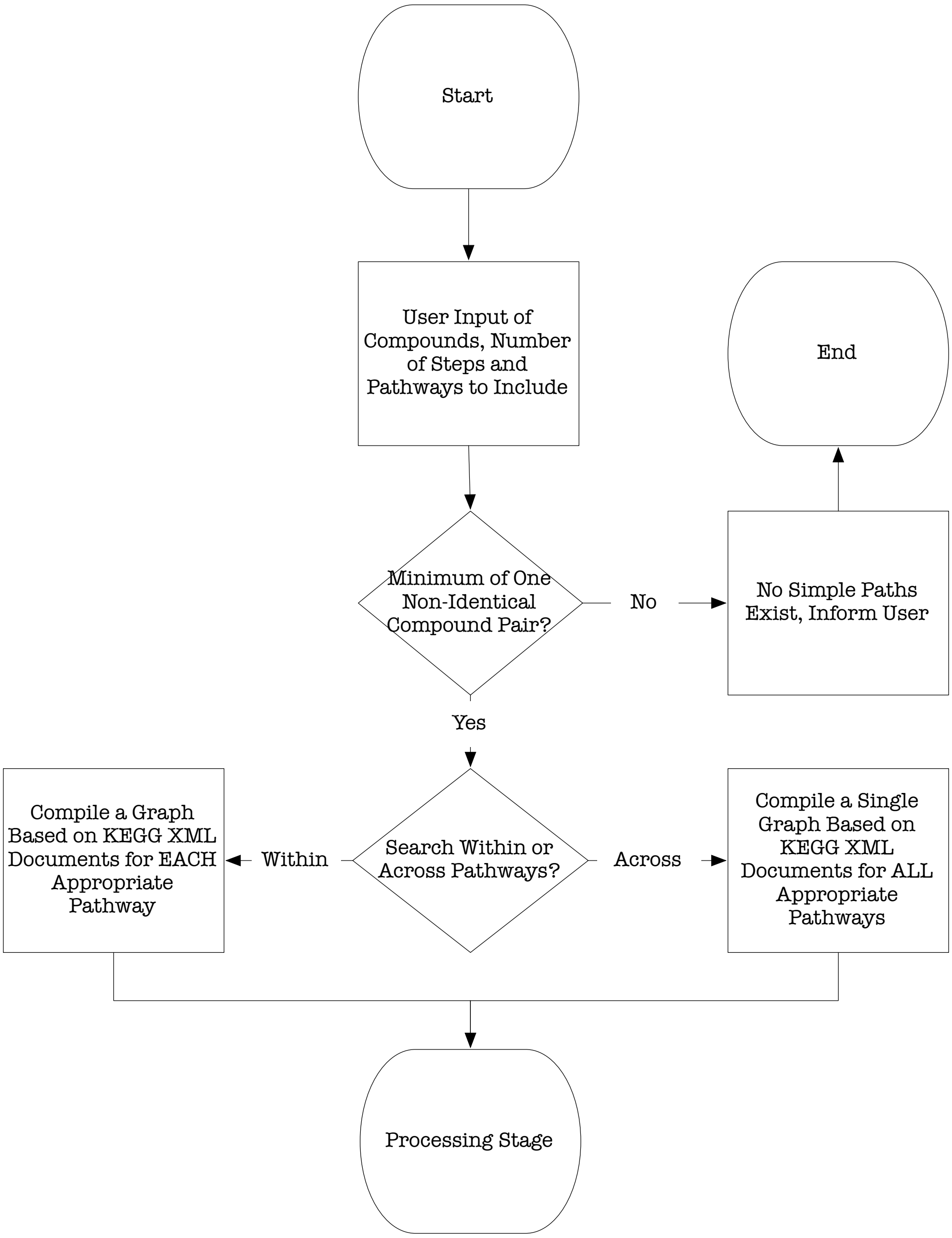

Start
User Input of Compounds, Number of Steps and Pathways to Include
Minimum of One Non-Identical Compound Pair?
No
No Simple Paths Exist, Inform User
End
Yes
Search Within or Across Pathways?
Within
Compile a Graph Based on KEGG XML Documents for EACH Appropriate Pathway
Across
Compile a Single Graph Based on KEGG XML Documents for ALL Appropriate Pathways
Processing Stage

Figure 5.4: Flowchart showing the pre-processing stage of the algorithm.

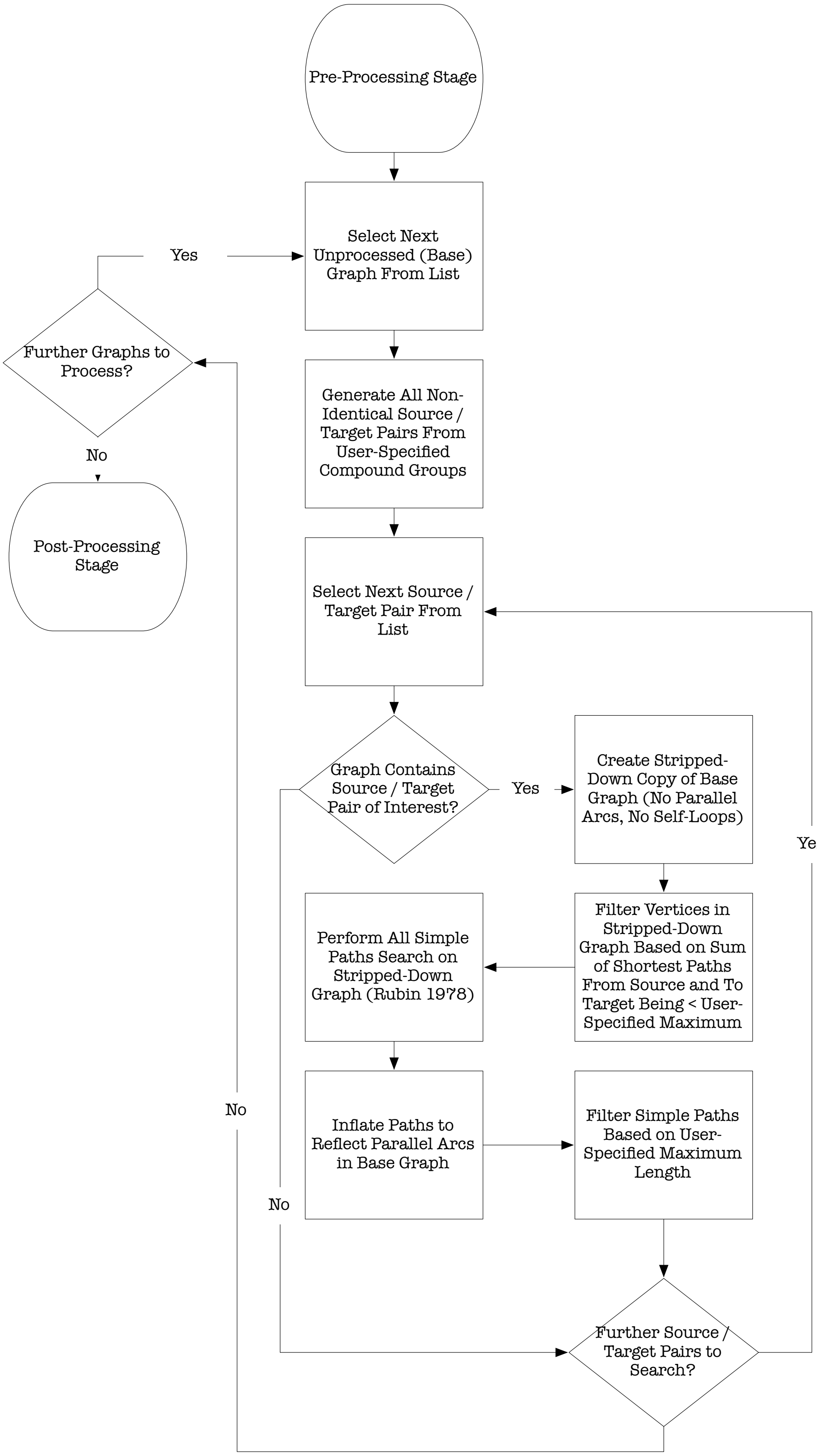

Pre-Processing Stage
Select Next Unprocessed (Base) Graph From List
Yes
Further Graphs to Process?
No
Post-Processing Stage
Generate All Non-Identical Source / Target Pairs From User-Specified Compound Groups
Select Next Source / Target Pair From List
Graph Contains Source / Target Pair of Interest?
Yes
Create Stripped-Down Copy of Base Graph (No Parallel Arcs, No Self-Loops)
Yes
Filter Vertices in Stripped-Down Graph Based on Sum of Shortest Paths From Source and To Target Being < User-Specified Maximum
Perform All Simple Paths Search on Stripped-Down Graph (Rubin 1978)
Inflate Paths to Reflect Parallel Arcs in Base Graph
Filter Simple Paths Based on User-Specified Maximum Length
No
No
Further Source / Target Pairs to Search?

Figure 5.5: Flowchart showing the main processing stage of the algorithm.

and the shortest path from the node of interest to the target (calculated individually for each node using Dijkstra's algorithm[95]) is within the maximum number of steps specified by the user. This filtering stage is safe, as if the shortest route from source to target via the node of interest is not within the maximum length, no simple paths of appropriate length will pass through the node. It should however be noted that inclusion at this point does not imply that all routes through a node will be valid, as many may be over-length. Since Rubin's algorithm does not support parallel arcs, only one arc (in each direction if appropriate) is included between compounds in the cut-down graph. Once the paths have been identified the arcs are then expanded recursively using the appropriate base graph in order to obtain the complete set of simple paths featuring all the possible combinations of reactions. Finally, any remaining routes that are longer than the maximum length are filtered out.

It should be noted that while a compound can feature in both the source and target groups, the algorithm will not search for routes between a compound and itself (as this would not be a simple path). Such paths may, however, appear in the results if a route with a compound as its source intersects with another route in the graph for which that same compound is a target.

5.2.3 Compilation and presentation of routes and directed metabolic graphs

In the final stage of the algorithm (Figure 5.6) the simple paths identified during the route search are returned to the user in several forms; as generic routes, in which objects (compounds or the reactions that link them) are not placed in any specific context, as pathway-specific routes, in which every object in the route is assigned to a particular pathway, and as directed metabolic graphs, which are combinations of the generic routes. While the first two forms can already be obtained via the KEGG PathComp tool, their identification is a necessary step in the generation of the networks and so it is appropriate to include them here.

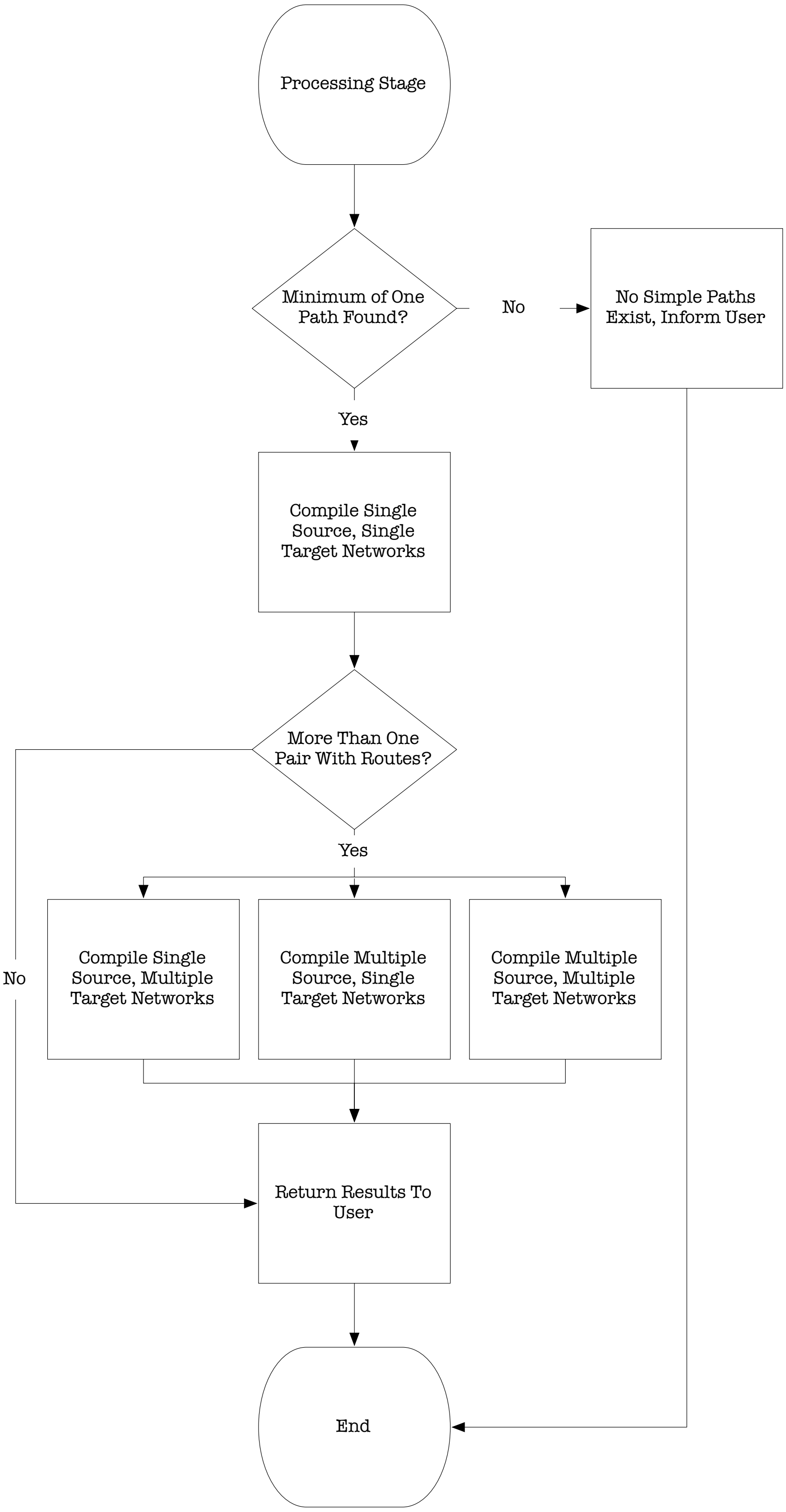

Processing Stage
Minimum of One Path Found?
No
No Simple Paths Exist, Inform User
Yes
Compile Single Source, Single Target Networks
More Than One Pair With Routes?
Yes
No
Compile Single Source, Multiple Target Networks
Compile Multiple Source, Single Target Networks
Compile Multiple Source, Multiple Target Networks
Return Results To User
End

Figure 5.6: Flowchart showing the post-processing stage of the algorithm.

Generic routes

The generic routes are the raw results of the route search, comprising chains of compounds and reactions with no additional contextual information. In many ways the generic routes are more useful than the more complicated pathway-specific routes that are generated from them, as they allow the user to easily see how many non-identical paths exist between the source and target compounds; a task that often becomes far more complex when considering routes that are effectively identical but are being presented from the point of view of a different metabolic pathway.

Pathway-specific routes

The pathway-specific routes are generated from the generic routes by the recursive expansion of each of the objects in a route according to the pathways in which it features. As any given reaction can occur in a number of different pathways, this process can result in a large number of pathway-specific routes being generated from each of the generic routes. Once identified (and before it is presented to the user), each of the pathway-specific routes is ranked based on the minimisation of a heuristic function:

$$Score = (\prod_{p=1}^{n} \frac{s_r}{s_p}) \times n - \frac{n}{s_r} \tag{5.1}$$

Where n is the number of metabolic pathways referred to by the route, s_p is the number of steps in metabolic pathway p, and s_r is the total number of steps in the route. The ranking is designed to promote shorter routes within a single pathway, routes involving small numbers of pathways or routes where the steps in the route are evenly distributed across the pathways involved, over those that feature larger numbers of steps spread over multiple pathway diagrams. The ranked pathway-specific routes are then presented to the user both as textual descriptions and coloured pathway diagrams obtained via the KEGG API.

Directed metabolic graphs

Directed metabolic graphs are generated from the generic routes based on the combination of compounds in the user-specified groups as previously discussed; single source to single target, multiple sources to single target, single source to multiple targets and multiple sources to multiple targets. The construction of the graphs themselves is a trivial process, purely a case of merging the appropriate generic routes. The directed metabolic graphs are returned to the user as images, in which compounds that are the source of routes are coloured green, targets are coloured red, and those compounds that are both sources and targets are yellow*. The remaining compounds are coloured blue. Graph layouts are calculated using the GEM algorithm[96].

Combined, the three stages form the complete algorithm (Figure 5.7).

5.3 Results

5.3.1 Comparison with KEGG PathComp

In order to obtain an estimate of the coverage of the tool, routes searches were performed between a series of compound pairs selected at random from the first one hundred compound codes in the KEGG database. It was necessary to use pairs of compounds as opposed to groups because KEGG PathComp, the tool against which the results were compared does not support searches between groups of compounds. Of the 200 compound pairs initially generated, 45 were rejected because they included a compound that did not feature in any of the metabolic

*In order to be coloured yellow a compound must be acting as a source and a target for routes in the graph, not just feature in both the user-specified source and target groups.

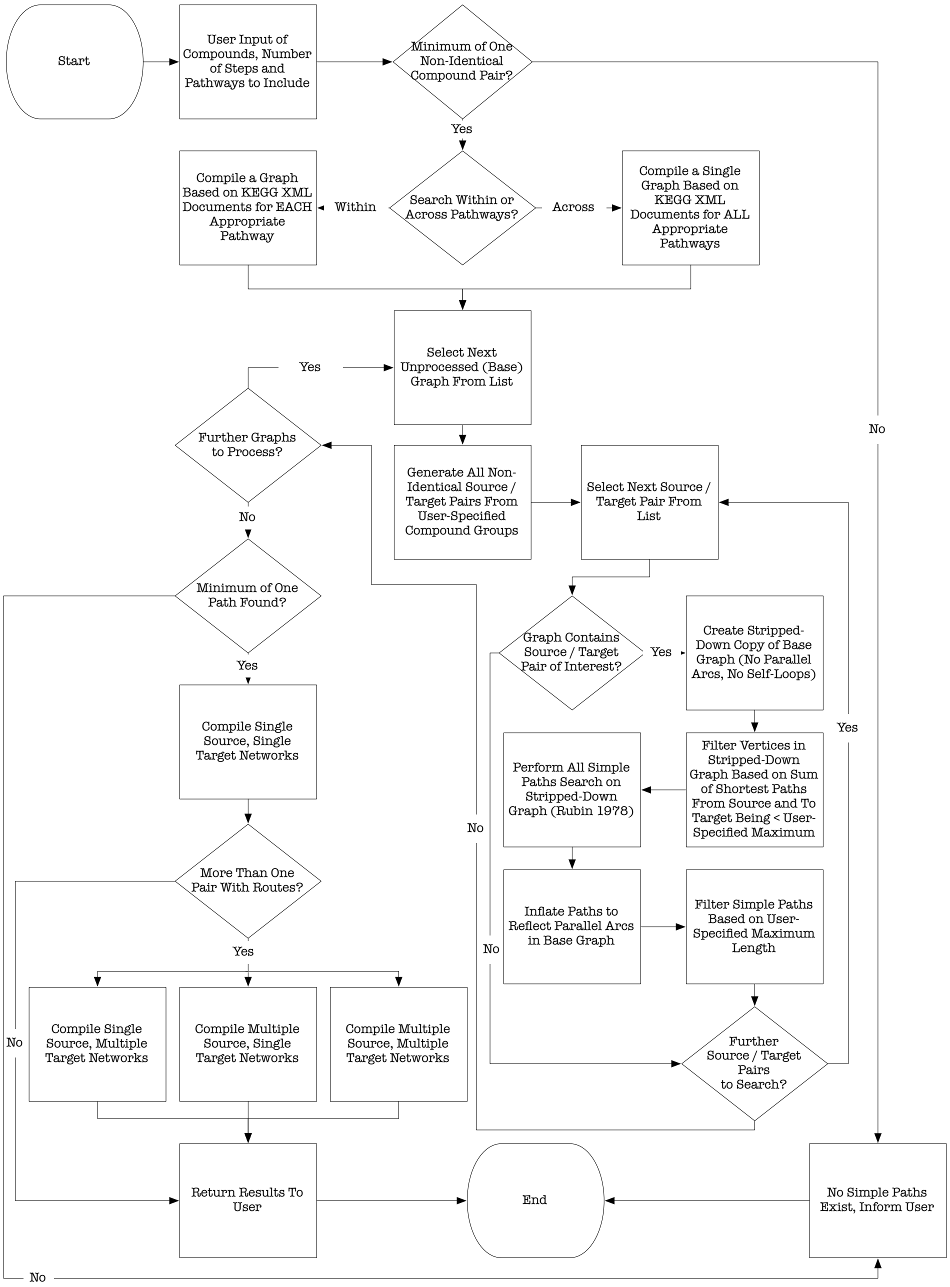

Start
User Input of Compounds, Number of Steps and Pathways to Include
Minimum of One Non-Identical Compound Pair?
Yes
No
Compile a Graph Based on KEGG XML Documents for EACH Appropriate Pathway
Within
Search Within or Across Pathways?
Across
Compile a Single Graph Based on KEGG XML Documents for ALL Appropriate Pathways
Select Next Unprocessed (Base) Graph From List
Yes
Further Graphs to Process?
No
Generate All Non-Identical Source / Target Pairs From User-Specified Compound Groups
Select Next Source / Target Pair From List
Minimum of One Path Found?
Yes
No
Graph Contains Source / Target Pair of Interest?
Yes
Create Stripped-Down Copy of Base Graph (No Parallel Arcs, No Self-Loops)
Compile Single Source, Single Target Networks
Perform All Simple Paths Search on Stripped-Down Graph (Rubin 1978)
Filter Vertices in Stripped-Down Graph Based on Sum of Shortest Paths From Source and To Target Being < User-Specified Maximum
Yes
No
More Than One Pair With Routes?
Yes
No
Inflate Paths to Reflect Parallel Arcs in Base Graph
Filter Simple Paths Based on User-Specified Maximum Length
Compile Single Source, Multiple Target Networks
Compile Multiple Source, Single Target Networks
Compile Multiple Source, Multiple Target Networks
Further Source / Target Pairs to Search?
No
Return Results To User
End
No Simple Paths Exist, Inform User
No

Figure 5.7: Flowchart showing the complete algorithm used by Linked Metabolites.

pathways. Each of the remaining 155 pairs was then randomly assigned a maximum path length of between four and eight steps. Searches for routes (up to the assigned length) between each compound pair were then performed using Linked Metabolites and KEGG PathComp and the results summarised (Table 5.1, Figure 5.8).

Overall, Linked Metabolites found nearly twice as many routes between the 155 compound pairs as KEGG PathComp, although PathComp found routes between a higher proportion of pairs. In those cases where PathComp outperformed Linked Metabolites the cause was usually that PathComp was finding routes via compounds or reactions that were not included in the metabolic pathways and were, therefore, unavailable to Linked Metabolites when its searches were performed.

	KEGG PathComp	Linked Metabolites
Total number of routes found	628	1152
Percentage of pairs for which routes were found	29.7	19.4
Average number of routes per pair	4.1	7.4

Table 5.1: Summary of the results of the random route searches.

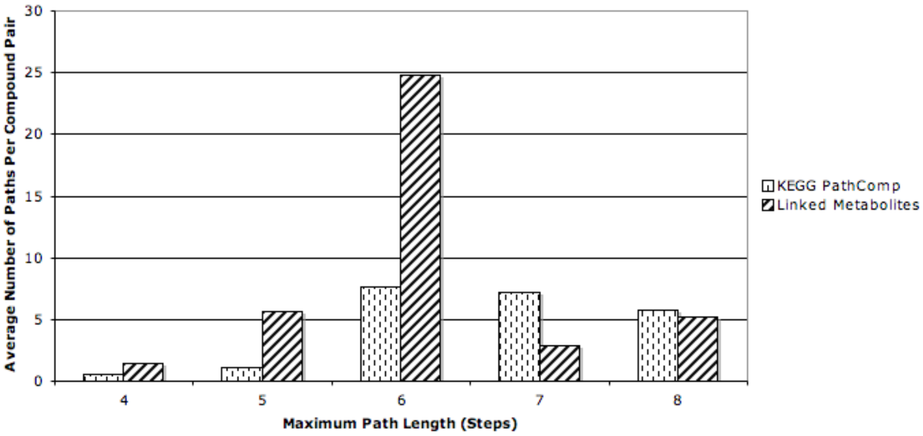


Figure 5.8: Comparison between Linked Metabolites and KEGG PathComp of average number of paths found per compound pair.

5.3.2 Example usage: Investigation of the biology underpinning a set of ratios used to differentiate between tumour types

In a study by Harris *et al.*[25] the ratios between the heights of peaks due to four metabolites in short-echo-time, single-voxel MRS spectra were used to differentiate between three types of childhood cerebellar tumour. Here we will use Linked Metabolites to investigate the relationships between those compounds in the metabolic pathways, in order to suggest why those ratios are able to discriminate between the different groups.

N-acetyl-L-aspartate and creatine

In their paper, Harris *et al.* used the N-acetyl-L-aspartate (NAA) to creatine (Cr) ratio to differentiate between pilocytic astrocytomas and the other tumour types being studied (medulloblastoma and ependymoma). The cut-off point was set at 4, with higher ratios indicating that the tumour was an astrocytoma. This could suggest that the rate of conversion from NAA to Cr is slower in astrocytomas than it is in medulloblastomas or ependymomas. The Linked Metabolites search for routes between NAA and Cr produced a relatively diverse directed metabolic graph (Figure 5.9) made up of more than a dozen generic routes, or over two hundred pathway-specific routes. Fortunately, a number of these can be rejected based on compartmentalisation issues. Figure 5.10 shows a route involving the alanine & aspartate and arginine & proline metabolism pathways. Within it, NAA is converted to aspartate and then on to Cr via L-argininosuccinate, arginine and guanidinoacetate. Based on the information available in KEGG, this route is quite plausible; all of the enzymes involved are coded for by the human genome, and it even follows the recognised entry points between the pathways rather than simply moving between instances of the same compound. However, the reactions between aspartate and arginine form part of the urea cycle, a pathway that takes place in the liver, and therefore is unlikely to be appropriate for

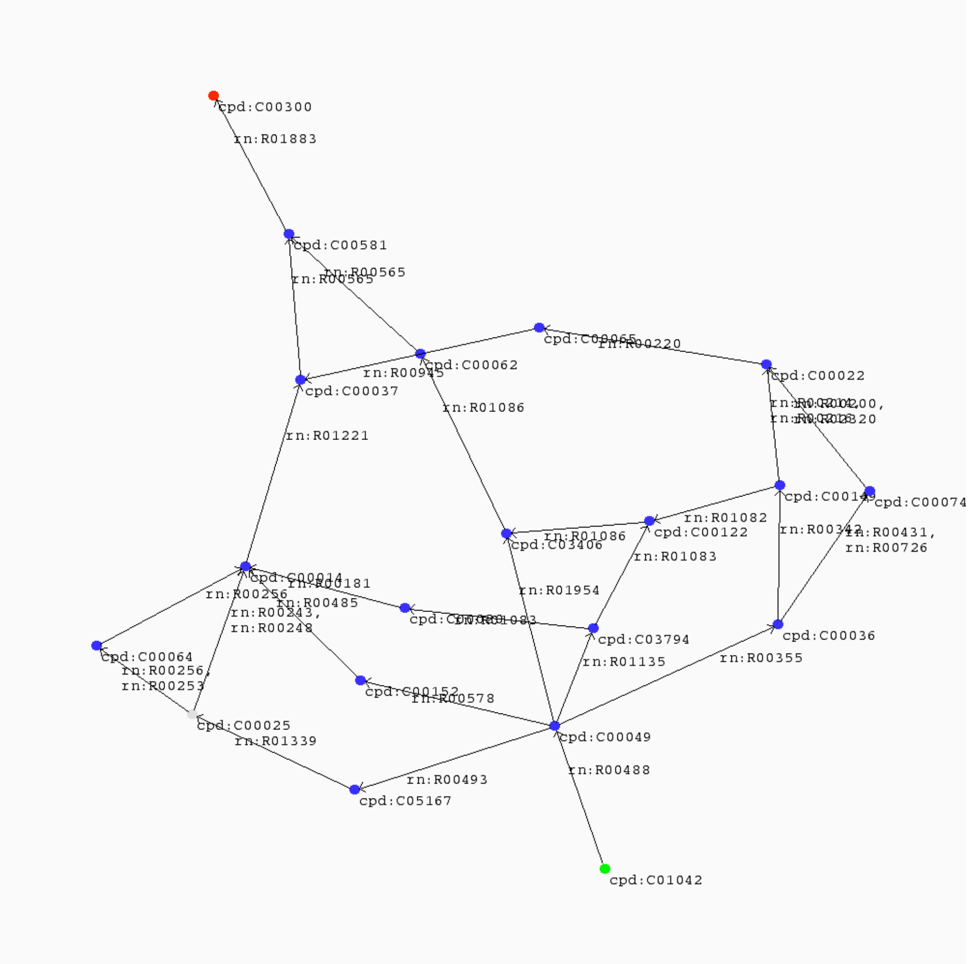


Figure 5.9: Directed metabolic graph produced by Linked Metabolites showing the routes up to a maximum of 8 steps long from N-acetyl-L-aspartate (cpd:C01042) to creatine (cpd:C00300).

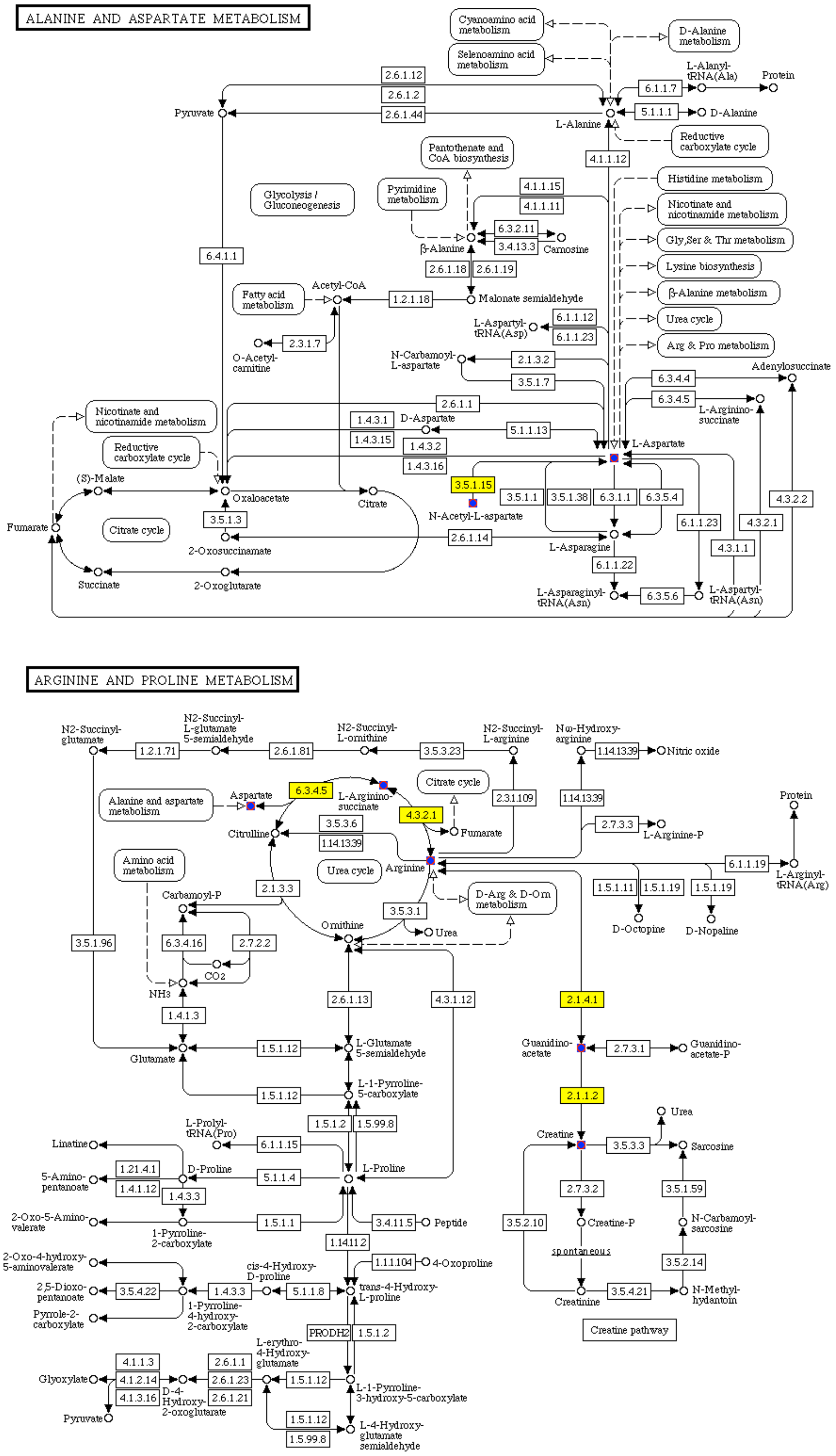


Figure 5.10: KEGG pathway diagrams showing a route found by Linked Metabolites between N-acetyl-L-aspartate and creatine. The route involves the alanine & aspartate metabolism and arginine & proline metabolism pathways.

use when explaining the behaviour of metabolites in the brain. While Linked Metabolites allows the user to specifically remove pathways from its search it would not have helped in this case, since KEGG's version of the arginine and proline metabolism pathway contains an instance of the urea cycle itself and that local version of the pathway could not be removed from the search without removing the entire arginine and proline metabolism pathway. The nesting of pathways in this way is a very significant issue, which has no easy resolution since the ways in which pathways are defined can have a huge impact on their usefulness in different types of task. KEGG's pathways are arranged in a way that best supports manual interpretation by a domain expert; however for computational tasks, redefining the pathways based on a more physical criteria such as the conservation of moiety used by Arita would be far more appropriate. This is also a good illustration of why the results of a Linked Metabolites search are only intended as a guide to possible chains of reactions that may be involved in an observed biological effect, the manual checking and interpretation of the results is still an important step.

Figures 5.11 & 5.12 show another two examples of pathway-specific routes between NAA and Cr. The route in figure 5.11 involves the alanine & aspartate metabolism, pyruvate metabolism and glycine, serine & threonine metabolism pathways, whereas the route in figure 5.12 replaces pyruvate metabolism with the glycolysis / gluconeogenesis pathway. However, as with the urea cycle previously, this is not because they are different routes; it is instead due to KEGG's version of the glycolysis / gluconeogenesis pathway including elements of pyruvate metabolism. Fortunately, unlike the earlier example all these pathways can take place in the brain and therefore the routes are still valid. The first point of interest to be considered here is that in the alanine & aspartate metabolism pathway, the route includes a compound (oxaloacetate) that is involved in the citrate cycle (also know as the TCA cycle), a major component of energy metabolism. Alterations to energy metabolism between tumour types are common, due to a wide range of factors such as the aggressiveness of the tumour, the type of tissue it originated in,

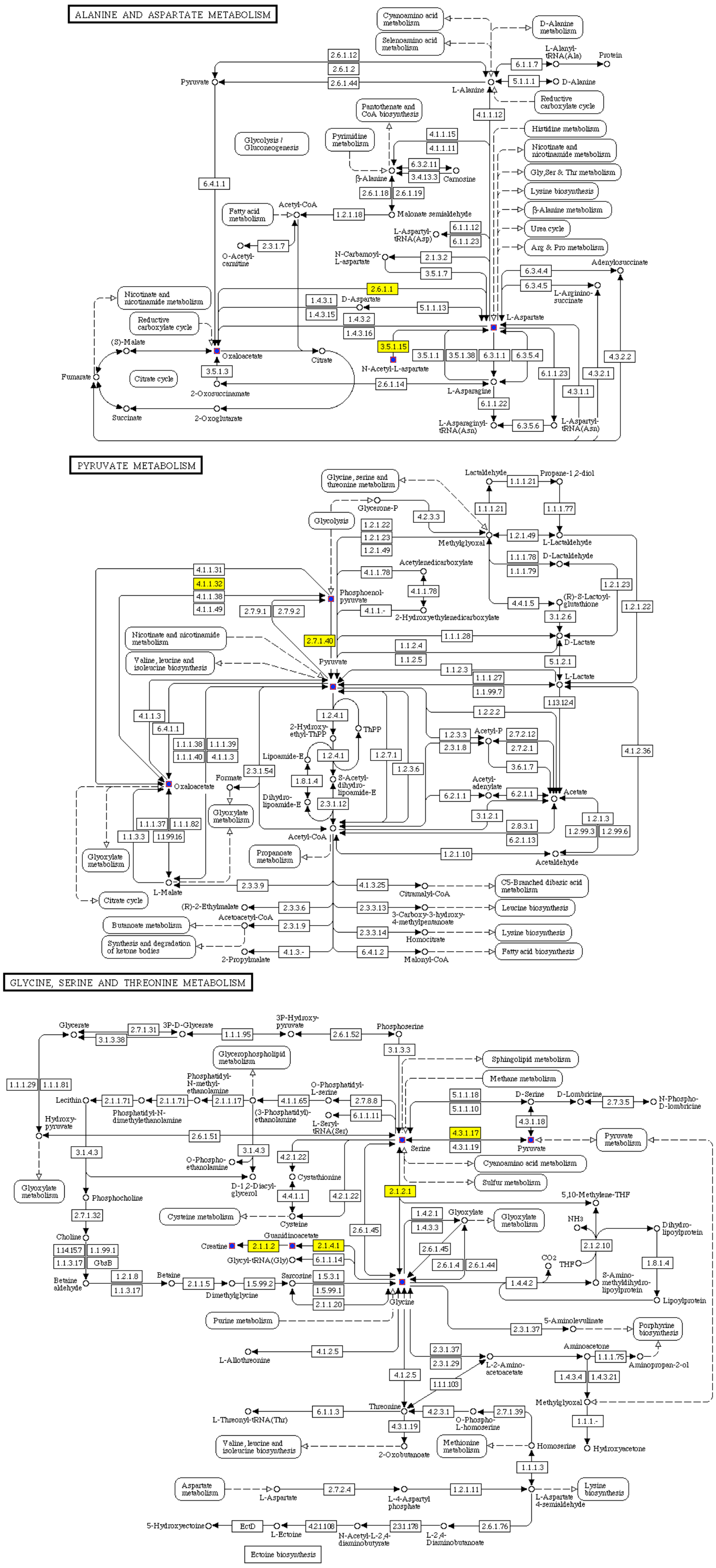


Figure 5.11: KEGG pathway diagrams showing a route found by Linked Metabolites between N-acetyl-L-aspartate and creatine. The route involves the alanine & aspartate metabolism, pyruvate metabolism and glycine, serine & threonine metabolism pathways.

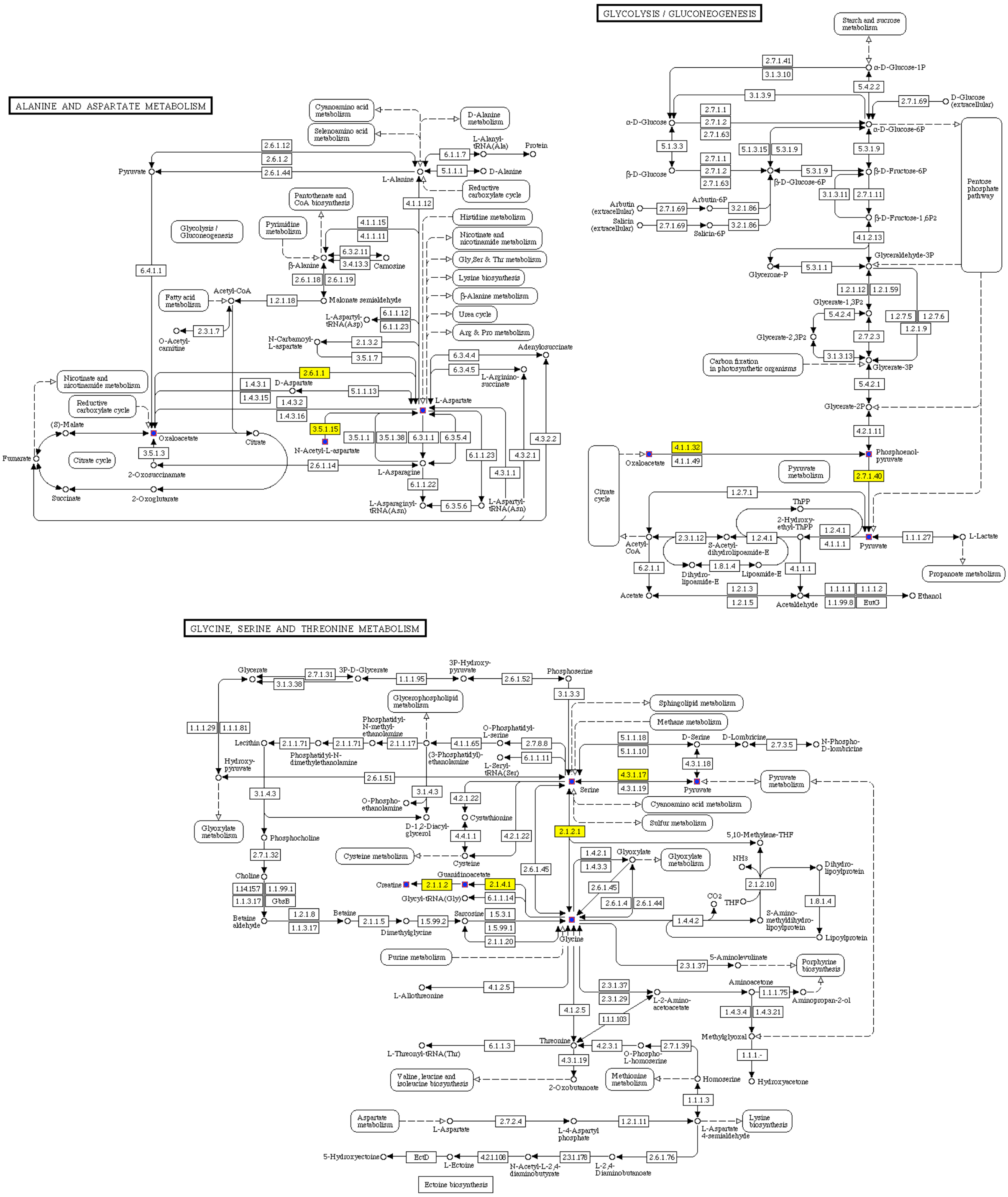


Figure 5.12: KEGG pathway diagrams showing a route found by Linked Metabolites between N-acetyl-L-aspartate and creatine. The route involves the alanine & aspartate metabolism, glycolysis / gluconeogenesis and glycine, serine & threonine metabolism pathways.

and the availability of sufficient resources for continued growth. It seems sensible therefore, that features in pathways responsible for energy metabolism would be able to differentiate between tumour types to some extent. This hypothesis would also apply to changes involving the glycolysis / gluconeogenesis pathway, as are suggested by the version of the route in figure 5.12.

Next, we should consider how a slowing of this pathway in pilocytic astrocytomas relative to medulloblastomas and ependymomas would affect the concentrations of compounds that feature within it, such as NAA and glycine. If there were a low rate of flow through the pathway (as we predict to be the case for pilocytic astrocytomas), then there should be a comparatively low concentration of glycine, owing to the smaller amount of serine, the compound required for its formation. We would also expect to see a buildup of NAA at the entry to the pathway. By comparison, in the cases where the pathway is supposed to be flowing more quickly (as we predict to be the case for medulloblastomas and ependymomas) we should see a larger concentration of glycine, because more serine will be available, but we should also see a drop in the concentration of NAA, because it is now able to to enter and be passed down the pathway. Precisely these results; high NAA and low glycine in astrocytomas, low NAA and high glycine in medulloblastoma and ependymomas, have been reported in connection with *in vitro* studies of these tumour types[97].

As this search is based on compounds that are being used in ratios rather than compounds selected based on correlation networks, it is quite possible that the difference in their relative concentrations is not due to an alteration in the biochemistry linking them and is instead due to some external process. Fortunately, an examination of the directed metabolic graph of these routes (Figure 5.9) can in this case show us other compounds that owing to the topology of the graph, should behave in the same way as the compounds involved in the ratios (assuming the change in relative concentration of the compounds is due to an alteration somewhere in the routes Linked Metabolites has identified). In this case, we can see that all the routes involve aspartate (cpd:C00049) and therefore it should also work as part of the ratio in place of NAA, although it will of course have a different cut-off value. Likewise guanidinoacetate (cpd:C00581) should act as an appropriate substitute for Cr. This information could also have been extracted from the directed metabolic graph automatically using a betweenness centrality algorithm (Section 2.2.5), a good example of the potential strength of the graph theory approach.

For completeness, a search was also performed for routes running from Cr to NAA although nothing appropriate was found.

Choline and creatine

Linked Metabolites also found a route between choline (Cho) and Cr, within the glycine, serine & threonine metabolism pathway (Figure 5.13). This route is well recognised in connection with tumours due to its role in the formation of S-adenosyl-L-methionine (SAM), a coenzyme involved in the transfer of methyl groups during DNA methylation (a process that can alter the regulation of gene expression)[2]. This was previously mentioned in the work on dab liver tumours (see section 4.3.2). In the classification chart devised by Harris *et al.* the threshold values for the Cr/Cho ratio differed significantly between the pilocytic astrocytomas (where the Cr/Cho ratio was specified as less than 0.35) and the other two tumour types (where the threshold value used to differentiate between them was 0.75). The lower threshold value for the pilocytic astrocytomas suggests that the difference in the relative amounts of Cho and Cr in this tumour type is greater than it is in the other two, and this could be due to either a larger amount of Cho, a smaller amount of Cr, or a combination of both of these changes. Looking at the route between the two compounds identified by Linked Metabolites (Figure 5.13), it can

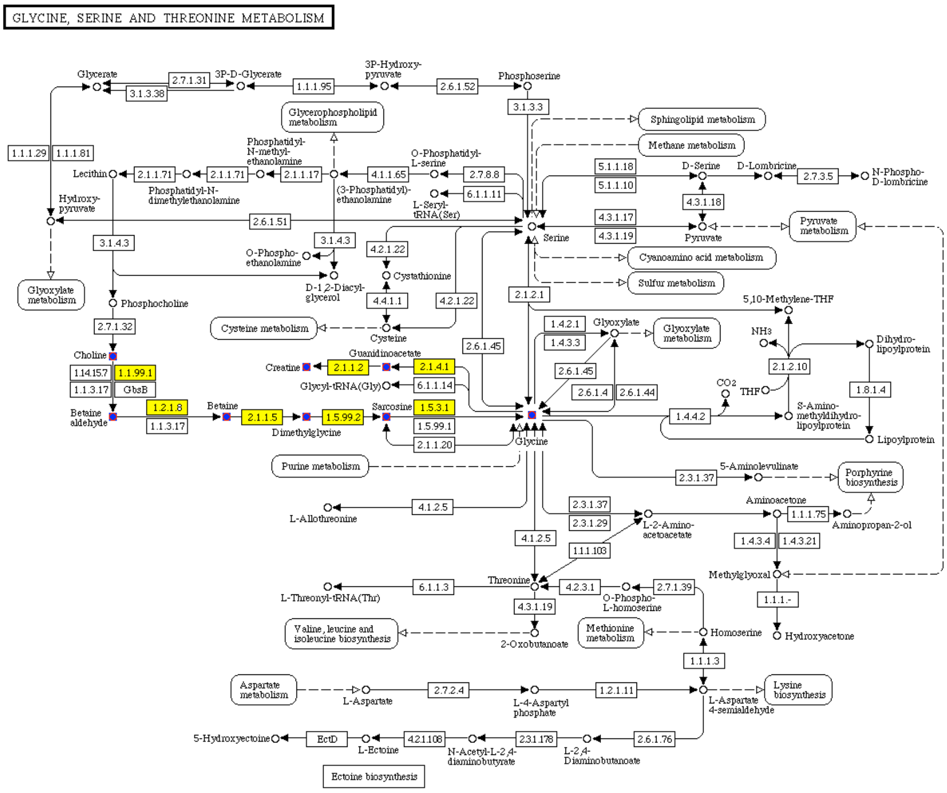


Figure 5.13: KEGG pathway diagrams showing a route found by Linked Metabolites between choline and creatine.

be seen that Cho forms Cr and thus it seems likely that the rate of flow through this pathway is slower in pilocytic astrocytomas than in the two other tumour types. A lower rate of flow would explain the greater difference in the relative amounts of the compounds (Cho and Cr) as a larger amount of Cho will build up at the entry to the pathway. Furthermore, the betaine to dimethylglycine step of the pathway is the link between it and the formation of SAM, as mentioned earlier. A lower rate of flow of this pathway in pilocytic astrocytomas, would imply a lower rate of formation of SAM, and therefore a lower rate of DNA methylation than in the two other tumour types. It has been reported that the number of hypermethylated (overmethylated) genes increases with the malignant potential of primary tumours[98], and this fits well with the low rate of methylation that was predicted (based on Linked Metabolites' route) for the pilocytic astrocytomas (a low-grade, and therefore relatively benign, glial tumour[22]) relative to the more aggressive medulloblastomas.

As the route between Cho and Cr intersects with those found previously between NAA and Cr, the directed metabolic graph for those routes can be expanded to include it (Figure 5.14). The fact that the routes intersect adds further weight to the idea that the alterations in the underlying biochemistry responsible for the difference between these tumour types may be related to the routes identified by Linked Metabolites. It also suggests that the point of intersection, arginine (cpd:C00062), may be an interesting target for further investigation.

The Cho and Cr peaks seen in a ^{1}H MRS experiment are actually a combination of the peaks for several different Cho and Cr containing compounds, merged due to the comparatively low resolution of the current technology. By extending the Linked Metabolites search from a maximum of 8 to a maximum of 9 steps we can include more of these constituent compounds (Figure 5.15) and as before, extend our directed metabolic graph (Figure 5.16).

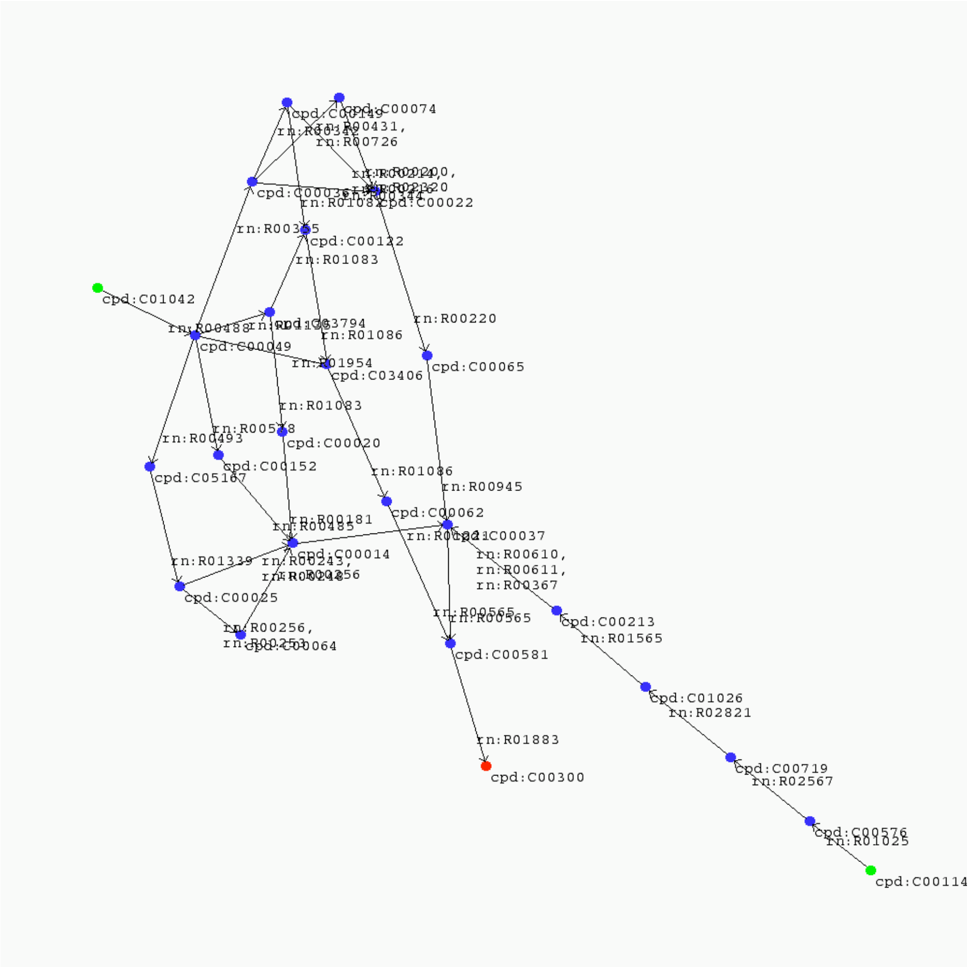


Figure 5.14: Directed metabolic graph produced by Linked Metabolites showing the routes up to a maximum of 8 steps long from N-acetyl-L-aspartate (cpd:C01042) and choline (cpd:C00114) to creatine (cpd:C00300).

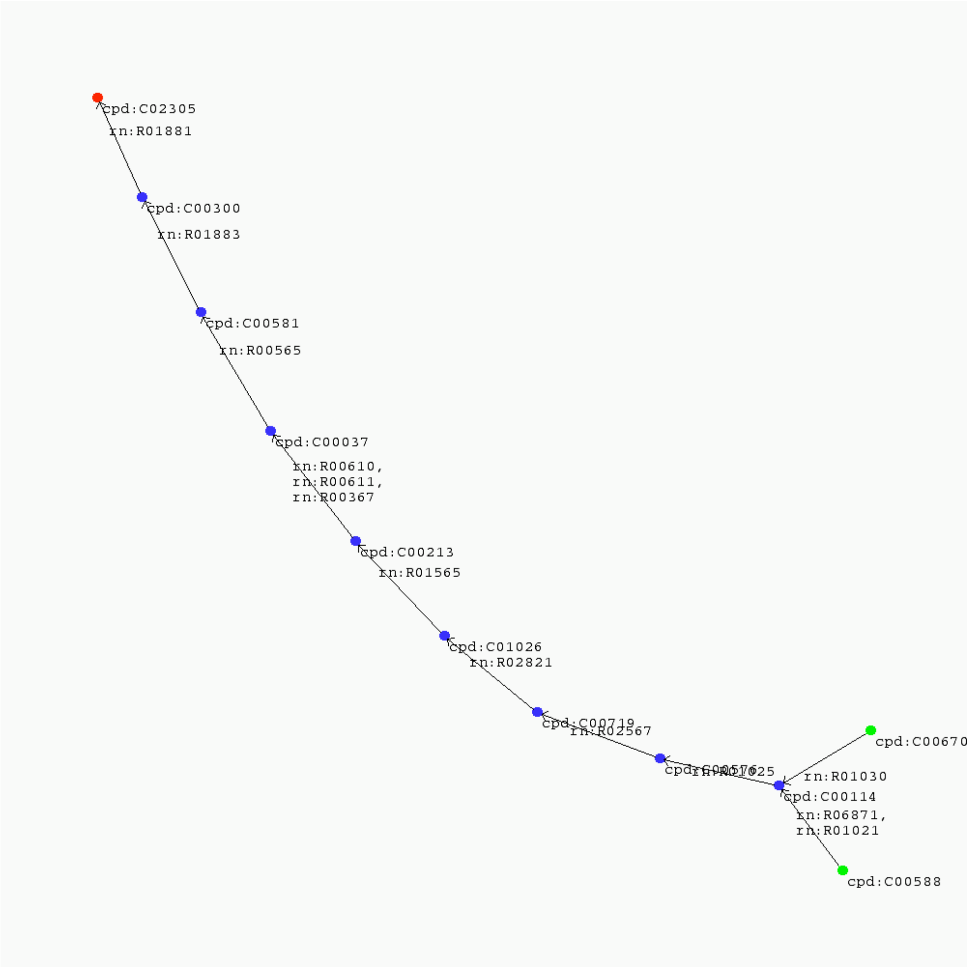


Figure 5.15: Directed metabolic graph produced by Linked Metabolites showing the routes up to a maximum of 9 steps long from phosphocholine (cpd:C00588) and glycerophosphocholine (cpd:C00670) to phosphocreatine (cpd:C02305).

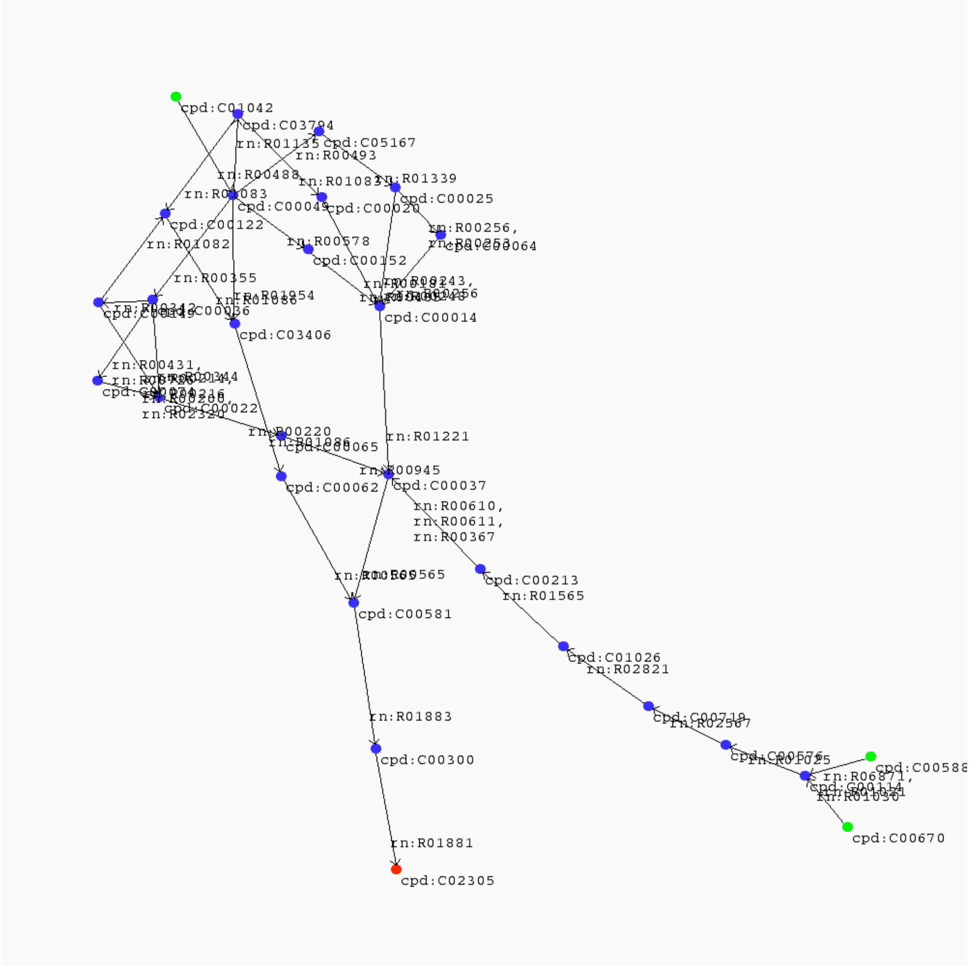


Figure 5.16: Directed metabolic graph produced by Linked Metabolites showing the routes up to a maximum of 9 steps long from phosphocholine (cpd:C00588), glycerophosphocholine (cpd:C00670) and N-acetyl-L-aspartate (cpd:C01042) to phosphocreatine (cpd:C02305).

5.4 Discussion

The directed metabolic graphs produced by Linked Metabolites have a range of potential applications. In metabolomics, multivariate analysis techniques such as PCA and partial least squares discriminant analysis are commonly used in combination with significance tests to produce a list of metabolites that contribute to the differences between samples. Subsequent interpretation of this list, with the aim of finding major metabolic perturbations, is often a time consuming process whereby the metabolites are "grouped" manually by inspection of known metabolic pathways. When used in conjunction with clustering metrics, the directed metabolic graphs produced by Linked Metabolites would offer a rapid, automated solution to this problem.

In the diagnosis and treatment of paediatric brain tumours, MRS can be used to gather quantitative information on the concentrations of key chemicals non-invasively. Analysis of this data will often indicate pairs of compounds that appear to be elevated in concentration for a particular tumour type, although small numbers of samples can make achieving statistical significance difficult. The directed metabolic graphs produced by Linked Metabolites can assist in the study of these compounds, firstly by providing information on possible biochemical relationships between them and secondly by helping to identify other compounds that may also have altered in concentration but that can not be seen using MRS.

Correlation networks, such as those discussed in the previous chapter, are often used to identify compounds within a dataset that may be involved in common processes. Unfortunately the presence of a correlation between two compounds does not imply that there is an obvious link in the underlying biochemistry; as Steuer, Kurths, Fiehn and Weckwerth remarked "...there is no straightforward connection between the observed correlations and the underlying reaction network. We observe strong correlations between seemingly distant metabolites, whereas metabolites sharing a common reaction are not necessarily correlated."[83] By combining information from multiple metabolic pathways and presenting only those routes of potential relevance to the compounds under investigation, Linked Metabolites can greatly reduce the complexity of the search for chains of reactions that may be responsible for an observed correlation.

When considering the potential impact of Linked Metabolites, one must first consider how a list of metabolites can be grouped in a way that is appropriate for answering a particular type of question. In the simplest case, where purely exploratory work is being undertaken and no further information is available, all the compounds could be placed in both groups. This would result in the complete set of routes from every compound to every other being generated, allowing the broadest possible base for the application of metrics such as motif search. Alternatively, one may wish to investigate how changes in the concentration of one compound might affect the concentrations of each of the others; this could be achieved by building a series of directed metabolic graphs with a single sources and multiple targets, rotating the source compound for each graph until the entire list of compounds has been covered. In situations where an earlier analysis of the experimental data has revealed compounds whose concentrations are elevated relative to another set of compounds, one might wish to use those groups as the basis for two directed metabolic graphs, one focusing on the routes from the group of compounds with elevated concentrations to the rest and *vice versa*. If the previous analysis involved the construction of correlation networks then several possibilities exist; a directed metabolic graph of all the possible routes between the compounds could be created as described above and then compared to the correlation network by searching for common subgraphs. If Pearson correlations were used, then it may also be appropriate to build graphs showing routes between the positively and negatively correlated compound groups for each compound. If clearly defined clusters exist within the correlation network, then these too could be used as the basis for a search.

At this point it is also necessary to consider the limitations of Linked Metabolites and the areas in which there is room for development in the future. Firstly, since KEGG does not contain information on the tissues or cellular compartments in which particular reactions take place, none of this information is considered during the route search. This is not an issue when the user has chosen to search for routes that take place within individual metabolic pathways, as the experimentally determined nature of the reference pathways means that compartmentalisation is already effectively dealt with; however, when searching across a set of pathways the way in which Linked Metabolites condenses instances of the same compound from a range of pathways into a single node means that a pathway that produces a given compound in one type of tissue could then be linked to a pathway that breaks the same compound down in a completely different area of the body, without any guarantee that there is a suitable transportation method between the two. This problem is compounded when in the last stage of the algorithm the identified routes are combined to form the directed metabolic graphs, again potentially mixing processes taking place in completely different areas of the body. There are both pros and cons to this behaviour. The obvious disadvantage is that routes and networks are likely to be generated that contain errors, linking compounds via intermediaries that can never interact, at least in the context from which the experimental data was gathered. On the other hand, the original motivation for the creation of Linked Metabolites was that under certain circumstances such as disease or external stress, changes to body chemistry might result in novel pathways temporarily becoming significant and as such "blue-sky" suggestions for interactions that might possibly take place could provide useful information. This would be particularly relevant in diseases such as cancer, where the activation and suppression of particular genes may drastically alter the enzymes present in a particular tissue type, and hence the reactions that could be taking place. Despite this, there is a strong argument for some form of compartmentalisation to be incorporated into the search, at least at the tissue level. This could be an optional criteria in the same vein as the selection of pathways for inclusion, thus allowing the current behaviour to remain if required.

Unfortunately, such a modification would require the database on which Linked Metabolites' search is based to be changed from KEGG to something more specialised such as Reactome, making it more difficult to add support for additional species; a feature that would certainly be of interest to the biological community, although possibly of less interest to those involved in medical research. Next we have the issue of how to make the graphs produced by Linked Metabolites available outside the program, as currently they can only be accessed through the internal view created after the search has been performed. It would be desirable for the directed metabolic graphs to be exported in a well-supported graph language, such as the graph modeling language (GML) so that graph theory tools can be easily applied in other applications; however, in order to facilitate the testing of the practicalities of the routes, it would also be advantageous to export the graphs in a format such as SBML, allowing them to be imported into kinetic modeling packages. In this case it would be sensible to include some other basic information in the model such as rate constants for each of the reactions, an operation that would not be difficult if the correct database was used.

The incorporation of the building of correlation networks into the workflow of Linked Metabolites would be an important step in its development for the biological and medical research communities, allowing researchers to enter their data, view the correlations, and then base their searches on this information. Figure 5.17 shows how this might be achieved in a future version of Linked Metabolites, while appendix B contains prototypes for some of the key screens.

One of the key contributions of Linked Metabolites to the field is in the way that it searches for routes between compounds. Many of the existing database search tools use an all-pairs shortest paths algorithm for this process, as it is relatively quick to compute. Linked Metabolites by comparison uses an all simple paths search, which is far more computationally intensive (indeed for many searches the memory requirements will make the problem intractable) but will find

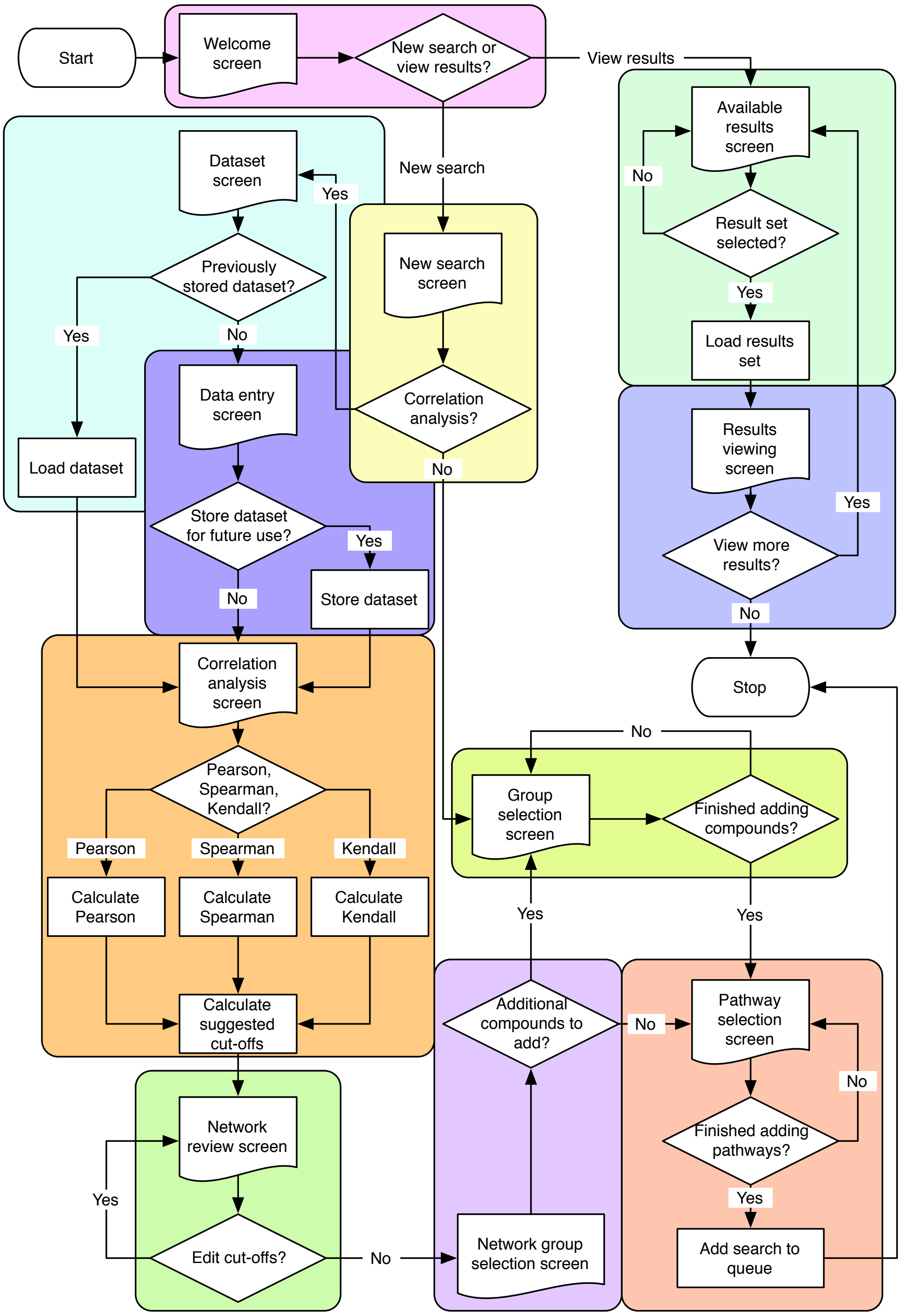


Figure 5.17: Flowchart showing how the workflow for a future version of Linked Metabolites might incorporate the construction of metabolic correlation networks. The coloured blocks represent different screens in the interface.

not only the shortest route between a pair of compounds, but also the next shortest and so on. All of these additional routes, which would be missed by a route search using only an all-pairs shortest paths algorithm, may in fact contain important information; particularly when considering situations in which pathways that would normally be dominant are being altered by disease or stress conditions.

5.5 Conclusions

This chapter has introduced Linked Metabolites, a tool that allows researchers to examine the relationships between groups of metabolites in the context of the Homo sapiens specific versions of the metabolic pathways stored in the KEGG database. It has shown a comparison between the route search capabilities of Linked Metabolites and those of another tool based on the same database, KEGG PathComp. It gave an example of how a researcher might use Linked Metabolites to investigate the relationships in the underlying biochemistry between compounds that appear to differentiate between tumour types. Finally, it has shown how despite some areas that still need development, the approach used by Linked Metabolites could be applied in a variety of different fields and how one might go about representing problems in ways that would be compatible with the use of the tool.

CHAPTER 6

DIRECTED METABOLIC NETWORKS FROM GENE EXPRESSION DATA

6.1 Background

Early in this thesis, we described metabolomics as "complementary" to other -omics technologies. So far we have seen how metabolomics datasets can be used to construct directed metabolic graphs, with the aim of identifying key features in the underlying biochemistry. These can then be used to help explain the observed biological changes in the subjects. In this chapter we will consider how this approach might be extended in the future, using additional information from other -omics technologies to help us limit the scope of our search of the metabolic pathways still further. Specifically, we will see how a directed metabolic graph might be used to visualise the potential metabolic changes resulting from differential gene expression, and how we might then go about integrating that information into the results of searches from Linked Metabolites.

6.2 Method

6.2.1 Data

As gene expression data is costly to gather and the early-stage nature of this work required the use of a relatively small microarray, the raw data used in this chapter was taken from three previously published studies by Grützmann *et al.*[99], Groene *et al.*[100] and Smirnov *et al.*[101], and downloaded from the European Bioinformatics Institute's (EBI) ArrayExpress service[102]. In all three cases, the data was gathered using either the Affymetrix Human Genome Focus (HGF) array or a larger array that contains the probes of the HGF array as a proper subset. In the latter case, the HGF array probes were extracted from the larger dataset and then used for the construction of the graphs.

6.2.2 Construction of the directed metabolic graphs

The creation of the directed metabolic graphs is performed in two stages; firstly, differentially expressed genes are selected from the datasets and associated with the enzymes that they encode. Differential expression is determined through the use of p values (preferably calculated by a non-parametric test such as the Kruskal-Wallis test) and optionally a two-fold change criteria. Data processing associated with the microarrays was performed using R[86] and the Bioconductor package[103]. In the second step, the enzymes associated with the differentially expressed genes are used to add reactions to the directed metabolic graph based on their involvement as a catalyst for the reaction. At the time this work was performed, pathway specific directional information for each reaction was not available in a computer-readable form from the KEGG database, requiring an enhanced version of the database including this information to be created. The labour-intensive nature of this process was the main reason for the use of the HGF array.

Now, the same functionality could be achieved using the XML versions of the pathways, making the choice of array far less significant.

Since multiple probes on the array might be associated with a single enzyme, reactions are normally only added to the graph if all the probes associated with them are significantly altered in expression in the same direction. In the case of particularly noisy datasets meeting this criteria may be a problem, so a “grey area” could be introduced into the construction process. In this case, reactions could be added to the graph as long as at least one of the associated probes is significantly up/downregulated and the others all fall inside the grey area (for example, the threshold for significance could be set at p of 0.05 but the grey area might extend to p of 0.075). Compounds are added to the directed metabolic graph if they are either a substrate or a product for one of the reactions being added. Three directed metabolic graphs are created for each dataset; one containing those reactions associated with upregulated probes, one containing the reactions associated with downregulated probes, and a third that contained both the upregulated and the downregulated reactions. The reactions in the latter graph are colour-coded to indicate if they were more frequently upregulated (green), downregulated (red) or up/downregulated with the same frequency (yellow). A flowchart was created to summarise this process (Figure 6.1).

6.3 Results

In order to validate the construction method, directed metabolic graphs were built for three different types of cancer, pancreatic cancer, colorectal cancer and metastatic breast cancer. Each graph was built using preexisting, publically available data as described previously, and the graphs produced were checked for compounds linked to the cancer types in the literature.

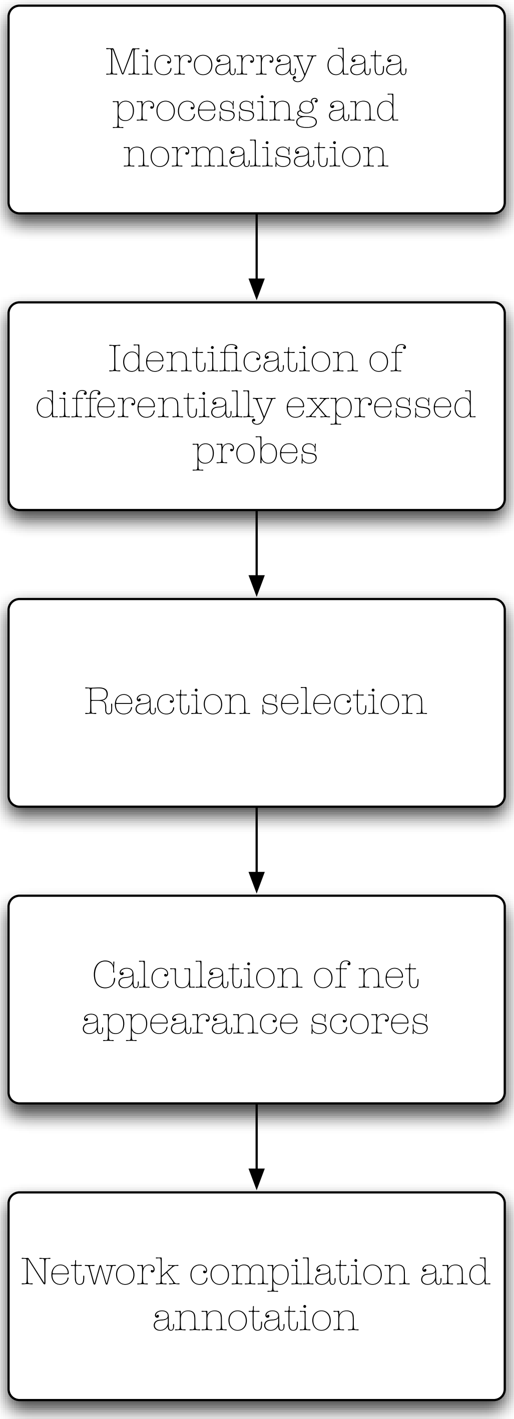

Microarray data processing and normalisation
Identification of differentially expressed probes
Reaction selection
Calculation of net appearance scores
Network compilation and annotation

Figure 6.1: Process used for the generation of the directed metabolic graphs from gene expression data.

Threshold	Mean Node-Node Distance, l	Network Diameter
0.01	4.2748	11
0.02	3.9109	11
0.03	3.5565	8
0.04	3.6836	9
0.05	3.8576	9

Table 6.1: The mean node-node distances for the directed metabolic graphs at each p value along with the graph diameter.

6.3.1 Pancreatic cancer

Pancreatic cancer was the ninth most common malignancy in the United States in 1994[104] and is one of the most dangerous, with nearly 100% of patients developing metastases and dying within 5 years[105]. The most common form of the cancer, Pancreatic Ductal Adenocarcinoma (PDAC), accounts for 90% of all cases[104].

Existing microarray data produced as part of a study into PDAC by Grützmann *et al.*[99] was used to generate directed metabolic graphs for the condition. As the data was gathered using the Affymetrix U133A microarray, only the subset of the probes that corresponded to the HGF array were included in the construction of the graphs. The graphs were produced for a range of p values from 0.01 to 0.05; the graph for p = 0.03 containing reactions associated with both the upregulated and downregulated probes is shown (Figure 6.2).

Looking briefly at the topology of the graph, the mean distance between node pairs (Table 6.1) is comparatively small when compared to the network diameters suggesting that these graphs are small worlds. The degree distribution for the graph shown also displays a power-law with exponent ≈ 2, in line with figures previously reported for metabolic networks[43].

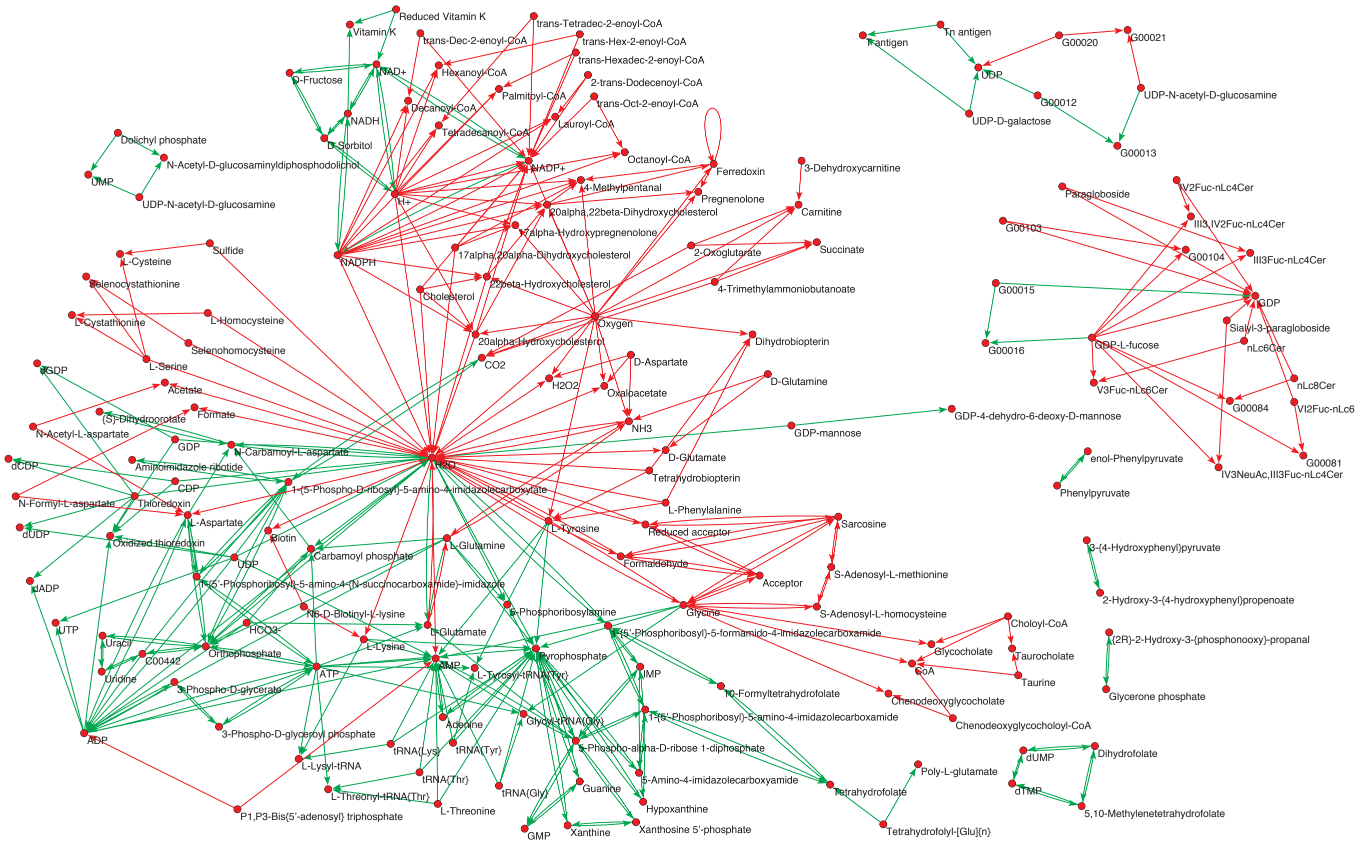


Figure 6.2: A directed metabolic graph produced from microarray data of gene expression in Pancreatic Ductal Adenocarcinoma.

Amongst the important nodes in the system is the H^+ ion, which has a large number of outgoing links all associated with downregulated probes. This connection pattern would lead to an increase in the concentration of H^+ ions in the system and is consistent with the low extracellular pH frequently seen in expanding tumours[105]. The amino acid glycine also features, with a pattern of connections suggestive of a drop in its concentration (although without rate constants in the graph this is a matter of debate); this is consistent with the mutation of the K-ras oncogene that is seen in the vast majority of patients suffering from this condition[106]. While valine, one of the other possible results of the K-ras mutation is present in the graph, the clear suggestion is that its concentration would drop. This may however be due to a lack of coverage of appropriate enzymes (and hence reactions) on the HGF array. Other vertices of high degree include "usual suspects" such as water, and energy-metabolism related compounds such as ATP, AMP, pyrophosphate, orthophosphate, $NADP^+$ and NADPH.

6.3.2 Colorectal cancer

The stage of a malignancy is an important determinant of survival. Amongst patients with colorectal cancer for example, stage 2 tumours have a 20-25% chance of recurrence within 5 years. For stage 3 tumours the rate nearly doubles, rising to 40%[100]. Existing data taken from a study be Groene *et al.*[100] was used to construct directed metabolic graphs showing potential differences in metabolism between stage 2 and stage 3 tumours.

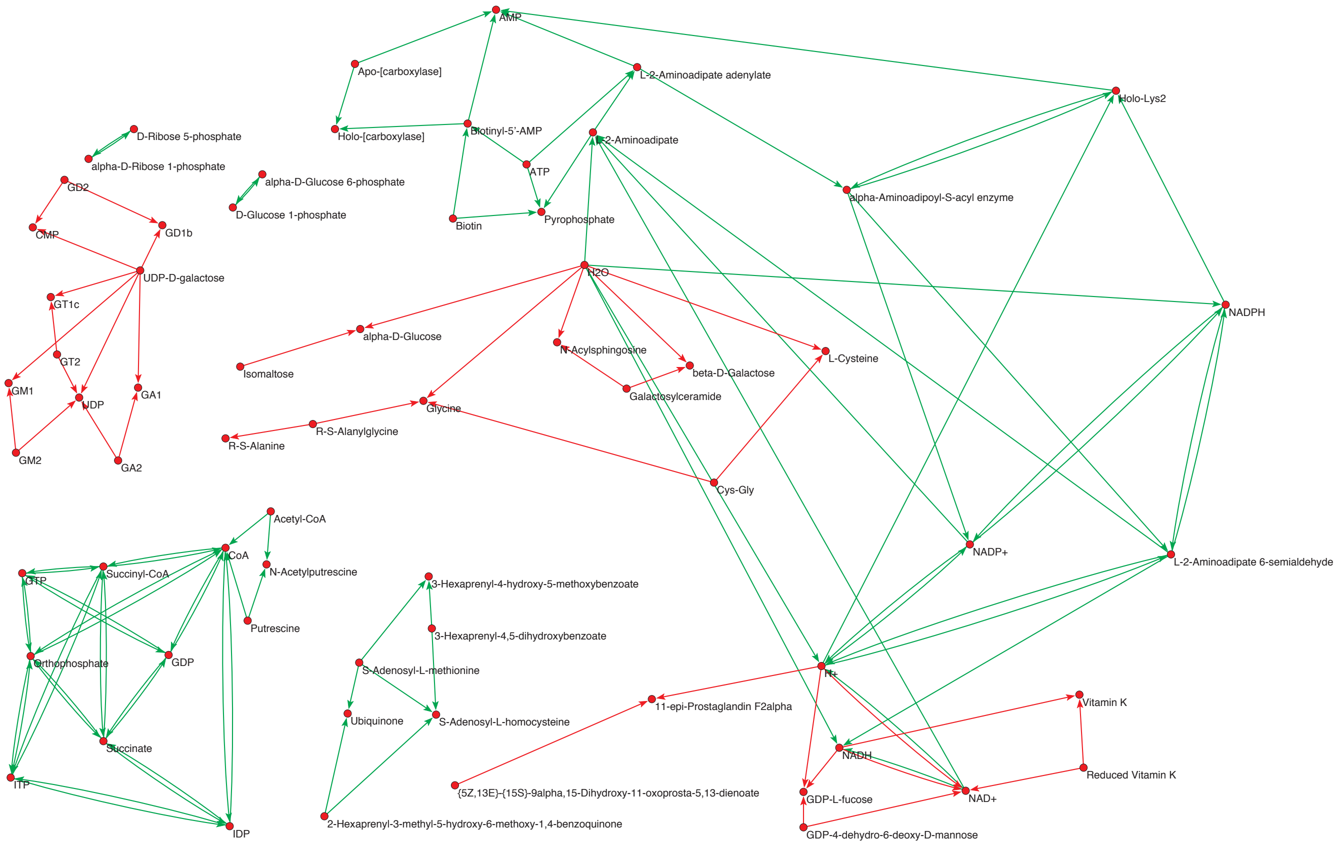


Figure 6.3: A directed metabolic graph produced from gene expression data showing potential differences in metabolism between stage 2 and stage 3 colorectal cancers.

The list of important nodes for the graph at p = 0.06 (Figure 6.3) is dominated by compounds linked to the TCA cycle (CoA, succinate, succinyl-CoA) and to energy production (ATP, AMP, NAD^+, NADH, $NADP^+$, NADPH), with connection patterns indicative of reductions in the concentrations of high-energy species such as ATP and increases in the concentrations of lower-energy species such as AMP. This suggests that the energy requirements of stage 3 colorectal cancers are very different to those of stage 2, most likely due to the generally accepted view that stage 3 tumours are more aggressive than stage 2 tumours[100]. As with the PDAC graphs, H^+ ions are shown to be of increasing concentration. The connection pattern surrounding S-adenosyl-L-methionine indicates a reduction in concentration, and it has been suggested that this could lead to a decrease in DNA methylation and a loss of the normal controls on proto-oncogene expression[107]. This process would result in an increased number of oncogenes being over expressed, which would be consistent with the more aggressive tumour type. The alteration in DNA methylation is a good example of how important biological processes can still be identified in graphs despite the type of data used to generate them; here, we see it in a directed metabolic graph generated from gene expression data, and earlier (in Section 4.3.2) the same process was seen in the metabolic correlation networks derived from dab liver samples as an alteration in choline metabolism.

6.3.3 Metastatic breast cancer

An existing HGF array dataset by Smirnov *et al.*[101] was used to generate a set of directed metabolic graphs showing potential differences in metabolism between circulating endothelial cells (a type of cell that forms the internal surface of blood and lymphatic vessels) taken from healthy volunteers and from patients with metastatic breast cancer.

For the graph at p = 0.05, interesting features include the presence of spermidine and putrescine;

both of these chemicals are known biomarkers found in the plasma and urine of cancer patients and are indicative of tumour growth and cell turnover[108]. Folate, a deficiency of which is increasingly thought to be a factor in the carcinogenesis of many different tumour types including breast cancer[109, 110] also features, along with several chemicals (sarcosine, glycine, and threonine) which are involved in choline metabolism and are immediately downstream of the choline/betaine oxidation known to be enhanced in malignant breast tumours[111]. Finally, hydrogen peroxide, which displays a pattern of connections suggestive of an increasing concentration, has been shown to be produced in human tumour cell lines at a far greater rate than normal[112].

6.4 Discussion

Whilst Linked Metabolites allows researchers to compile directed metabolic graphs showing how metabolites of interest in a particular condition relate to each other, it is difficult to say how those relationships behave dynamically. By building additional graphs from gene expression data, the same information can be viewed from the point of view of the control of the metabolic pathways, indicating which of the enzymes (that catalyse the metabolic reactions) may be altered in concentration and therefore cause an associated change in rates of metabolic reactions. It is important to note that gene expression data is just one more facet of a very complex, integrated system; however, there is no reason that other data types could not be compiled into directed metabolic graphs in similar ways.

A major advantage of this approach is that both the metabolic data, from Linked Metabolites, and the regulatory information, as determined from the gene expression data are in the same form; as a result they could be combined and processed as a single object. The way in which this type of data integration is handled will impact on the types of metrics that could be applied to the data in the future, and therefore it is important to consider these factors. By simply taking the areas of intersection between the graphs, for example, you could greatly reduced the number of chains of reactions that would need to be investigated later by the researcher. Care must be taken however, as different experimental techniques will have very different coverage of the system and it is likely that a simple intersection would therefore result in important information being missed. Instead, it may be more sensible to adopt an approach similar to Arita's, in which those routes suggested by both methods would receive a more favourable weighting in the combined graph. It is also likely that the areas of intersection between the graphs will not correspond to complete chains of reactions but instead effectively form hotspots, where attention should be focused and the most likely routes are likely to fall. Again, this would suggest that an approach that somehow favourably weighted & highlighted the areas of intersection between the graphs within the final, combined graph, would be the best solution to this problem. Unfortunately, a lack of metabolomic and gene expression datasets for the same conditions prevent us from exploring these ideas any further at present.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The visualisation and interpretation of data from metabolomics experiments is a complicated topic. When studies are known to involve a single or a small number of pathways, the colouring of metabolic pathway diagrams based on metabolite concentrations can be a very effective approach; allowing the researcher to study the relationships between compounds by eye, and in the context of the underlying biochemistry. However, such studies are far from being the norm and the low-coverage of the metabolome by many current experimental techniques often mean that the identification of markers by this approach is ineffective. While statistical tools such as PCA and significance tests are a great aid to biomarker identification, reducing the number of metabolites being considered by perhaps an order of magnitude depending on the technology, the compounds they select as most responsible for the separation between sample groups will not necessarily be the most appropriate for use as biomarkers and therefore still require validation with reference to the metabolic pathways.

Metabolic correlation networks present interesting possibilities for both the visualisation of metabolic datasets and the early stages of the identification of biomarkers. By compiling networks for each of the groups of samples in a study, relationships between the various metabolites can be determined and areas of difference or intersection between the networks used to identify potential biomarkers. Correlations might also be used to indicate groups of metabolites that are potentially involved in the same metabolic pathways, and these can then be used as the basis for a search. Care must be taken however, since as with other statistical approaches there is no guarantee that the differences in the networks are specific to the condition being investigated. The issue of "coverage" of the metabolic pathways also poses problems, and can have a dramatic impact on the pattern of connections between a group of metabolites in the network. This difficulty in particular currently makes a more automated approach to biomarker discovery based on metabolic correlation networks and the metabolic pathways impractical; although as experimental techniques improve, and the coverage of each of the metabolic pathways by experimental data becomes more predictable, the situation will change.

In the medium term, metabolic correlation networks may have a much more significant role to play in the analysis of metabolomics datasets, particularly while experimental limitations prevent the use of time-series methods for the determination of biochemical network structure. This may well begin with the extension of the underlying correlation analysis to groups rather than pairs of compounds, creating hyperedges (edges that link more than two compounds together), which would provide additional information and allow the scope of a search of the metabolic pathways to be further refined. It is also possible that correlation networks have a role to play in peak assignment in experimental techniques such as MRS, where issues of resolution mean that it is not always possible to be sure which compound is responsible for a feature in the spectrum. By building correlation networks for that data, and determining which other metabolites form a cluster with the unknown peak, it should be reasonably straightforward to identify it given there would be a limited number of possible options in that area of the spectrum.

As the major contribution of this work to the field, Linked Metabolites attempts to ease the task of researchers looking to validate potential biomarkers identified using multivariate statistical analysis techniques such as PCA, correlation analysis or other similar approaches. Linked Metabolites has two key advantages over other similar tools that are already used for this task; firstly, it offers the possibility of searching for routes between two groups of metabolites. This is an important advance, since as it becomes increasingly apparent that single compounds are unlikely to prove specific enough to act as biomarkers on their own, the idea that a researcher might wish to investigate the links between the compounds forming one “composite” biomarker and another becomes a distinct possibility and one which would not be possible using single-source compound or single-target compound searches. Secondly, Linked Metabolites uses an all simple paths search for the identification of paths between the compound groups. This is an important advantage over the all simple paths algorithm used by similar packages, since although it is slower in many cases, it has the potential to find a far more diverse range of possible solutions as demonstrated by its performance relative to that of KEGG PathComp. The way in which Linked Metabolites combines the metabolic pathways before searching is also unusual, as it collapses all instances of a particular compound into a single node rather than linking the pathways only at specific points. This behaviour has both positive and negative aspects, making “novel” chains of compounds more likely, but increasing the likelihood that errors will be introduced. It is important to stress, that while Linked Metabolites is supposed to ease the task of relating compounds identified in experimental data via the metabolic pathways, the results still need to be checked and interpreted by a domain expert.

The directed metabolic graphs produced by the Linked Metabolites package open up several interesting avenues for the advancement of the work. Most promising is the possibility that through the use of directed metabolic graphs data from other parts of an integrated genomics study could be merged with the results of the metabolomics study. This would allow information on the current state of the cell or tissue, as stored in the metabolome, to be combined with regulatory information, from gene expression, refining the sets of possible reactions further.

The prospect of using metrics from graph theory, such as betweenness centrality, to identify other compounds that are likely to be of altered concentration but not necessarily visible to the experimental technique being used is also important, and could provide an interesting source of compounds whose behaviour could be used for experimental validation of the biomarkers.

The evolution of the Linked Metabolites software itself would be an important next step in the process. The development of a tool, probably web-based, which could perform the entire analysis pipeline is of potentially great use to researchers in the area. As a specific example, it would be advantageous to be able to move from raw, experimental data, through the production of correlation networks, and the running of searches against the metabolic pathways in a single tool. The graphs produced could then be exported in a standard form for use either as the basis for simulation or fed into other software packages for further analysis, for example motif search. A new version of the tool would also require a careful rethink of its expected areas of use; in particular whether a sufficiently large group of users would require the flexibility of switching between organisms, which is potentially available through the use of the KEGG database as opposed to the advantages in terms of cellular compartmentalisation that would be gained by moving to say Reactome. It may be that multiple databases would need to be supported, that way researchers involved in the medical community could perform more restricted searches that included compartmentalisation, while others who were interested in species other than Homo sapiens or who wanted a more “exploratory” answer could use KEGG. With that in mind, further development of the software will require extensive user assessment by the medical community and biologists to determine the best way forward and test the current system on a range of datasets. This work is currently being considered in collaboration with researchers at the Birmingham Children’s Hospital.

APPENDIX A

TIME-SERIES METHODS FOR BIOCHEMICAL NETWORK DISCOVERY

The following appendix summarises a number of different methods for the reconstruction of biochemical reaction networks from time-series concentration data for the participating compounds. While they are currently of limited applicability to metabolomics (as they require the reacting system to be forced away from its steady state and as such can only really be used on data gathered *in vitro*) they are included here as an interesting future alternative to the use of metabolic correlation networks and reference pathways.

Chevalier's method

In 1993, Chevalier *et al.*[113] published a paper that examined several different experimental methods for the discovery of unknown or partially unknown reaction mechanisms and compared their performance based on a model of an oscillating chemical system. Each of the methods represented a different way to gather time-series information on the concentrations of the compounds making up the system but they all shared the same goal; the formation of a Jacobian matrix that described how an alteration in the concentration of one compound would impact on the concentrations of the others, and from which the reaction network could be determined. The three methods investigated were pulse perturbation (the sudden introduction of a given compound into a system in a steady state), concentration shift experiments (in which the concentration of a compound feeding into the system is altered) and delayed feedback experiments (where one of the feeds into the system reflects the concentration of the same compound at an earlier point in time). Of the three methods, only the delayed feedback experiments failed to produce an acceptable approximation of the Jacobian for the system. Concentration shift experiments produced the most accurate version of the Jacobian. Despite producing accurate models of the system under study, Chevalier's methods suffered from several flaws that made them infeasible for large-scale practical use. Firstly and foremost amongst these was the need to perform multiple experiments in order to allow the perturbation of every compound within the system. Perturbations also had to be of known magnitude, making it difficult to perturb the system in ways that were not direct alterations to the concentrations of compounds (such as temperature changes).

The CMC method

In 1995, Arkin and Ross[114] proposed a method for the identification of reaction pathways based on multiple correlation analyses of time-series data. The Correlation Metric Construction (CMC) method differed from Chevalier's in many ways, but the most significant alteration was in the way in which the system was forced from the steady state. In Chevalier's method this was either done through the use of a pulse perturbation or a series of sustained concentration shifts, whereas in CMC those compounds that served as the inputs to the reaction pathway were continuously driven to random concentrations throughout the recording process. The CMC method consists of seven steps:

1. A set of measurements of the concentrations of all the compounds in the system are taken over an extended period, the sampling interval is set to be approximately the time that it takes for the system to settle back into its steady state and the inputs of the system are forced to random (Gaussian distributed) concentrations each time a sample is made. The total experimental time should be long enough to ensure that all possible combinations of the input compounds have been sampled.

2. The correlation matrix, $R(\tau) = (r_{ij}(\tau))$ is calculated from the time-series using the equations:

$$S_{ij}(\tau) = \langle (x_i(t) - \bar{x}_i)(x_j(t+\tau) - \bar{x}_j) \rangle \tag{A.1}$$

$$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}(\tau) S_{jj}(\tau)}} \tag{A.2}$$

   Where angle brackets denote an average, $x_i(t)$ is the tth timepoint of the time-series for compound i and $\bar{x}_i$ is the average of the ith time-series.

3. A clustering algorithm is used to group the compounds in the system based on maximum correlations between all species, this produces a representation of the system in which every compound is connected to at least one other.

4. The correlation matrix, $R(\tau)$ is converted into a Euclidean distance matrix, $D = (d_{ij})$ using the transform:

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}} \tag{A.3}$$

$$c_{ij} = \max |r_{ij}(\tau)|_\tau \tag{A.4}$$

For those pairs of variables which had a low correlation value the corresponding Euclidean distance is large and *vice-versa*.

5. The classical multidimensional scaling method is applied to the distance matrix producing a consistent configuration of the points, while at the same time establishing the dimensionality of the data. Once projected in 2-D the points will be separated by a distance which is less than or equal to the actual distance in the distance matrix.

6. An optimisation-based multidimensional scaling algorithm is applied to the dataset which produces an alternative representation in which the distances between the points may be less than, equal to or greater than the actual distance values. The optimisation processes is based on the minimisation of a stress function using a simulated annealing algorithm. The stress value for both representations is then calculated to determine which is the better solution.

7. Finally, a cluster analysis is performed on the distance matrix to establish a hierarchy amongst the interactions of each of the sub-systems within the pathway.

Despite the advantages of Arkin's method, which not only establishes the structure of the reaction pathway but also highlights areas of tight regulation (the compounds involved lying very close to each other in the 2-D projection) difficulties do exist; these include the need to know the inputs to the system under study *a priori*, the need to continuously alter the concentration of those compounds throughout the acquisition of the time-series, the selection of the number of datapoints required for analysis and the problem of assessing the significance of the correlations calculated.

Díaz-Sierra's method

In 1999, Díaz-Sierra *et al.*[115] published an improved version of Chevalier's pulse perturbation method which when applied to well-defined data led to not only the identification of the chemical processes at work, but also the stoichiometric coefficients and rate constants for each of the reactions.

The method Chevalier had used to obtain an estimate for the Jacobian was sometimes numerically problematic; an alternative that had since been proposed by Sorribas *et al.*[116] used an intermediate matrix Φ as the basis for the calculation of the eigenvalues of the Jacobian. Φ was calculated from the equation:

$$\delta X(t_i + h) = \Phi \cdot \delta X(t_i) \tag{A.5}$$

Unfortunately, the calculation of the eigenvalues of Φ was sensitive to noise and rounding errors in the experimental data. To avoid this problem, Díaz-Sierra adapted Sorribas' method by using Φ as a term in:

$$J = \frac{1}{h}[I + (\Phi - I)] \tag{A.6}$$

When expanded in its Taylor series, the above could then be used to calculate the Jacobian as long as the sampling period, h was constant and long enough to capture the greatest relaxation time for the system.

$$J = \frac{1}{h}[(\Phi - I) - \frac{(\Phi - I)(\Phi - I)}{2} + \ldots] \tag{A.7}$$

The reaction mechanisms are determined from the Jacobian as follows:

1. The matrix of coefficients, $J_{ij}X_{j,0}$ is created by the element-wise multiplication of each column of the Jacobian by its equivalent steady state concentration. The number of generic reaction steps can then be determined by counting the number of independent entries. For each of the remaining entries, linear relationships should be expressed in the form:

$$\sum_{ij}^{n} I_{ij}^{(r)}(J_{ij}X_{j,0}) = 0 \qquad r = 1, \ldots, p \tag{A.8}$$

where $I_{ij}^{(r)}$ are low integers and p is the number of dependent entries.

2. The linear equations are then substituted into:

$$\sum_{s=1}^{m} \alpha_{js}\gamma_{is}\nu_{s,0} = J_{ij}X_{j,0} \qquad i, j = 1, \ldots, n \tag{A.9}$$

to obtain equations in the form:

$$\sum_{ij}^{n} I_{ij}^{(r)}(\alpha_{js}\gamma_{is}) = 0 \qquad r = 1, \ldots, p, \qquad s = 1, \ldots, m \tag{A.10}$$

3. While these equations can not be used to directly infer the values of α and γ, they do imply that α and γ can be expressed as parameter-dependent functions of the form:

$$\alpha_{is}(\mu_1, \ldots, \mu_R), \qquad \gamma_{is}(\mu_1, \ldots, \mu_R) \tag{A.11}$$

The values of μ_R can then be determined by the substitution of the above into the first set of equations in step 2, followed by the use of the known rates at which compounds are being fed into the system.

Schmidt's method

Recently, Schmidt *et al.*[117] have produced a method specifically tailored towards the identification of small scale biochemical networks based on time-series in which both the magnitude and effects of the perturbation are unknown. Their method has several important characteristics in that it not only produces an estimate for the continuous time Jacobian of the system, but also produces estimates for the perturbation to individual system components and can indicate elements of the system possessing faster dynamics than can be captured by the experimental sampling time; however to date it has only been applied to *in silico* systems.

The method assumes a fixed sampling time ΔT and that the perturbation to the system is constant between two sampling points. It also assumes that in a system consisting of n components, the concentration of every component is measured at each timepoint and that a minimum of $n+2$ timepoints are recorded with the first being taken before the perturbation is applied.[*]

The estimates for A_d, the discrete time Jacobian of the system and Δu, the constant unknown perturbation to the system are obtained simultaneously from Δx_k, the measurement vectors using the equation:

$$[\hat{A}_d, \Delta\hat{u}] = RM^T(MM^T)^{-1} \tag{A.12}$$

Where R, the result matrix assuming m relative timepoints is given by:

$$R = [\ \Delta x_m \quad \Delta x_{m-1} \quad \dots \quad \Delta x_2\ ] \tag{A.13}$$

[*] This is the absolute minimum and in real-world systems more timepoints are needed and multiple experiments should be performed from the same steady state. The first recorded timepoint is only used as a reference from which to calculate the relative values, so $n+2$ recorded timepoints will result in $n+1$ relative timepoints in the calculations.

and M, the measurement matrix is given by:

$$M = [\ \Delta x_{m-1} \quad \Delta x_{m-2} \quad \dots \quad \Delta x_1\ ] \tag{A.14}$$

Since all of the values are relative, a system from which you have recorded information at six timepoints will therefore result in (4)×(4) result and measurement matrices.

When r experiments are performed the result and measurement matrices are constructed as follows:

$$R = [\ R_1 \quad \dots \quad R_r\ ] \qquad\qquad M = \begin{bmatrix} M_1 & M_2 & \dots & M_r \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{A.15}$$

and the estimate becomes:

$$[\hat{A}_d, \Delta\hat{u}_1, \dots, \Delta\hat{u}_r] = RM^T(MM^T)^{-1} \tag{A.16}$$

where $\Delta\hat{u}_r$ is the estimate of the perturbation in experiment r.

$\hat{A}_{zoh}$, the estimate of the continuous time Jacobian is then calculated from the estimated discrete time Jacobian using the inverse form of the zero-order hold discretisation:

$$\hat{A}_{zoh} = \frac{1}{\Delta T} \log_m \left(\hat{A}_d\right) \tag{A.17}$$

where $\log_m \left(\hat{A}_d\right)$ denotes the matrix logarithm of the estimation of the discrete time Jacobian.

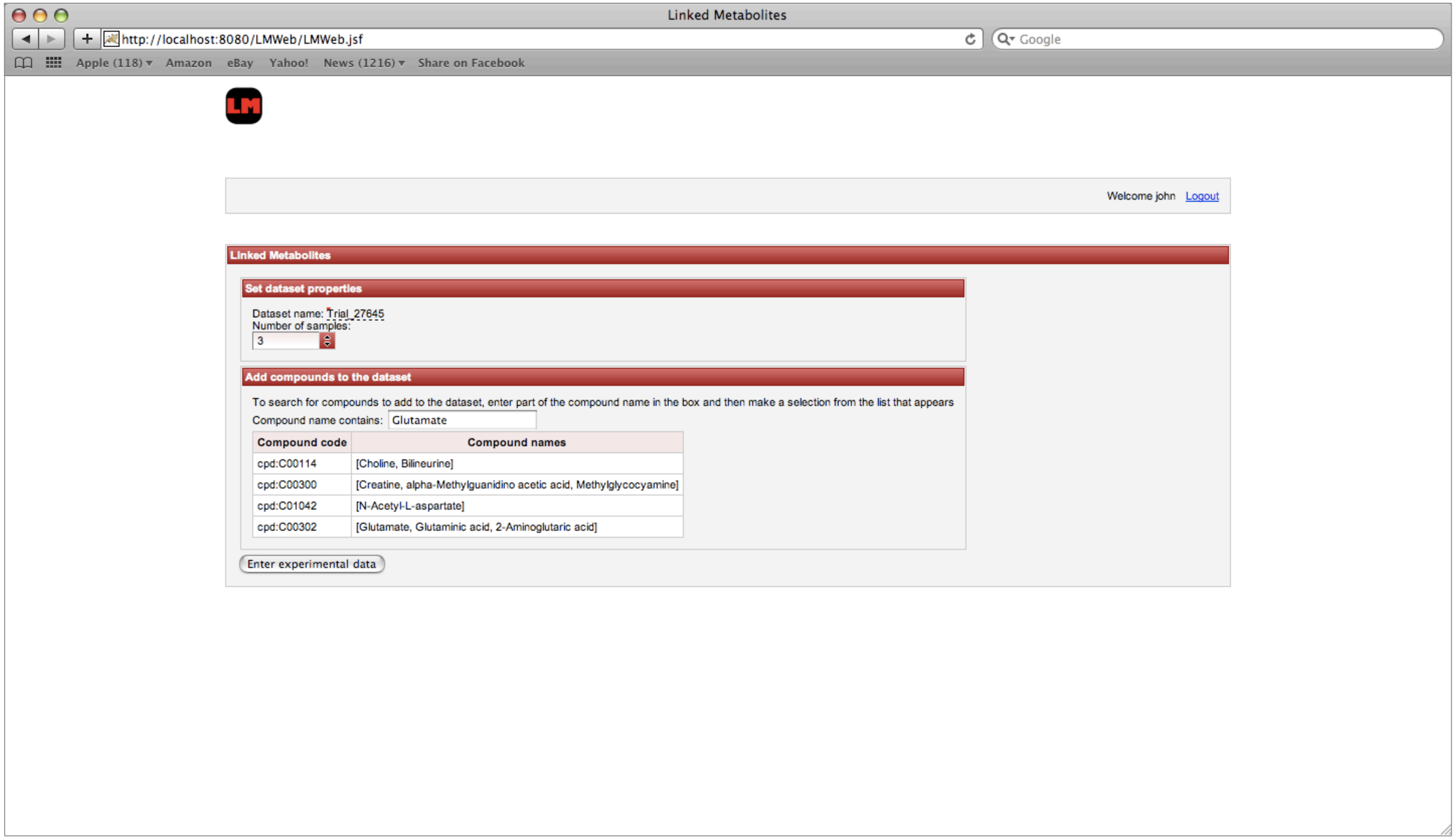
The identification of system components with dynamics that are faster than can be captured by the chosen experimental sampling time is an important issue since if not removed from the analysis they lead to inaccurate representations of the system. The identification process is relatively straightforward, first the singular values of the measurement matrix, M are calculated. If θ_1 is relatively close to zero compared to the other singular values of the system then the chosen sampling time is too short relative to the system's dynamics. Next, the singular vector which corresponds to the singular value identified must be examined, the most dominant element of which indicates the component with the fast dynamic. This component is then removed from the system and the network for the remaining elements calculated as normal.

The extraction of the network underlying the system is now performed in the same way as for the other methods.

APPENDIX B

PROTOTYPE SCREENS FOR THE WEB VERSION OF LINKED METABOLITES

The following appendix presents a number of prototype screens for a web-based version of Linked Metabolites. The new version could expand on the functionality of its predecessor by allowing the entry of experimental datasets and the construction of metabolic correlation networks, thus guiding the users selection of search groups.



Linked Metabolites

Set dataset properties

Dataset name: Trial_27645
Number of samples:
3

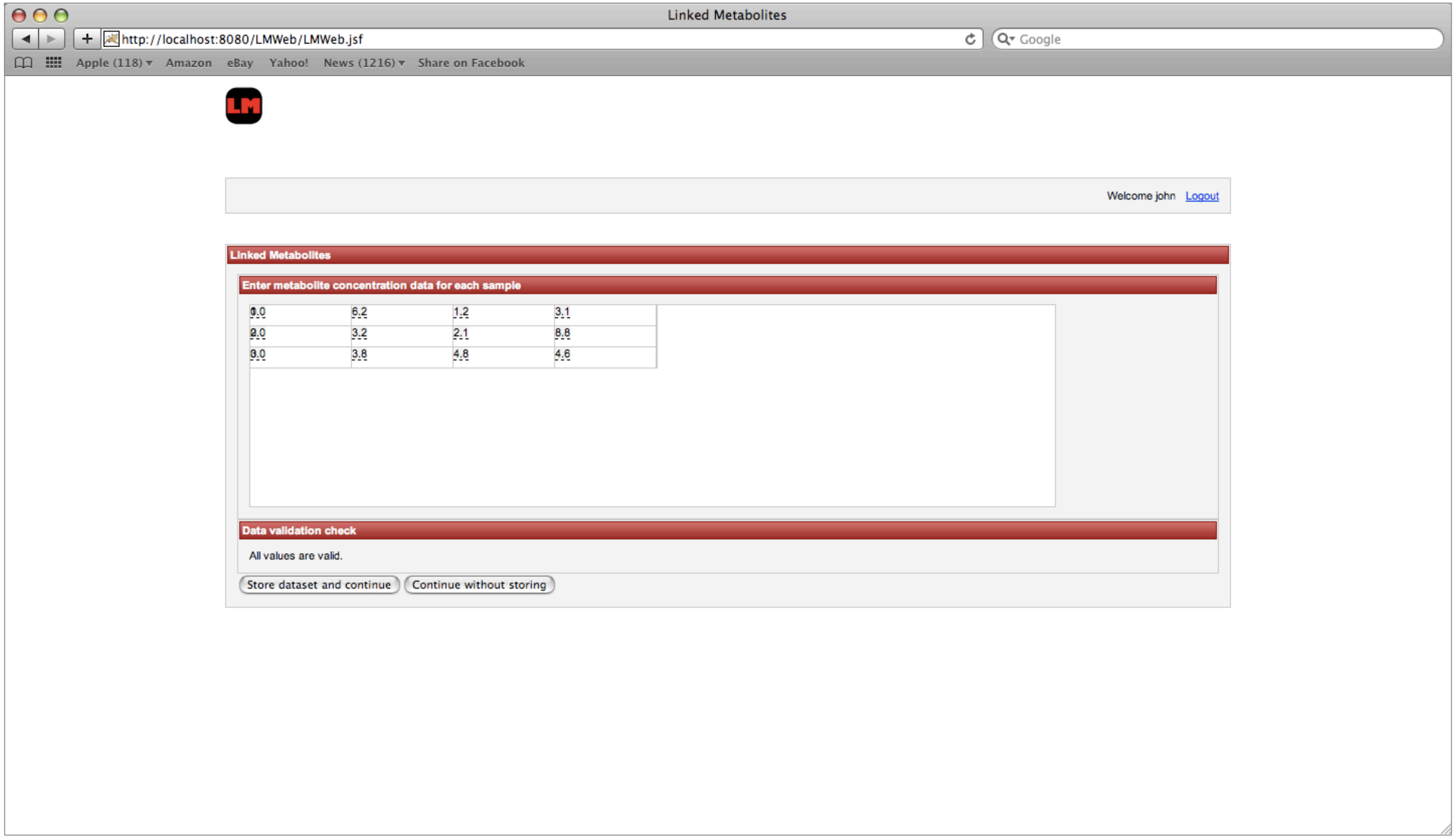
Add compounds to the dataset

To search for compounds to add to the dataset, enter part of the compound name in the box and then make a selection from the list that appears

Compound name contains: Glutamate

Compound code	Compound names
cpd:C00114	[Choline, Bilineurine]
cpd:C00300	[Creatine, alpha-Methylguanidino acetic acid, Methylglycocyamine]
cpd:C01042	[N-Acetyl-L-aspartate]
cpd:C00302	[Glutamate, Glutaminic acid, 2-Aminoglutaric acid]

Figure B.1: The data entry screen (part 1). Here the user defines the dataset by specifying a name, the number of samples and the metabolites measured.



Linked Metabolites

Enter metabolite concentration data for each sample

1.0	6.2	1.2	3.1
2.0	3.2	2.1	8.8
3.0	3.8	4.8	4.6

Data validation check

All values are valid.

Figure B.2: The data entry screen (part 2). Here the user enters the concentrations for each metabolite.

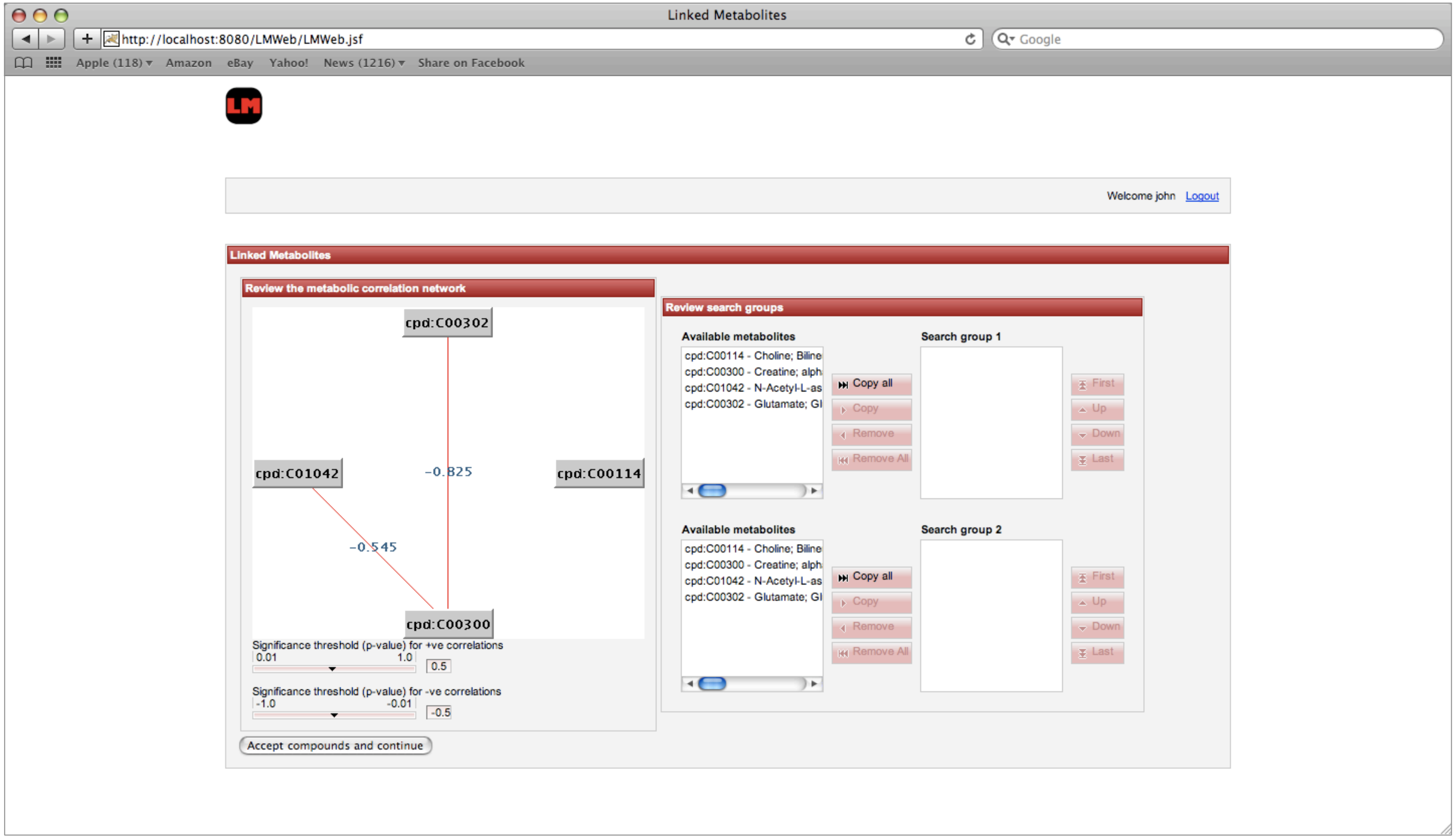
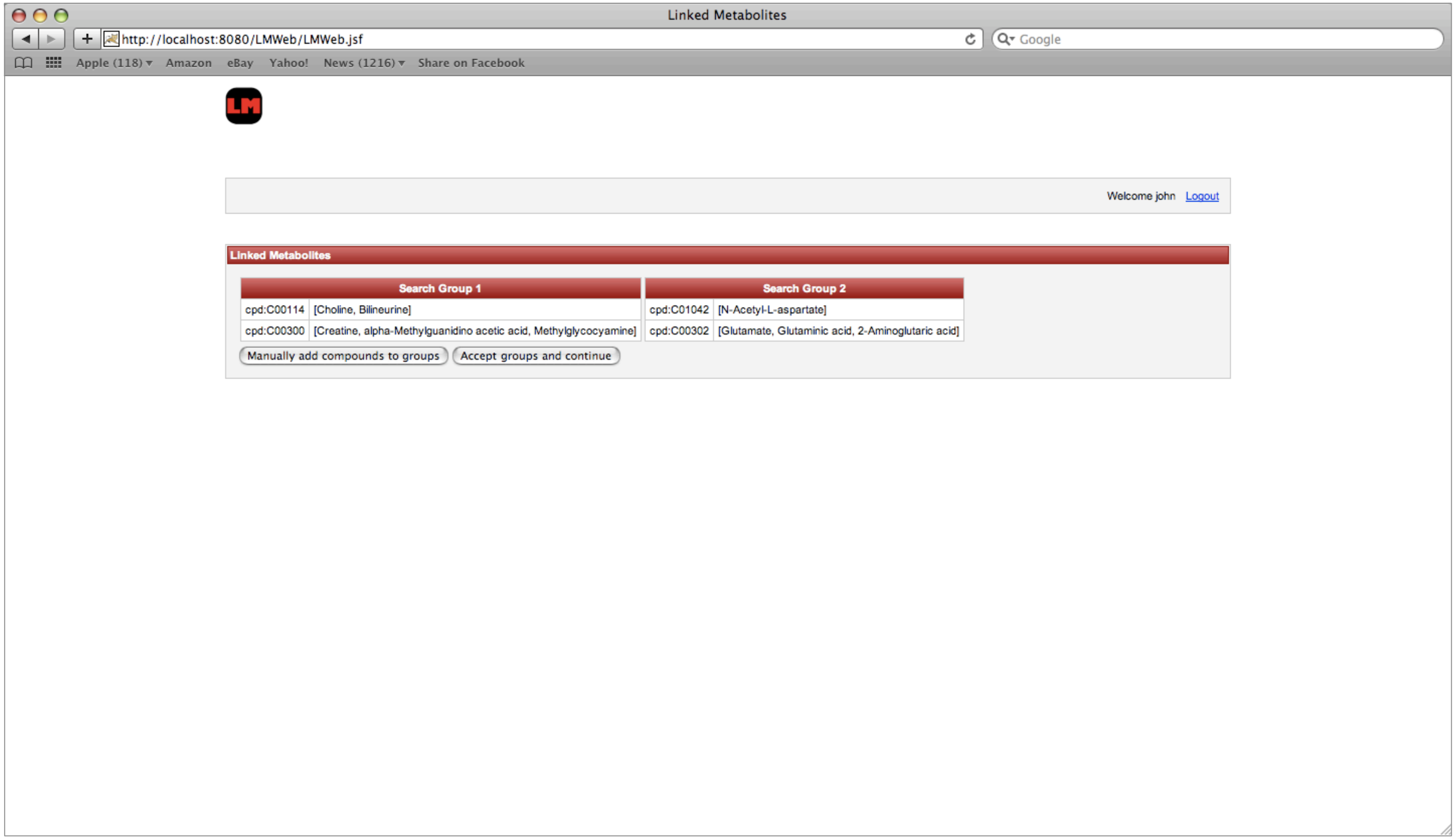

Linked Metabolites
http://localhost:8080/LMWeb/LMWeb.jsf
Google
Apple (118) ▾ Amazon eBay Yahoo! News (1216) ▾ Share on Facebook
LM
Welcome john Logout
Linked Metabolites
Review the metabolic correlation network
cpd:C00302
cpd:C01042
cpd:C00114
cpd:C00300
−0.825
−0.545
Significance threshold (p-value) for +ve correlations
0.01
1.0
0.5
Significance threshold (p-value) for -ve correlations
-1.0
-0.01
-0.5
Accept compounds and continue
Review search groups
Available metabolites
cpd:C00114 - Choline; Biline
cpd:C00300 - Creatine; alph
cpd:C01042 - N-Acetyl-L-as
cpd:C00302 - Glutamate; Gl
Copy all
Copy
Remove
Remove All
Search group 1
First
Up
Down
Last
Available metabolites
cpd:C00114 - Choline; Biline
cpd:C00300 - Creatine; alph
cpd:C01042 - N-Acetyl-L-as
cpd:C00302 - Glutamate; Gl
Copy all
Copy
Remove
Remove All
Search group 2
First
Up
Down
Last

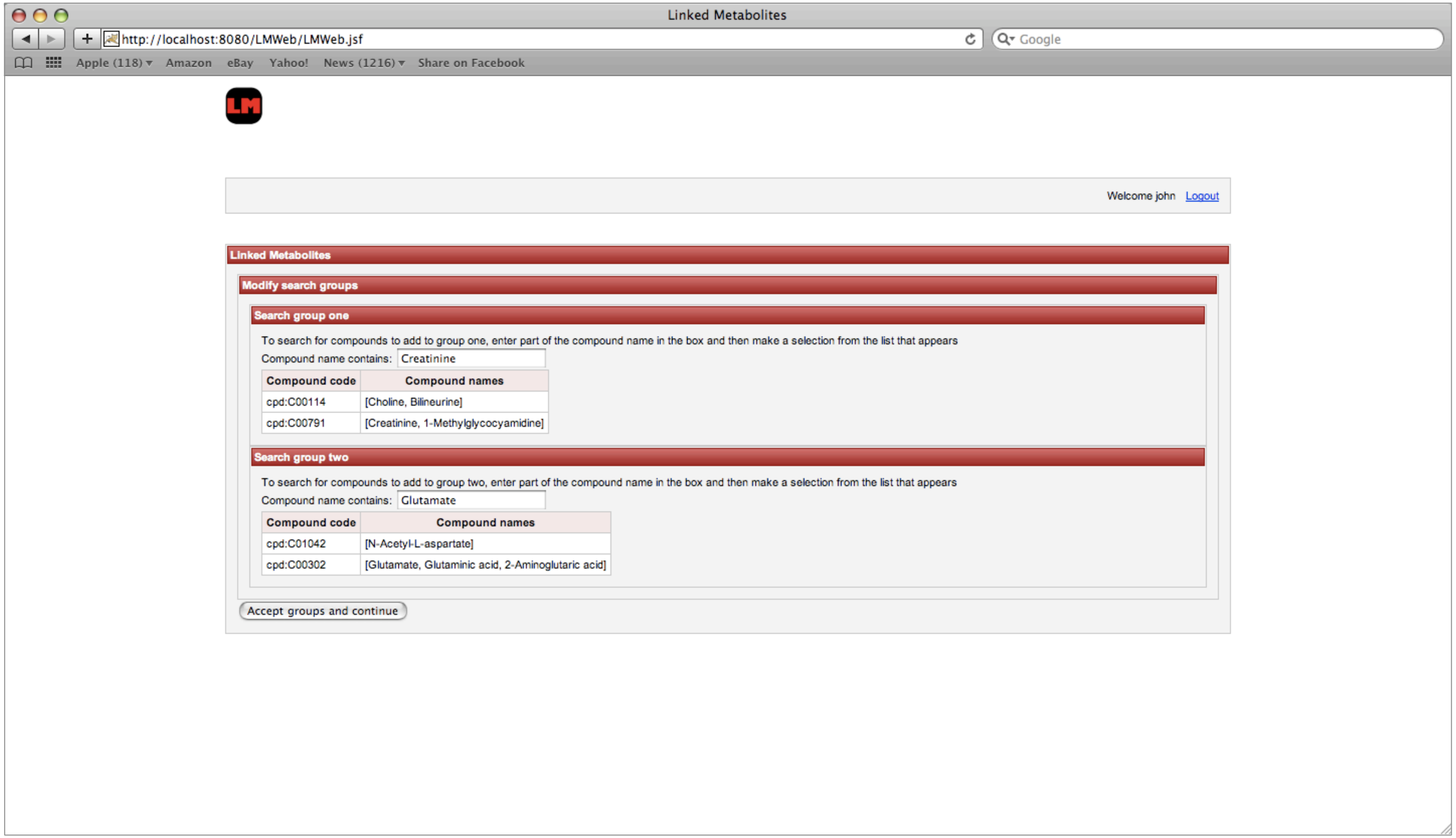
Figure B.3: The network review screen. Here the user can view the correlation network for their dataset. The threshold values can be selected using the sliders. On the right hand side of the screen the user can edit the search groups while still viewing the correlation network.



Linked Metabolites

Search Group 1		Search Group 2	
cpd:C00114	[Choline, Bilineurine]	cpd:C01042	[N-Acetyl-L-aspartate]
cpd:C00300	[Creatine, alpha-Methylguanidino acetic acid, Methylglycocyamine]	cpd:C00302	[Glutamate, Glutaminic acid, 2-Aminoglutaric acid]

Figure B.4: The network group selection screen. Here the user checks the groups they have set. They may then either confirm that they want to proceed with the search, or choose to manually add additional metabolites.



Linked Metabolites

Modify search groups

Search group one

To search for compounds to add to group one, enter part of the compound name in the box and then make a selection from the list that appears

Compound name contains: Creatinine

Compound code	Compound names
cpd:C00114	[Choline, Bilineurine]
cpd:C00791	[Creatinine, 1-Methylglycocyamidine]

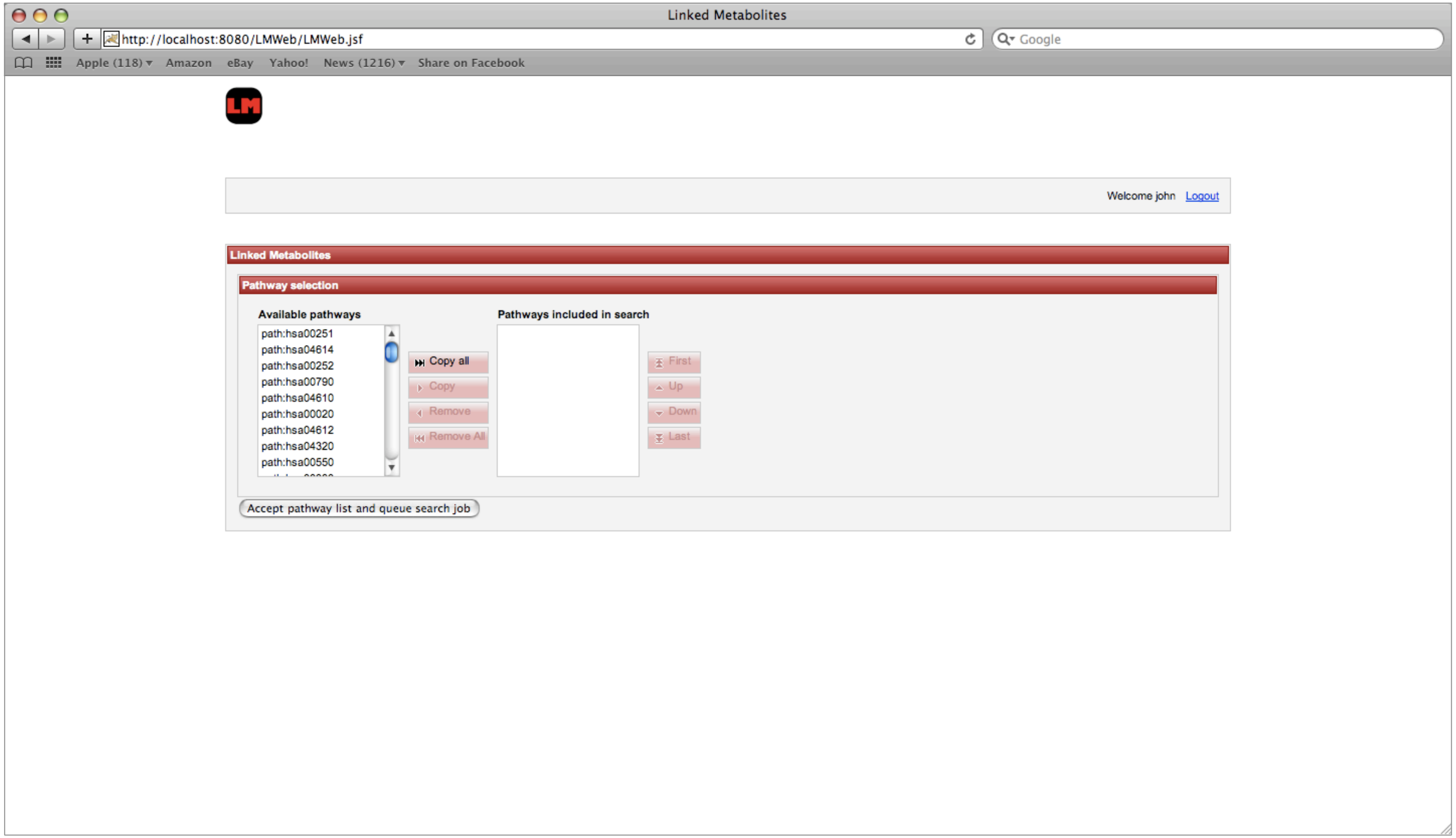
Search group two

To search for compounds to add to group two, enter part of the compound name in the box and then make a selection from the list that appears

Compound name contains: Glutamate

Compound code	Compound names
cpd:C01042	[N-Acetyl-L-aspartate]
cpd:C00302	[Glutamate, Glutaminic acid, 2-Aminoglutaric acid]

Figure B.5: The group selection screen. Here the user may edit the search groups independently of the correlation analysis.



Available pathways

path:hsa00251
path:hsa04614
path:hsa00252
path:hsa00790
path:hsa04610
path:hsa00020
path:hsa04612
path:hsa04320
path:hsa00550

Pathways included in search

Figure B.6: The pathway selection screen. Here the user selects the pathways that will be included in the search.

GLOSSARY

-

-omics An area of study in biology focusing on a particular set of compounds such as genomics (the study of the genome) or metabolomics (the study of the metabolome).

A

apoptosis The process of cell death.

B

betweenness centrality The proportion of the paths through a graph that feature a given node or edge. High betweenness centrality values may indicate a bottleneck.

biofluid A fluid such as blood or urine that can be taken from a biological specimen.

biomarker A compound or group of compounds that in particular concentrations can be used to determine an organism's condition (e.g. determine whether it has a particular disease).

BioPAX A data exchange format for metabolic pathway information.

Bonferroni correction A method of correcting for multiple significance tests by dividing the significance value by the number of tests to be performed against it.

C

clustering coefficient The extent to which the neighbours of a node are also connected to each other.

correlation coefficient A measure of how closely two vectors are related. In biology this is usually the concentrations of two compounds across a set of samples.

D

degree The total number of edges or arcs incident on a node in a graph.

degree distribution The range of probabilities showing how likely it is a node in the graph will be connected to a given number of edges.

directed metabolic graph A graph showing metabolites and the directed flow of reactions between them.

F

false discovery rate A method of selecting the significance threshold based on an expected frequency of false-positive results.

G

glycolysis A metabolic pathway in which sugars are broken down to form energy.

graph theory A branch of mathematics that studies the properties of graphs.

I

in vitro Within the glass. *In vitro* experiments are performed on biological samples that are outside an animal or plant.

in vivo Within an organism. *In vivo* imaging techniques include x-rays, where bones are examined without removing them from the individual.

indegree The number of arcs heading into a node in a graph.

K

KEGG See Kyoto Encyclopedia of Genes and Genomes.

Kruskal-Wallis A non-parametric, variance-based test for statistical significance. Since it is non-parametric, the Kruskal-Wallis test has no underlying assumption that the data follows a normal distribution.

Kyoto Encyclopedia of Genes and Genomes A biological database that contains, amongst other things, a set of metabolic pathway diagrams.

L

Linked Metabolites A graphical tool for route search within the metabolic pathways and the construction of directed metabolic graphs.

M

Magnetic Resonance Imaging An *in vivo* branch of NMR that uses the water within tissues to create images of the body.

Magnetic Resonance Spectroscopy An *in vivo* branch of NMR that focuses on the detection of metabolites.

metabolic correlation network A graph in which nodes, representing metabolites, are linked by edges that represent a significant correlation between the concentrations of the two in an experimental dataset.

metabolic pathway A set of reactions that perform a particular metabolic function such as the release of energy from sugars.

metabolome The complete set of low molecular weight compounds involved in metabolism.

metabolomics The study of the set of low molecular weight compounds involved in metabolism.

microarray A piece of apparatus that can be used to measure gene expression.

motif search The extension of triadic census to n-node subgraphs.

MRI See Magnetic Resonance Imaging.

MRS See Magnetic Resonance Spectroscopy.

N

necrosis The premature death of cells and tissue.

NMR See Nuclear Magnetic Resonance.

Nuclear Magnetic Resonance A way of determining the molecules that are present in a sample based on the magnetic properties of some atomic nuclei (e.g. hydrogen).

O

outdegree The number of arcs leaving a node in a graph.

P

Pajek A program for the analysis of networks.

PCA See Principal Components Analysis.

Principal Components Analysis A statistical technique that describes a dataset with a large number of dimensions in terms of a smaller number of vectors that represent the variance in the original data.

Python A scripting language.

R

Reactome A metabolic pathway database.

S

SBML See the Systems Biology Markup Language.

scale-free network A graph with a degree distribution that follows a power-law. Amongst other interesting features, scale-free systems display a higher than normal resilience to random node failures.

simple path A path between two nodes that does not loop back on itself at any point.

small-world network A graph with a small average path length between nodes compared to that of a random graph of similar size and density.

Systems Biology Markup Language An xml standard for the exchange of biological models.

T

TCA cycle A metabolic pathway, also known as the citric acid cycle or the Krebs cycle, that is key to energy metabolism and electron transport.

triadic census A survey of a graph for the frequencies of the sixteen possible non-identical, three-node subgraphs and the comparison of those frequencies against a set of expected values.

BIBLIOGRAPHY

[1] D. Robertson, “Metabonomics in toxicology: a review,” *Toxicological Sciences*, vol. 85, no. 2, pp. 809–822, 2005.

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular biology of the cell.* Garland Publishing, Inc., 3rd ed., 1994.

[3] V. Hatzimanikatis, C. Li, J. Ionita, C. Henry, M. Jankowski, and L. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, 2005.

[4] O. Fiehn, “Metabolomics - the link between genotypes and phenotypes,” *Plant Molecular Biology*, vol. 48, pp. 155–171, 2002.

[5] P. Mendes, “Emerging bioinformatics for the metabolome,” *Briefings in Bioinformatics*, vol. 3, no. 2, pp. 134–145, 2002.

[6] S. Oliver, “Functional genomics: lessons from yeast,” *Philosophical Transactions of the Royal Society B*, vol. 357, pp. 17–23, 2002.

[7] J. Nicholson, J. Lindon, and E. Holmes, “‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data,” *Xenobiotica*, vol. 29, no. 11, pp. 1181–1189, 1999.

[8] J. Griffin and J. Shockcor, “Metabolic profiles of cancer cells,” *Nature Reviews Cancer*, vol. 4, pp. 551–561, 2004.

[9] R. Goodacre, S. Vaidyanathan, W. Dunn, G. Harrigan, and D. Kell, “Metabolomics by numbers: acquiring and understanding global metabolite data,” *Trends in Biotechnology*, vol. 22, no. 5, pp. 245–252, 2004.

[10] M. Viant, “Metabolomics of aquatic organisms: the new ‘omics’ on the block,” *Marine Ecology Progress Series*, vol. 332, pp. 301–306, 2007.

[11] M. Viant, E. Rosenblum, and R. Tjeerdema, "NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health," *Environmental Science and Technology*, vol. 37, no. 21, pp. 4982–4989, 2003.

[12] J. Bundy, D. Spurgeon, C. Svendsen, P. Hankard, J. Weeks, D. Osborn, J. Lindon, and J. Nicholson, "Environmental metabonomics: applying combination biomarker analysis in earthworms at a metal contaminated site," *Ecotoxicology*, vol. 13, pp. 797–806, 2004.

[13] C. Lin, M. Viant, and R. Tjeerdema, "Metabolomics: methodologies and applications in the environmental sciences," *Journal of Pesticide Science*, vol. 31, no. 3, pp. 245–251, 2006.

[14] V. Go, R. Butrum, and D. Wong, "Diet, nutrition, and cancer prevention: the postgenomic era," *The Journal of Nutrition*, pp. 3830S–3836S, 2003.

[15] J. Nicholson, J. Connelly, J. Lindon, and E. Holmes, "Metabonomics: a platform for studying drug toxicity and gene function," *Nature Reviews Drug Discovery*, vol. 1, pp. 153–162, 2002.

[16] British Broadcasting Corporation News, "Six taken ill after drug trials." Webpage: http://news.bbc.co.uk/1/hi/england/london/4807042.stm. Last accessed 04/05/2009.

[17] J. Brindle, H. Antti, E. Holmes, G. Tranter, J. Nicholson, H. Bethell, S. Clarke, P. Schofield, E. McKilligin, D. Mosedale, and D. Grainger, "Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics," *Nature Medicine*, vol. 8, no. 12, pp. 1439–1444, 2002.

[18] L. Kenny, W. Dunn, D. Ellis, J. Myers, P. Baker, the GOPEC Consortium, and D. Kell, "Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning," *Metabolomics*, vol. 1, no. 3, pp. 227–234, 2005.

[19] M. Coen, M. O'Sullivan, W. Bubb, P. Kuchel, and T. Sorrell, "Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis," *Clinical Infectious Diseases*, vol. 41, pp. 1582–1590, 2005.

[20] S. Carraro, S. Rezzi, F. Reniero, K. Héberger, G. Giordano, S. Zanconato, C. Guillou, and E. Baraldi, "Metabolomics applied to exhaled breath condensate in childhood asthma," *American Journal of Respiratory and Critical Care Medicine*, vol. 175, pp. 986–990, 2007.

[21] A. Tate, J. Griffiths, I. Martinez-Pérez, A. Moreno, I. Barba, M. Cabañas, D. Watson, J. Alonso, F. Bartumeus, F. Isamat, I. Ferrer, F. Vila, E. Ferrer, A. Capdevila, and C. Arús, "Towards a method for automated classification of ^{1}H MRS spectra from brain tumours," *NMR in Biomedicine*, vol. 11, pp. 177–191, 1998.

[22] R. Packer, "Brain tumours in children," *Archives of Neurology*, vol. 56, pp. 421–425, 1999.

[23] F. Howe and K. Opstad, "^{1}H MR spectroscopy of brain tumours and masses," *NMR in Biomedicine*, vol. 16, pp. 123–131, 2003.

[24] M. Preul, Z. Caramanos, D. Collins, J. Villemure, R. Leblanc, A. Oliver, R. Pokrupa, and D. Arnold, "Accurate, non-invasive diagnosis of human brain tumours by using proton magnetic resonance spectroscopy," *Nature Medicine*, vol. 2, pp. 323–325, 1996.

[25] L. Harris, N. Davies, L. MacPherson, K. Foster, S. Lateef, K. Natarajan, S. Sgouros, M. Brundler, T. Arvanitis, R. Grundy, and A. Peet, "The use of short-echo-time ^{1}H MRS for childhood cerebellar tumours prior to histopathological diagnosis," *Pediatric Radiology*, vol. 37, no. 11, pp. 1101–1109, 2007.

[26] A. Llombart-Bosch, J. López-Guerrero, and V. Felipo, eds., *New trends in cancer for the 21st century*, ch. MRS as endogenous molecular imaging for brain and prostate tumors: FP6 project "eTUMOR", pp. 285–302. Springer-Verlag, 2006.

[27] H. González-Vélez, M. Mier, M. Julià-Sapé, T. Arvanitis, J. García-Gómez, M. Robles, P. Lewis, S. Dasmahapatra, D. Dupplaw, A. Peet, C. Arús, B. Celda, S. V. Huffel, and M. Lluch-Ariet, "Healthagents: distributed multi-agent brain tumor diagnosis and prognosis," *Applied Intelligence*, 2007.

[28] K. Warren, "NMR spectroscopy and pediatric brain tumors," *The Oncologist*, vol. 9, pp. 312–318, 2004.

[29] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, *Multi and megavariate data analysis : principles & applications.* Umetrics Academy, 2001.

[30] J. Gross and J. Yellen, *Graph theory and its applications.* CRC Press, 1999.

[31] M. Newman, A.-L. Barabási, and D. Watts, eds., *The structure and dynamics of networks.* Princeton University Press, 2006.

[32] I. Pool and M. Kochen, "Contacts and influence," *Social Networks*, vol. 1, no. 1, pp. 5–51, 1978.

[33] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.

[34] A.-L. Barabási, *Linked.* Plume, 2003.

[35] M. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[36] B. Tjaden, "The oracle of bacon." Homepage: http://www.oracleofbacon.org. Last accessed 04/05/2009.

[37] "IMDB: The internet movie database." Homepage: http://www.imdb.com. Last accessed 04/05/2009.

[38] C. Goffman, "And what is your Erdös number?," *The American Mathematical Monthly*, vol. 76, no. 7, p. 791, 1969.

[39] J. Grossman and P. Ion, "The Erdös number project." Homepage: http://www.oakland.edu/enp. Last accessed 04/05/2009.

[40] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[41] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, pp. 130–131, 1999.

[42] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, 2000.

[43] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, pp. 651–654, 2000.

[44] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A*, vol. 311, pp. 590–614, 2002.

[45] F. Liljeros, C. Edling, L. N. Amaral, H. E. Stanley, and Y. Åberg, "The web of human sexual contacts," *Nature*, vol. 411, pp. 907–908, 2001.

[46] V. Krebs, "Uncloaking terrorist networks," *First Monday*, vol. 7, April 2002.

[47] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[48] S. Bornholdt and H. Schuster, eds., *Handbook of graphs and networks.* WILEY-VCH GmbH & Co., 2003.

[49] B. Bollobás and O. Riordan, "The diameter of a scale-free random graph," *Combinatorica*, vol. 24, no. 1, pp. 5–34, 2004.

[50] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551–1555, August 2002.

[51] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[52] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.

[53] S. Dorogovtsev and J. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[54] P. Holland and S. Leinhardt, "A method for detecting structure in sociometric data," *American Journal of Sociology*, vol. 76, no. 3, pp. 492–513, 1970.

[55] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824–827, 2002.

[56] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nature Genetics*, vol. 31, pp. 64–68, May 2002.

[57] J. Berg and M. Lässig, "Local graph alignment and motif search in biological networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 41, pp. 14689–14694, 2004.

[58] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, *Mfinder tool guide*, 2002. Current version available from http://www.weizmann.ac.il/mcb/UriAlon/NetworkMotifsSW/mfinder/MfinderManual.pdf. Last accessed 04/05/2009.

[59] F. Schreiber and H. Schwöbbermeyer, "MAVisto: a tool for the exploration of network motifs," *Bioinformatics*, vol. 21, no. 17, pp. 3572–3574, 2005.

[60] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[61] M. Kanehisa, "A database for post-genome analysis," *Trends in Genetics*, vol. 13, no. 9, pp. 375–376, 1997.

[62] M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, pp. D354–D357, 2006.

[63] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[64] C. Krieger, P. Zhang, L. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Rhee, and P. Karp, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research*, vol. 32, pp. 438–442, 2004.

[65] R. Caspi, H. Foerster, C. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Rhee, A. Shearer, C. Tissier, T. Walk, P. Zhang, and P. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 36, pp. D623–D631, 2008.

[66] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, and P. Karp, “Computational prediction of human metabolic pathways from the complete human genome,” *Genome Biology*, vol. 6, no. 1, 2004.

[67] M. Arita, “In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism,” *Genome Research*, vol. 13, pp. 2455–2466, 2003.

[68] P. Karp, M. Riley, M. Saier, I. Paulsen, J. Collado-Vides, S. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro, “The EcoCyc database,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 56–58, 2002.

[69] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D’Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, “Reactome: a knowledgebase of biological pathways,” *Nucleic Acids Research*, vol. 33, pp. 428–432, 2005.

[70] I. Vastrik, P. D’Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein, “Reactome: a knowledge base of biological pathways and processes,” *Genome Biology*, vol. 8, no. 3, 2007.

[71] P. Mendes, “Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3,” *Trends in Biochemical Sciences*, vol. 22, pp. 361–363, 1997.

[72] M. Tomita, K. Hashimoto, K. Takahashi, T. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. Venter, and C. Hutchison, “E-CELL: software environment for whole-cell simulation,” *Bioinformatics*, vol. 15, no. 1, pp. 72–84, 1999.

[73] L. Loew and J. Schaff, “The Virtual Cell: a software environment for computational cell biology,” *Trends in Biotechnology*, vol. 19, no. 10, pp. 401–406, 2001.

[74] J. Luciano, “PAX of mind for pathway researchers,” *Drug Discovery Today*, vol. 10, no. 13, pp. 937–942, 2006.

[75] M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, A. Cuellar, S. Dronov, E. Gilles, M. Ginkel, V. Gor, I. Goryanin, W. Hedley, T. Hodgman, J.-H. Hofmeyr, P. Hunter, N. Juty, J. Kasberger, A. Kremling, U. Kummer, N. L. Novére, L. Loew, D. Lucio, P. Mendes, E. Minch, E. Mjolsness, Y. Nakayama, M. Nelson, P. Nielsen, T. Sakurada, J. Schaff, B. Shapiro, T. Shimizu, H. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.,” *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[76] O. Ruebenacker, I. Moraru, J. Schaff, and M. Blinov, "Kinetic Modeling using BioPAX ontology," in *2007 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 339–346, 2007.

[77] C. Klukas and F. Schreiber, "Dynamic exploration and editing of KEGG pathway diagrams," *Bioinformatics*, vol. 23, no. 3, pp. 344–350, 2007.

[78] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, pp. 2498–2504, 2003.

[79] Z. Hu, J. Mellor, J. Wu, and C. DeLisi, "VisANT: an online visualisation and analysis tool for biological interaction data," *BMC Bioinformatics*, vol. 5, 2004.

[80] M. Nakao, H. Bono, S. Kawashima, T. Kamiya, K. Sato, S. Goto, and M. Kanehisa, "Genome-scale gene expression analysis and pathway reconstruction in KEGG," No. 10, pp. 94–103, Genome Inform Ser Workshop Genome Inform, 1999.

[81] T. Dwyer, H. Rolletschek, and F. Schreiber, "Representing experimental biological data in metabolic networks," in *Proceedings of the second conference on Asia-Pacific bioinformatics*, vol. 55 of *ACM International Conference Proceeding Series*, 2004.

[82] B. Junker, C. Klukas, and F. Schreiber, "VANTED: a system for advanced data analysis and visualization in the context of biological networks," *BMC Bioinformatics*, vol. 7, p. 109, 2006.

[83] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, "Observing and interpreting correlations in metabolic networks," *Bioinformatics*, vol. 19, no. 8, pp. 1019–1026, 2003.

[84] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[85] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, "Interpreting correlations in metabolic networks," *Biochemical Society Transactions*, vol. 31, no. 6, pp. 1476–1478, 2003.

[86] R Development Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN: 3-900051-07-0.

[87] W. de Nooy, A. Mrvar, and V. Batagelj, *Exploratory social network analysis with Pajek.* Cambridge University Press, 2005.

[88] US National Cancer Institute, "Definition of alanine transferase." Webpage: http://www.cancer.gov/Templates/db_alpha.aspx?CdrID=372944.

[89] Reactome, "Alanine metabolism." Webpage: http://www.reactome.org/cgi-bin/eventbrowser?DB=gk_current&ID=70525&.

[90] A. Southam, J. Easton, G. Stentiford, C. Ludwig, T. Arvanitis, and M. Viant, "Metabolic changes in flatfish hepatic tumours revealed by NMR-based metabolomics and metabolic correlation networks," *Journal of Proteome Research*, vol. 7, no. 12, pp. 5277–5285, 2008.

[91] J. Siek, L. Lee, and A. Lumsdaine, "Boost graph library." Homepage: http://www.boost.org/libs/graph/. Last accessed 24/08/2009., 2000.

[92] F. Rubin, "Enumerating all simple paths in a graph," *IEEE Transactions on Circuits and Systems*, vol. 25, no. 8, pp. 641–642, 1978.

[93] R. Bellman, "On a routing problem," *Quarterly Applied Mathematics*, vol. 16, pp. 87–90, 1958.

[94] L. Ford and D. Fulkerson, *Flows in networks.* Princeton University Press, 1962.

[95] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[96] A. Frick, A. Ludwig, and H. Mehldau, "A fast adaptive layout algorithm for undirected graphs," in *Proceedings of the DIMACS International Workshop on Graph Drawing* (R. Tamassia and I. Tollis, eds.), no. 894 in Lecture Notes in Computer Science, pp. 388–403, Springer-Verlag, 1994.

[97] Y. Kinoshita and A. Yokota, "Absolute concentrations of metabolites in human brain tumours using *in vitro* proton magnetic resonance spectroscopy," *NMR in Biomedicine*, vol. 10, pp. 2–12, 1997.

[98] M. Esteller, "Aberrant DNA methylation as a cancer-inducing mechanism," *Annual Review of Pharmacology and Toxicology*, vol. 45, pp. 629–656, 2005.

[99] R. Grützmann, C. Pilarsky, O. Ammerpohl, J. Lüttges, A. Böhme, B. Sipos, M. Foerder, I. Alldinger, B. Jahnke, H. Schackert, H. Kalthoff, B. Kremer, G. Klöppel, and H. Saeger, "Gene experssion profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays," *Neoplasia*, vol. 6, no. 5, pp. 611–622, 2004.

[100] J. Groene, U. Mansmann, R. Meister, E. Staub, S. Roepcke, M. Heinze, I. Klaman, T. Brümmendorf, K. Hermann, C. Loddenkemper, C. Pilarsky, B. Mann, H. Adams, H. Buhr, and A. Rosenthal, "Transcriptional census of 36 microdissected colorectal cancers yields a gene signiture to distinguish UICC II and III," *International Journal of Cancer*, vol. 119, no. 8, pp. 1829–1836, 2006.

[101] D. Smirnov, B. Foulk, G. Doyle, M. Connelly, L. Terstappen, and S. O'Hara, "Global gene expression profiling of circulating endothelial cells in patients with metastatic carcinomas," *Cancer Research*, vol. 66, pp. 2918–2922, March 2006.

[102] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. Lara, A. Oezcimen, P. Rocca-Serra, and S. Sansone, "ArrayExpress - a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[103] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004.

[104] M. Murr, M. Sarr, A. Oishi, and J. V. Heerden, "Pancreatic cancer," *CA - A Cancer Journal for Clinicians*, vol. 44, pp. 304–318, 1994.

[105] D. Li, K. Xie, R. Wolff, and J. Abbruzzese, "Pancreatic cancer," *The Lancet*, vol. 363, pp. 1049–1057, March 2004.

[106] V. Smit, A. Boot, A. Smits, G. Fleuren, C. Cornelisse, and J. Bos, "KRAS codon 12 mutations occur very frequently in pancreatic adenocarcinomas," *Nucleic Acids Research*, vol. 16, no. 16, pp. 7773–7782, 1988.

[107] E. Giovannucci, M. Stampfer, G. Colditz, E. Rimm, D. Trichopoulos, B. Rosner, F. Speizer, and W. Willett, "Folate, methionine, and alcohol intake and risk of colorectal adenoma," *Journal of the National Cancer Institute*, vol. 85, no. 11, pp. 875–883, 1993.

[108] D. Russell, "Clinical relevance of polyamines as biochemical markers of tumor kinetics," *Clinical Chemistry*, vol. 23, no. 1, pp. 22–27, 1977.

[109] L. Sharp, J. Little, A. Schofield, E. Pavlidou, S. Cotton, Z. Miedzybrodzka, J. Baird, N. Haites, S. Heys, and D. Grubb, "Folate and breast cancer: the role of polymorphisms in methylenetetrahydrofolate reductase (MTHFR)," *Cancer Letters*, vol. 181, pp. 65–71, 2002.

[110] S. Zhang, W. Willett, J. Selhub, D. Hunter, E. Giovannucci, M. Holmes, G. Colditz, and S. Hankinson, "Plasma folate, vitamin B_6, vitamin B_{12}, homocysteine, and risk of breast cancer," *Journal of the National Cancer Institute*, vol. 95, no. 5, pp. 373–380, 2003.

[111] R. Katz-Brull, D. Seger, D. Rivenson-Segal, E. Rushkin, and H. Degani, "Metabolic markers of breast cancer: enhanced choline metabolism and reduced choline-ether-phospholipid synthesis," *Cancer Research*, vol. 62, pp. 1966–1970, April 2002.

[112] N. Brown and R. Bicknell, “Hypoxia and oxidative stress in breast cancer. Oxidative stress: its effects on the growth, metastatic potential and response to therapy of breast cancer,” *Breast Cancer Research*, vol. 3, pp. 323–327, 2001.

[113] T. Chevalier, I. Schreiber, and J. Ross, “Towards a systematic determination of complex reaction mechanisms,” *The Journal of Physical Chemistry*, vol. 97, pp. 6776–6787, 1993.

[114] A. Arkin and J. Ross, “Statistical construction of chemical reaction mechanisms from measured time-series,” *The Journal of Physical Chemistry*, vol. 99, pp. 970–979, 1995.

[115] R. Díaz-Sierra, J. Lozano, and V. Fairén, “Deduction of chemical mechanisms from the linear response around steady state,” *The Journal of Physical Chemistry A*, vol. 103, pp. 337–343, 1999.

[116] A. Sorribas, J. Lozano, and V. Fairen, “Deriving chemical and biochemical model networks from experimental measurements,” *Recent Research Developments in Physical Chemistry*, vol. 2, pp. 553–573, 1998.

[117] H. Schmidt, K. Cho, and E. Jacobsen, “Identification of small-scale biochemical networks based on general type system perturbations,” *The FEBS Journal*, vol. 272, pp. 2141–2151, 2005.