#### 2. Shortest Path Analysis in Metabolic Networks

#### 2.1. Biological Network

The state of a cell consists of all those parameters--both internal and external - which determine its behaviour. Following the **Central Dogma** of molecular biology, the activity of a cell is determined by which of its genes are, and are not being expressed. If a particular gene is being expressed, its DNA is transcribed into complementary messenger RNA (mRNA), which is then translated into the specific protein the gene codes for. We can measure the level of expression of each gene by measuring how many mRNA copies are present in the cell.

Apart from the manifestation of the classical central dogma, of **DNA** -> **mRNA** -> **protein** pathway, it is important to note that the genes in the DNA are themselves regulated by the presence or absence of certain proteins. Furthermore, many of the interactions going on in the cell occur entirely at the protein level, which can cause significant discrepancies between protein and mRNA levels. In a recent comparison of selected mRNA and protein abundances in the human liver, a correlation of only 0.48 was observed between the two (Anderson and Seilhamer 1997). Clearly, protein levels form an important part of the internal state of a cell.

In addition to mRNA and protein levels, one could imagine measuring a number of other parameters, including cell volume, growth rate, methylation states of DNA, phosphorylation state of proteins, localization of proteins and mRNA within the cell, ion levels, etc. One class of data which prove be very useful is metabolite and nutrient levels.

In the post genomic era the key aim for researchers is to understand the interactions between various biological processes within the living cell (Hartwell, Hopfield et al. 1999; Wagner and Fell 2001; Barabasi and Oltvai 2004; Xia, Yu et al. 2004). Biological networks can be broken down into genetic (Hasty, McMillen et al. 2002; Wall, Hlavacek et al. 2004), protein (Koonin, Wolf et al. 2002) and metabolic (Bray 2003) networks. Various types of interaction webs, or networks, (including protein– protein interaction (Lappe and Holm 2004), metabolic, signalling and transcriptionregulatory networks (Ideker 2004)) emerge from the sum of these interactions (Kitano 2002; Oltvai and Barabasi 2002). None of these networks are independent, instead they form a 'network of networks' that is responsible for the behaviour of the cell (Oltvai and Barabasi 2002; Alon 2003). A major challenge of contemporary biology is to embark on an integrated theoretical (Kelley, Sharan et al. 2003) and experimental programme to map out, understand and model in quantifiable terms the topological and dynamic properties (Barabasi and Albert 1999) of the various networks that control the behaviour of the cell (Wagner and Fell 2001; Papp, Pal et al. 2004).

Currently, most studies trying to infer expression mechanisms from cell state data use mRNA levels, because they are the easiest to measure (especially with the new large-scale gene expression technologies). Large-scale protein measurements tend to be incomplete (typically only measuring the highest abundant proteins), but can be supplemented with more exact measurements of individual proteins, which are known to play an important role. If most protein levels turn out to be exactly correlated with the corresponding mRNA levels, they can always be left out of the model. Similarly, when measuring gene expression data in a process involving cellular metabolism an effort should be made to quantify the most important metabolite and nutrient levels.

For example, in the diauxic shift, yeast switch from anaerobically fermenting glucose to form ethanol to aerobically consuming this ethanol when the glucose is exhausted. Analysis of the diauxic shift using DNA microarrays revealed that very many genes change their expression during this switch in nutrient utilization. The interrelationship between specific gene expression changes and metabolic changes however, remains unclear primarily because the temporal course of changes in metabolite concentrations and fluxes has not been studied adequately.

For example, by using information on the network topology from genome-scale protein & metabolic reconstruction, it is possible to reveal patterns in these **networks** that follow a common transcriptional response. The gene expression pattern is a strong indicator of perturbations induced in specific part(s) of the metabolic network. Any changes due to the perturbation are then propagated through the biological network because of the highly connected nature of metabolism.

In the following section I will introduce some biological networks and their design principles.

# 2.1.1. Gene Network

"The mRNA levels sensitively reflect the state of the cell, perhaps uniquely defining cell types, stages, and responses. To decipher the logic of gene regulation, we should aim to be able to monitor the expression level of all genes simultaneously ... " [Eric S. Lander] (Lander 1996)

The gene network inference techniques tend to be data-hungry. Measuring gene expression time series has the fastidious feature of yielding lots of data. However, all the data points in a single time series tend to be about a single dynamic process in the cell, and are related to the surrounding time points. A data set of ten expression measurements under different environmental conditions, or with different mutations, will actually contain more information than a time series of ten data points on a single phenomenon. The advantage of the time series is that it can provide crucial insights in the dynamics of the process.

In a genetic network genes are nodes and their interaction is given by edges. Most of the genetic networks are Boolean in nature thereby the weights on the edges can be assigned a weight of either 0 or 1. Both types of data, and multiple data sets of each, will likely be needed to unravel the regulatory interactions of the genes. Indeed, to correctly infer the regulation of a single gene, we need to observe the expression of that gene under many different combinations of expression levels of its regulatory inputs. This implies a wide variety of different environmental conditions and perturbations.

#### 2.1.2. Protein Networks

Proteins perform distinct and very well-defined tasks, but little is known about how interactions among them are structured at the cellular level. In a protein interaction network each link in the network is assigned a length of 1. In principle, if a protein interacts with its partner, the link was designed as one. If however, a protein does not interact with other proteins, zero was given in the link. For constructing the network, the basic principle therefore follows adjacent matrix that is a matrix with rows and columns labeled by graph nodes with a 1 or 0 in position (i, j) according to whether i and j are adjacent or not. Protein complexes and modules are derived from clustering the protein interaction network.

A study of protein interaction network in yeast cells found that they are not random, but well organized (Fraser, Hirsh et al. 2002). In another study by Schwikowski *et.al* (Schwikowski, Uetz et al. 2000) a global analysis of 2,709 published interactions between proteins of the yeast *Saccharomyces cerevisiae* was performed, enabling the establishment of a single large network of 2,358 interactions among 1,548 proteins. They found that proteins of known function and cellular location tend to cluster together, with 63% of the interactions occurring between proteins with a common functional assignment and 76% occurring between proteins found in the same subcellular compartment. They suggested that possible functions can be assigned to a protein based on the known functional category for 72% of the 1,393 characterized proteins with at least one partner of known function, and 364 previously uncharacterized proteins.

Apparently in scale-free protein–protein interaction networks, or 'interactome' networks (Jeong, Mason et al. 2001; Li, Armstrong et al. 2004), most proteins interact with few partners, whereas a small but significant proportion of proteins, the 'hubs', interact with many partners (Figure 8). Both biological and non-biological scale-free networks are particularly resistant to random node removal but are extremely sensitive to the targeted removal of hubs (Jeong, Mason et al. 2001). A link between the potential scale-free topology of interactome networks and genetic robustness (Winzeler, Shoemaker et al. 1999; Gu, Steinmetz et al. 2003) seems to exist, because knockouts of genes (Winzeler, Shoemaker et al. 1999; Giaever, Chu et al. 2002)

encoding for hubs are approximately **three** times more likely to confer lethality than those of non-hubs (Jeong, Mason et al. 2001). Han *et.al* (Han, Bertin et al. 2004) investigated how hubs might contribute to robustness and other cellular properties for protein—protein interactions dynamically regulated both in time and in space. They uncovered two types of hub: 'party' hubs, which interact with most of their partners simultaneously, and 'date' hubs, which bind their different partners at different times or locations. Both, *in silico* studies of network connectivity and genetic interactions described *in vivo* support a model of organized modularity in which date hubs organize the proteome, connecting biological processes—or modules (Hartwell, Hopfield et al. 1999), to each other, whereas party hubs function inside modules.



Figure 8. Map of protein-protein interactions in yeast. Each point represents a different protein and each line indicates that the two proteins are capable of binding to one another. Only the largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown)

Adapted from: [H. Jeong, S. P. Mason, A.-L. Barabási & Z. N. Oltvai, "Lethality and centrality in protein networks", Nature, Vol 411, p41, 3 May 2001]

#### 2.1.3. Metabolic Networks

The **metabolism** of an organism is the basic biochemical system that generates essential components such as amino acids, sugars and lipids and the energy required to synthesize and use them in creating proteins and cellular structures. *This system of connected chemical reactions is known as metabolic network* (Figure 9). The metabolites correspond to nodes in the graph, and reactions correspond to connections between these nodes.



Figure 9. General Metabolic Pathway. Adapted from http://www.genome.jp/kegg/pathway/map/map01100.html

The metabolic network can be broken down into sub-networks consisting of sets of metabolites, reactions and enzymes specialized in certain biochemical process. Below is an example of such a network. The partitioning of the metabolic network into pathways is not always straightforward (Schuster, Fell et al. 2000). As a result there may be correlations that are not visible in a pathway-oriented perspective which emerge in a whole-network oriented view. There is also an element of arbitrariness involved in assigning promiscuity to compounds (compounds involved in many reactions, e. g. H<sub>2</sub>O, ATP and other cofactors). This is the subsystem *for the synthesis of the amino acids valine, leucine and isoleucine*. This image (Figure 10) was taken from the KEGG (Kanehisa, Goto et al. 2004) database. The green boxes indicate enzymes, which have been, identified in the organism, in this case *Corynebacterium glutamicum*.



Figure 10. Valine, Leucine and Isoleucine in Corynebacterium glutamicum. Adapted from KEGG http://www.genome.jp

With the help of such a graphic representation, Barabasi and his coworkers have studied the structure of metabolic networks using methods adopted from studies of the world wide web (Adamic, Huberman et al. 2000). It was found that the metabolic network of 43 distinct organisms are organized into many small, highly connected topologic modules that combine in a hierarchical manner into larger, less cohesive units, with their number and degree of clustering following the power law (Ravasz, Somera et al. 2002). They found that like all the other networks studied, metabolic networks exhibited typical characteristics of small world networks, thus implying that most nodes have a low connection degree, while few have a very high connection degree.

Indeed, recent studies have demonstrated that the probability that a substrate can react with *k* other substrates [the degree distribution P(k) of a metabolic network] decays as per the power law  $P(k)\sim k^{\gamma}$  with  $\gamma \cong 2.2$  in all organisms (Jeong, Tombor et al. 2000; Wagner and Fell 2001), suggesting that metabolic networks have a scale-free topology (Barabasi and Albert 1999). A distinguishing feature of such scale-free networks is the existence of a few highly connected nodes (e.g., pyruvate or coenzyme A), which participate in a very large number of metabolic reactions. It was found that within *Escherichia coli*, the uncovered hierarchical modularity closely overlaps with known metabolic functions. Further, it was speculated that the identified network architecture may be generic to system-level cellular organization (Ravasz, Somera et al. 2002).

#### 2.1.4. Network Analysis Emphasizing Shortest Path

Previously mentioned studies based on graph theory were shown to be useful for network structure analysis. It was found that the high degree nodes dominate the network structure and are called hubs of the network. Most of the nodes are connected through them by a relatively short path (Strogatz 2001; Albert and Barabasi 2002). For metabolic networks, Jeong *et al.* (H. Jeong 2001) calculated the average path length (AL) for 43 organisms and found that AL was almost the same (approximately 3.2) for all the organisms. This means that most of the metabolites can be converted to each other in approximately only 3 steps. From the biochemical point of view this result is surprising because often a path longer than 3 steps is used for the synthesis of many metabolites.

#### 2.1.4.1. Existing Shortest Path Analysis Model for Metabolic Networks

A metabolic network is a directed graph, implying that reactions involved in the metabolic pathways are both reversible and irreversible. Hence, information about reaction reversibility is important in network analysis. The study conducted by Jeong et al. (H. Jeong 2001) lacked this information. There is still no metabolic reaction database that gives clear and sufficient information about it. Also, in the study by Jeong et al. (H. Jeong 2001), even promiscuous metabolites i.e. ATP, ADP were regarded as nodes in the network. This resulted in an unrealistic definition of the path length in many cases as illustrated with a part of the Glycolysis pathway (Figure 11). It is obvious that the path length (number of reaction steps in the pathway) from glucose to pyruvate should be nine in terms of biochemistry. However, having included ATP and ADP as nodes in the network and due to the functions of ATP and ADP as cofactors in many reaction steps, the path length between glucose and pyruvate was concluded as being 2 steps long (pink line in (Figure 11)). This calculation of path length is obviously not meaningful in biochemical terms. Similar problems also exist regarding the consideration of other current metabolites such as ATP, NADH, NAD+ etc.



Figure 11. Shortest Path analysis for the Glycolysis pathway: The green lines represent one of the valid paths between 'alpha-D-glucose' and 'pyruvate', while the pink line represents a bio-chemically invalid shortest path via ADP

#### 2.1.5. Characterizing the Links - Manually Curated Models for Path Finding

Ma *et al.* (Ma and Zeng 2003) used a revised bio-reaction database to reconstruct the metabolic networks of 80 fully sequenced organisms from genome data. They distinguished the major co-factor metabolites from normal metabolites and considered the reaction direction. They were able to show that quantitative differences exist in the network structure of the three domains of organisms, namely *eukaryotes, archaea* and *bacteria*.

Although the network showed small world phenomenon as it followed the power law, there was a change in the top 10-metabolite hubs (as many cofactors i.e. ATP, NADH were already removed). The average path length from glucose to all the reachable metabolites was calculated using "Breadth First Search (BFS)" as 7.68. The average path length (AL) for the whole network for E. coli was 8.20. Another structure parameter is network diameter. It is defined as the path length of the longest pathway among all the shortest pathways (Batagelj and Mrvar 1998). The longest shortest path length of this pathway is 15. The network diameter can be calculated by comparing the longest pathway for all the metabolites. It is 23 for E. coli. AL and network diameter tend to increase with an increase in network scale, especially for small-scale networks (such as networks with less than 300 nodes). Parasites were found to have small-scale and not well-connected networks and they contained many separated small networks or pathways. This resulted in a relatively short average path length. These results are consistent with the fact that the parasitic organisms have lost a great number of metabolic genes in the evolutionary process to adapt to the environments of host cells (Podani, Oltvai et al. 2001). For the relatively large networks (i.e. networks with more than 300 nodes), the relation between AL and network scale (node number) is not very clear. Even for networks with a similar scale, the AL values vary greatly. For example, Rattus norvegicus (rno) and Vibrio cholerae (vch) have a similar scale, while their AL values are 10.99 and 7.64 respectively. These results clearly differ from the results of Jeong *et al.* (Jeong, Tombor et al. 2000) as they found a nearly constant and much shorter average path length for different kinds of organisms when using promiscuous metabolites as connections.

*The eukaryotes and archaea have a longer AL than bacteria*. The average AL values for these three domains of organisms are 9.57, 8.50 and 7.22 (7.73 for bacteria without considering parasites), respectively. Average diameter values are 33.1, 23.4 and 20.6, respectively. Eukaryotes and archaea have a longer path length and a larger network diameter than bacteria, indicating a difference in the structure of their respective metabolic networks (Ma and Zeng 2003).

This indicates that although the fundamental structure of metabolic network is similar for all the organisms, they do exhibit quantitative differences in the metabolic network structure as described by the parameters average path length and network diameter. For example, it may be interesting to identify the shortcuts that lead to the lower AL for bacteria. The reactions (enzymes and genes) corresponding to these shortcuts can then be determined. These enzymes may be specific for bacteria and important for metabolic conversion or pathogenic functions. This class of knowledge can be used for strain improvement through metabolic engineering or for selecting pathway targets to develop drugs against pathogenic bacteria. Hence, it is clear from the above study (Ma and Zeng 2003) that chemically meaningful connectivity in the biochemical networks could lead to a better understanding of the network function and organization in various organisms.

# 2.1.6. Adding Biochemical Meaning: Finding Biochemically Valid Shortest Paths (SP) in the Network

A metabolite (often called a small-molecule) consists of several atoms namely carbon, hydrogen etc. and their orientation in space embed certain structural patters like a carbon-ring structure in case of glucose. Although several groups confirmed the small-world property of small-molecule metabolisms in multiple data sources, the details of their results differ depending on the purpose of the analysis and its data-preparation scheme (Fell and Wagner 2000; Jeong, Tombor et al. 2000; Wagner and Fell 2001; Ravasz, Somera et al. 2002; Ma and Zeng 2003; Ma and Zeng 2003).

- Jeong *et al.* (Jeong, Tombor et al. 2000) computed the proximity of metabolites by regarding all substrates and products in the same reaction as adjacent.
- Wagner and Fell (Wagner and Fell 2001) computed stoichiometric relationships to estimate the transmission degree of perturbations in the metabolic network. They used the metrics with and without coenzymes such as ATP and NAD in both substrate- and reaction-based networks to compare their differences.
- Ma and Zeng (Ma and Zeng 2003) manually specified links in each reaction, aiming to delineate only physical relationships responsible for biosynthesis and degradation. To reproduce biochemical pathways in the traditional metabolic map however, metabolites to be linked cannot be defined *per se* by compounds or reactions. The biochemical link between metabolites is contextsensitive; it depends on the conserved structural moieties in the adjacent reactions.

Chapter 2

All of these studies used the same algorithmic procedure, and discrepancies are ascribable to the varied aims of their network analyses. The main differences are attributable to the reversibility of enzymatic reactions and to the treatment of metabolically ubiquitous compounds referred to as coenzymes or inorganic compounds. Many studies have been conducted to elucidate the organizational and evolutionary principles of the metabolic network using graph-theoretic analysis of large biochemical networks responsible for the synthesis and degradation of cellular building blocks (Jeong, Tombor et al. 2000; Wagner and Fell 2001; Ma and Zeng 2003). In such studies, the global properties of the network are computed by considering enzymatic reactions as links between metabolites. However, the pathways computed in this manner do not conserve their structural moieties and therefore do not correspond to biochemical pathways on the traditional metabolic map.

Arita (Arita 2004) proposed a new model to reassess earlier results by digitizing carbon atomic traces (i.e. atomic position pairs between substrates and products corresponding to the sub-structural moieties conserved in each reaction (Figure 12) in metabolic reactions annotated for *Escherichia coli*. In this study a metabolic pathway (shortest path) from metabolite X to Y is defined as a sequence of biochemical reactions through which at least one carbon atom in X reaches Y. Only carbon atoms are considered throughout this study. A metabolite Y is considered reachable from X if there is a pathway from X to Y. He found that the average path length (AL) of Escherichia coli metabolism is much longer than previously thought and that the metabolic world of this organism is not small in terms of biosynthesis and degradation (Table 2) .The metabolite "hubs" was defined by the number of times a metabolite structure pattern was involved in a path rather than being defined solely by connectivity. With this information, he reassessed the global properties of metabolic networks with special emphasis on the small-world hypothesis. This analysis gave a completely new dimension to path finding in the bio-chemical network and it was widely accepted both in the biochemical society as well as the non-biochemical community. Table 2 summarizes differences in the major analyses and compares the AL and hub metabolites they identified.



Figure 12. Adapted from Arita's (Arita 2004) paper. Two ways to represent the reaction of EC 2.3.1.35. In this reaction, the acetyl moiety of N-acetyl L-ornithine is transferred to L-glutamate to form N-acetyl L-glutamate. (Lower Left) In the scheme of *Jeong et al. (Jeong, Tombor et al. 2000)*, its two substrates and two products are equally linked to the object representing the EC number, irrespective of their structural changes. (Lower Right) In *Arita's (Arita 2004)* scheme, conserved sub-structural moieties, coded in different colors, are computationally detected and each link is associated with the information of which atom goes where.

Table 2. Comparison of four E. coli network analyses. The top 10 hub metabolites and ALs reported in each study. Wagner and Fell (Wagner and Fell 2001) computed several versions of the network. The one shown here is the substrate-based network where ATP, ADP, NAD, NADP, NADH, NADPH, carbon dioxide, ammonia, sulfate, thioredoxin, (ortho) phosphate (P), and pyrophosphate (PP) are removed.

Top 10 hubs	Jeong et al. 2000,	Wagner and Fell 2001,	Ma and Zeng 2003, directed	Arita 2004,
	directed	undirected		(un)directed
1	H <sub>2</sub> O	L-glutamate	glycerate 3P	carbon dioxide
2	ADP	Pyruvate	D-ribose 5P	Pyruvate
3	Р	CoA	acetyl CoA	acetyl CoA
4	ATP	α-keto glutarate	pyruvate	ATP
5	L-glutamate	L-glutamine	D-xylulose 5P	D-glucose
6	$NADP^+$	L-aspartate	D-fructose 6P	L-glutamate
7	РР	acetyl CoA	5P-D-ribose 1PP	D-galactose
8	$\mathrm{NAD}^{+}$	phosphoribosyl PP	L-glutamate	CoA
9	NADPH	Tetrahydrofolate	D-glyceraldehyde 3P	S-adenosyl L-methionine
10	NADH	Succinate	L-aspartate	D-5-phosphoribosyl-1P
AL	3.2	3.8	8.2	8.4 (8.0)

The artefact in Arita's algorithm was that it does not allow the same compound to appear iteratively in the same pathway although every reaction is reversible in the graph (this was done to avoid combinatory explosion). The cyclic paths are a hindrance to such an analysis and the gaps in the network annotation are another bottleneck. There thus exists the need for an algorithm, which supplies appropriate atomic-level information about the reactions. This vacuum in the computational analysis of metabolic networks provided me with the motivation for my study.

#### 2.2. Thermodynamic Feasibility of the Metabolic Pathways

The feasibility and reversibility of a reaction is determined by its equilibrium constant and the concentrations of its reactants (also called substrates) and products. Because intracellular concentrations vary within limited ranges (e.g.,  $1\mu$ M to 5mM), the equilibrium constant alone is sufficient for reaching a qualitative conclusion on a reaction's feasibility. In general, a feasible and irreversible reaction is characterized by an equilibrium constant, K, much larger than 1. A feasible and reversible (i.e., feasible in both the forward and reverse directions) reaction is characterized by an equilibrium constant of the order of 1. A reaction that is infeasible in the forward direction but feasible in the reverse direction is characterized by equilibrium constant much smaller than 1. The quantitative interpretation of these criteria depends on the range of permissible intracellular concentrations for metabolites.

The standard Gibbs free energy (2.1) of formation of a compound is the change of Gibbs free energy that accompanies the formation of 1 mole of that substance from its component elements, at their standard states (the most stable form of the element at 25 degrees Celsius and 100 kilopascals).

$$\Delta G = \Delta G^{\circ} + RT \ln K \tag{2.1}$$

If the free energy change in a chemical reaction is negative, the reaction can occur spontaneously. If the free energy change in a chemical reaction is positive, the reaction cannot occur spontaneously. This is obvious if you remember that free energy changes correspond to work being done. When  $\Delta G$  is negative, the system will give you work, do work for you. When  $\Delta G$  is positive, you have to do work on it. Suppose you are pushing a large rock up a hillside. As long as you keep pushing (doing work on it, putting work into it) it continues to move uphill. But let go, and it will roll down the hill and, if you are standing in the way, it will do considerable work on you in the process.

A negative  $\Delta G$  corresponds to a "downhill" direction while a positive  $\Delta G$  corresponds to an "uphill" direction. And, just as up is opposite to down, so the

direction of a chemical reaction for which  $\Delta G$  is negative, as for example C(s) + O<sub>2</sub>(g) --> CO<sub>2</sub>(g) which takes place spontaneously, is opposite to the direction of the reaction CO<sub>2</sub>(g) --> C(s) + O<sub>2</sub>(g) for which  $\Delta G$  is positive and which we know does not take place spontaneously.

$$\Delta G = \sum \Delta G_{f(PRODUCTS)} - \sum \Delta G_{f(REACTANTS)}, \qquad (2.2)$$

The values for  $\Delta G$ , since (2.2) are numerically the same but opposite in sign for these two reactions. The value of the free energy change of a chemical reaction changes only in sign when the direction of the reaction is reversed<sup> $\omega$ </sup>.

<sup>&</sup>lt;sup>60</sup> (refer <u>http://www.intute.ac.uk/sciences/reference/plambeck/chem2/p02052.htm</u>)

#### 2.3. Motivation and Proposed Model

With the advent of the "omics" era more and more system-based approaches to biological functions are being developed. Metabolome analysis and metabolomics are gaining higher attention and help to understand the complexity of the underlying cellular networks in organisms (Kromer, Sorgenfrei et al. 2004; Strelkov, von Elstermann et al. 2004; Verhoeckx, Bijlsma et al. 2004; Oliver 2006). The sequencing of a large number of genomes has made the comparative study of genomes possible at different levels (Kromer, Sorgenfrei et al. 2004; Catchpole, Beckmann et al. 2005). One means of gaining a deeper insight into the sequenced genomes is to analyse the underlying metabolic network(s) and its topology in different genomes (Cakir, Patil et al. 2006). For example, a virtue of studying metabolism in systems like yeast is that metabolites are remarkably conserved between species-even more so, in fact, than genes. Thus, lessons learned in unicellular organisms can rapidly inform medical issues, such as the molecular basis of metabolic malfunctioning that occurs in human disease. Several databases provide information about metabolic pathways i.e. KEGG (Kanehisa, Goto et al. 2004), BioCyc (Karp, Ouzounis et al. 2005), BRENDA (Schomburg, Chang et al. 2004) and PROSITE (Hulo, Sigrist et al. 2004).

A global view of the metabolic network highlights the contribution and usage of metabolites. Shortest path analysis (Fell and Wagner 2000; Jeong, Tombor et al. 2000; Wagner and Fell 2001; Ravasz, Somera et al. 2002; Ma and Zeng 2003; Simeonidis, Rison et al. 2003; Arita 2004; Croes, Couche et al. 2006) is one of the best-defined methods to analyse a graph (metabolic pathways) at different levels in terms of local and global connectivity. Fact is that in the metabolic process, not just one atom (in a compound) but usually a set of atoms are involved in the reaction process. Utilising information about the overall change(s) in the structural pattern for the metabolite(s) will improve the connectivity depicted in metabolic networks based on substrate-level network topology does not always correspond to metabolic pathways. Hence, structural information of metabolites is indispensable for computing biochemical pathways. Thus, atomic-level analysis complements virtually any

metabolism-related study, from gene annotation to network evolution, to potential drug targeting.

We propose an *ab-initio* algorithm to compute k-shortest paths (more than one competing shortest path) in the given metabolic network. We developed a novel algorithm to computationally map substrates to the products in the biochemical reaction. This will help us to dynamically reconstruct the network connectivity and avoid false links in the pathway analysis. In order to make the path finding biochemically meaningful we track the overall change(s) in the metabolite structural pattern both at the local and global level. Further, the path finding algorithm is able to handle constraints like 'via' and 'not via' certain enzyme(s)/metabolite(s). It can calculate various statistical measures on the network topology i.e Average Shortest Path (AL), Degree distribution etc. Our algorithm was coded in the new tool which was termed Pathway Hunter Tool (PHT) (Rahman, Advani et al. 2005) The database for PHT has more than 200 organisms and their pathway maps. The basic version of PHT is available on the Internet, while the more comprehensive stand-alone version is in use in-house and by close collaborators. The output of this tool connects metabolic information back to the enzymes and genes of the genome. PHT is a robust and userfriendly "Systems Biology"-based Bioinformatics tool to process biologically relevant information about shortest path analysis.

# 2.4. Method

This study addresses the bio-chemically valid k-shortest path discovery problem in the context of biological networks. We can model a metabolic network as a bipartite graph where metabolites are connected to reach other via a reaction/enzyme (refer to definition 2 & 6). Here we will consider a metabolic network as an unweighted directed graph (refer to definition 6). The input to the problem is a set of graphs in which nodes correspond to metabolites and edges correspond to interactions between these metabolites.

**Problem:** Given a graph G = (V, E, L), find the *k*-shortest path between source *u* and sink *v*, such that  $d_G(u, v)$  of two vertices  $u, v \in V$  is the minimum number of edges of shortest paths between *u* and *v* in *G*. Each  $v \in V(G)$  is associated with a set of reaction pair  $L(v) \subseteq L$ . Each edge  $uv \in E(G)$  represents a reaction between *u* and *v*.

# 2.4.1. Database and Input

The knowledge about genes, enzymes, reactions, compounds and pathways in various organisms was obtained from two very comprehensive databases KEGG (Kanehisa, Goto et al. 2004) and BRENDA (Schomburg, Chang et al. 2004).

# 2.4.1.1. KEGG

KEGG provides a knowledge-based approach for network prediction by reference knowledge base that integrates current knowledge on molecular interaction networks such as pathways and complexes generated by genome projects and information about biochemical compounds and reactions. The RPAIR contains curated chemical structure transformation patterns extracted from known enzymatic reactions (Kanehisa, Goto et al. 2006). Additionally, drug information is now stored separately and linked to new KEGG DRUG structure maps.

# 2.4.1.2. BRENDA

The enzyme centric database BRENDA (BRaunschweig ENzyme DAtabase) (Schomburg, Chang et al. 2004) represents a comprehensive collection of enzyme and metabolic information, based on primary literature. The database contains several enzymes (83,000 different enzymes) obtained from approximately 9800 different organisms. These enzymes have been classified in approximately 4200 EC numbers with chemical structure information (ligands), controlled vocabulary, ontology and a thesaurus for ligand names.

# 2.4.1.3. Enzymes

The enzymes in the metabolic pathway are divided into six different classes. Each class consists of certain reaction pattern as substrate(s) and product(s) share common molecular structures between them as depicted in Figure 13.





Interactions between these enzyme classes can give rise to a metabolic pathway.

# 2.4.1.4. Enzyme Centric View and Metabolic Centric View

Using graph theory we can define the system in terms of a bipartite graph (Figure 7), which can be reduced to an enzyme-centric graph (Figure 7) and a metabolic-centric graph (Figure 7). In a bipartite view, two nodes share a common enzyme and the edges define the biological relationship between a set of metabolites and enzymes. In the metabolic-centric view metabolites are nodes and reactions/enzymes are edges whereas in the enzyme-centric view, enzymes are nodes and metabolites are edges.

# 2.4.2. Modelling the Network

In the metabolic network, we are looking for bio-chemically valid interactions that are connected and exhibit similar patterns (common molecular structure or maximum common subgraph refer to to section 2.4.1.3). In the context of biological networks (metabolic pathways), labeling of nodes is based on the assessment of functional correspondence, as suggested by similarity of molecular structures or hierarchical classification of enzymes<sup>Y</sup>. Without loss of generality, nodes can be referred to as metabolites and labeled on the basis of the assignment of these enzymes and similarity of molecular structures.

Hence the problem can be reformulated and the path finding algorithm can be divided into two steps.

- Decomposing a given network into reaction subgraphs and finding meaningful interactions in the network by looking into the metabolic structure similarity.
- Since the above procedure will reduce the number of false interactions the pruned graph can then be used to find a biochemical k-shortest path between a substrate and a product.

Since the observed interactions of metabolites in the network represent a superimposition of dynamically organized molecular structural in spatial and temporal

Y http://www.chem.qmul.ac.uk/iubmb/

dimensions, this model accurately captures the dynamic and complex organization of cellular processes. This will improve the runtime of the algorithm and make path finding more meaningful.

The model of metabolic network can be defined as:

**Definition 7: Metabolic network model:** Given a set of metabolites V in a network and set of interactions E between these metabolites, and a many-to-many mapping of these metabolites L, the corresponding metabolic network is modeled using a bipartite graph G = (V, E, L). Each  $v \in V(G)$  is associated with a set of reaction pair  $L(v) \subseteq L$ . Each edge  $uv \in E(G)$  represents a reaction between u and v.

In molecular networks, a connected graph may be interpreted as a set of interactions related to each other through at least one molecule. Therefore, interactions that are related to a particular cellular process are expected to form a connected common subgraph or isomorphism. Such subgraphs may also be connected to each other as a reflection of a crosstalk between different processes.

#### 2.4.3. Decomposition of Network to Map Metabolites in the Reactions

The network can be decomposed into smaller subgraphs of reaction and metabolite pairs. This will help us find common structural pattern shared between substrate and product metabolites.

The metabolic network can be defined as set of subgraphs consisting of reactions and compounds (metabolites).

**Input:** A set of reaction networks  $G = \{G_1 = (V_1, E_1, L), G_2 = (V_2, E_2, L), ..., G_m = (V_m, E_m, L)\}$ , each belonging to a different reaction.

Each metabolite has a certain structural pattern and these patterns can be represented in the form of a graph. The input graph was transformed into 4 tuple definition graph in order to represent metabolites as sub-graphs (Le, Ho et al. 2004): **Definition 8:** A metabolic network based graph is a 4-tuple  $G = (V, E, \mu, v)$  where V is a set of finite vertices,  $E \subseteq V \times V$  is the set of edges,  $\mu: V \rightarrow L_V$  is a function assigning labels to the vertices and  $v: E \rightarrow L_E$  is a function assigning labels to the edges.

The connectivity between the metabolites (substrate and product) is defined as:

**Definition 9:** Graph G is called a connected graph if and only if there is at least one path between any vertex pair, where a path is a list of vertices such that there is an edge between two adjacent vertices.

**Definition 10a:** A graph **S** is a subgraph of reaction network **G**, i.e.,  $S \circ G$ , if there is an injective mapping  $f: V(S) \to V(G)$  such that for all  $v \in V(S)$ ,  $L(v) \subseteq L(f(v))$  and for all  $uv \subseteq E(S)$ ,  $f(u)f(v) \in E(G)$ . A subgraph **S** is connected if and only if for any subset  $U \subset V(S)$ ,  $\exists u \in U$  and  $v \in V(S) \setminus U$  such that  $uv \in E(S)$  or  $vu \in E(S)$ .

**Definition 10b:**  $G_i$  is called a common subgraph of G and G' when  $G_i$  is a subgraph of both G and G'.

# 2.4.4. Algorithm for Finding Valid Links Between Substrate and Product

Molecular fingerprints can be used to define metabolites (here treated as graphs) as a set of binary string representation called *bitstrings* (Holliday, Ranade et al. 1995; Flower 1998; Schuffenhauer, Gillet et al. 2000; Bender, Mussa et al. 2004). BitStrings consist of sequence of '0's and '1's which are binary representation of certain chemical feature present or absent in a molecule (allows fast structural comparison between two molecules). Though it is a heuristic method for representing biomolecular graph it captures various stereo-chemistry of the bio-molecules (Hubalek 1982; Schuffenhauer, Gillet et al. 2000; Hattori, Okuno et al. 2003; Whittle, Willett et al. 2003). Where the attribute values are restricted to 0 and 1, the expressions used for the various similarity and distance measures can often be substantially simplified. We have used the Chemistry Development Kit (CDK) (Steinbeck, Han et al. 2003)\* to calculate the binary fingerprints for molecules. CDK uses hashed fingerprints which is generated by calculating the possible linear path in a molecule between one to defined number of atoms (default is 7). Each of these paths in turn serves as the input to a second program that uses hashing procedure to set a small number of bit (default four or five) to "1" in the fingerprint bitstrings. Each bit in the fingerprint can be set by a different pattern and in case of collision it does not result in false negatives (during similarity search) but they result in more hits.

# 2.4.4.1. Calculation of Similarity Score between Substrate and Product based on Molecular Fingerprint

In this section the new algorithm for mapping substrate to the product is described.

# Definition 11: (Chemical structural similarity between molecules)

Given a subgraph  $G_i = (V_i, E_i, L)$  of a graph G.

<sup>\*</sup> Descriptors are similar to Daylight (<u>http://www.daylight.com/dayhtml/doc/theory/theory.finger.html</u>)

If each source  $u \in V(G_i)$  and  $v \in V(G_i)$  sink is associated with reaction  $L(v) \subseteq L$  and each edge  $uv \in E(G)$  represents the reaction between **u** and **v**.

#### Then let $S_{uv}$ represents the similarity be u and v.

For metabolites u and v characterized by vectors Xu and Xv containing n binary values (such as fingerprints) we can analytically formulate the fingerprint vector as:

$$a = \sum_{i=1}^{i=n} \chi u_i \quad number \text{ of bits "1" in } Xu \quad (2.3)$$

$$b = \sum_{i=1}^{i=n} \chi v_i \quad number \text{ of bits "1" in } Xv \quad (2.4)$$

$$c = \sum_{i=1}^{i=n} \chi u_i \chi v_i \quad number \text{ of bits "1" in both } Xu \text{ and } Xv \quad (2.5)$$

$$d = \sum_{i=1}^{i=n} (1 - \chi u_i - \chi v_i + \chi u_i \chi v_i) \quad number \text{ of bits "0" in both } Xu \text{ and } Xv$$

$$(2.6)$$

$$\therefore n = a + b - c + d$$
 (2.7)

If we define  $\chi u$  as the set of all elements  $\chi u_i$  in vector Xu whose value is 1 (the "on" bits) and  $\chi v$  as the set of all elements  $\chi v_i$  in vector Xv whose value is 1.

Then the above equations can be transformed using set-theoretic notation as

$$a = |\chi u| \quad number \ of \ bits \ "on" \ in \ Xu \tag{2.8}$$

$$b = |\chi v| \quad number \ of \ bits \ "on" \ in \ Xv \tag{2.9}$$

$$c = |\chi u \bigcap \chi v| \qquad number of bits "on" in both Xu and Xv$$
(2.10)

$$d = n - |\chi_u \bigcup \chi_v| \quad \text{number of bits "off" in both Xu and Xv}$$
(2.11)

$$\therefore a+b-c = \left| \chi u \bigcup \chi v \right|$$
 (2.12)

Similarity  $S_{u,v}$  (Flower 1998; Willett, Barnard et al. 1998; Willett 2003) between two metabolites can be defined using Tanimoto Coefficient (Hubalek 1982) or Jaccard indexes (Jaccard 1912).

# Formula for continuous variables (2.13)

$$S_{u,v} = \frac{\left[\sum_{i=1}^{i=n} \chi u_{i} \chi v_{i}\right]}{\left[\sum_{i=1}^{i=n} (\chi u_{i})^{2} + \sum_{i=1}^{i=n} (\chi v_{i})^{2} - \sum_{i=1}^{i=n} \chi u_{i} \chi v_{i}\right]}, \text{ where } -0.333 \le S_{u,v} \le 1$$
(2.13)

Formula for dichotomous variables (2.14)

$$S_{u,v} = \frac{c}{a+b-c}$$
, Where  $0 \le S_{u,v} \le 1$  (dichotomous) (2.14)

**Set-theoretic definition** (2.15)

$$Su, v = \frac{|\chi u \cup \chi v|}{|\chi u \cap \chi v|}$$
(2.15)

#### 2.4.4.2. Generation of Scoring Function

**Definition 12:** The Percentage Atomic Mass Contribution (PAMC) (2.16) for two competing substrate (v) and product (u) can be defined as hundred times the sum of mass for both the metabolites (u and v) divided by the total mass of the metabolites in that reaction ( $V_i$ ).

$$PAMCu,v = 100 \times \begin{bmatrix} (Mu + Mv) \\ \sum Vi \end{bmatrix}$$
(2.16)

**Definition 13:** The final mapping scoring function for given source (u) and sink (v) vertices is defined as the product of similarity score Su, v and percentage atomic mass contribution PAMCu, v in subgraph  $G_i = (V_i, E_i, \mu_i, v_i)$  of graph G where  $V_i \subseteq V, E_i = E \cap (V_i \times V_i)$ , and  $\mu_i$  and  $v_i$  are restriction on  $\mu$  and v to  $V_i$  and  $E_i$ .

$$\mu_{i}(\mathbf{v}) = \begin{cases} \mu(\mathbf{v}) & \text{if } \mathbf{v} \in V_{i} \\ \text{unidefined} & \text{otherwise} \end{cases}$$
(2.17)  
$$\mathbf{v}_{i}(\mu) = \begin{cases} \mathbf{v}(\mu) & \text{if } \mu \in \mathring{A}_{i} \\ \text{unidefined} & \text{otherwise} \end{cases}$$
(2.18)

Hence we can derive equation (2.19) from equations (2.16) and (2.15)

Score 
$$u, v = Su_i, v_i \times PAMCu_i, v_i$$
  
Where  $0 \le Su, v \le 1$  and  $0 \le PAMCu_v \le 100$ 

$$(2.19)$$

#### 2.4.4.3. Mapping the Links using Weighted Matrix

The derived scoring function was used to find a suitable mapping between substrate molecules and product molecules. A slightly modified form of game theory (http://www.gametheory.net/) was used to map the substrate to the product metabolite. The method consists of construction of a matrix of substrates as rows and products as columns with the score defined above as matrix elements. The score between any substrate or product whose extension is smaller than three bonds is set to zero. A substrate is mapped to a product when either the score dominates all other scores in the present row or column respectively.

**Input:** Give a set of vertices substrate  $\mu_i$  and  $v_i$  products for reaction graph  $G_i$  and their score Score  $u_i, v_i$ 

Problem: Find the most compatible edge(s) between  $\mu_i$  and  $v_i$  based on the scoring function *Score* u,v where  $V_i \subseteq V$ ,  $E_i = E \cap (V_i \times V_i)$ , and  $\mu_i$  and  $v_i$  are restriction on  $\mu$  and v to  $V_i$  and  $E_i$ . Each source  $u \in V(G_i)$  and  $v \in V(G_i)$  sink is associated with reaction  $L(v) \subseteq L$  and each edge  $uv \in E(G)$  represents the reaction between u and v.

**Definition 14:** Let  $M_{m,n} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$  be a matrix which represents the

weighted edges for vertices  $\mu_i$  and  $\nu_i$ . The search for the most similar structures is conducted by maximizing the value of  $M_i$  in mutual dependency of  $M_j$ . In other words choose (vertices) entries in the matrix  $\omega(G_1, G_2)$  As shown in (Figure 14) metabolites (substrates and products) in one of the reactions in glycolysis was mapped according to the algorithm explained above. Hence by this process we can reduce the number of unwanted or less compatible links in each reaction *subgraph*  $G_i$ .



Figure 14. Metabolite mapping obtained from our algorithm shows that ATP maps to ADP (green line) and D-Glucose maps to D-Glucose-6phosphate (red line).

# 2.4.5. Improvised Algorithm for Finding Valid Links Between Substrate and Product

#### 2.4.5.1. Maximal Common Subgraph (MCS) Approach

Since molecular fingerprint is a heuristic way of representing molecules (as we can not represent the exact molecule as graph using fingerprints), it is a very fast and efficient way to skim through various structural properties embedded in molecules. On the other hand maximal common substructure (MCS) (Raymond and Willett 2002) can be a very effective way of finding maximum possible overlapping regions (isomorphic) between molecules. As we know graph comparison methods such as MCS (Raymond and Willett 2002) are fundamentally difficult because graph isomorphism problem is NP-complete (Akutsu 2004) and the computational time will increase exponentially (Kann 1992).

Hence we propose a model that combines the heuristic and exhaustive method to get a reasonably accurate and fast solution of graph comparison.

In order to map the common structure between substrate and product in each reaction, the maximum common subgraph (McGregor. 1982; Raymond, Gardiner et al. 2002; Hattori, Okuno et al. 2003) needs to be found/calculated. Hence the each reaction network can be defined as

**Definition 15:** Given graph  $G = (V, E, \mu, \nu)$ , subgraph  $G_i = (V_i, E_i, \mu_i, \nu_i)$  of G is a graph where  $V_i \subseteq V$ ,  $E_i = E \cap (V_i \times V_i)$ , and  $\mu_i$  and  $\nu_i$  are restriction on  $\mu$  and  $\nu$  to  $V_i$  and  $E_i$  as described in equations (2.17) and (2.18).

Finding the MCS in the graph can be defined in the following steps:

**Definition 16:** (Maximum Clique (MCL)) A clique G(Q) of graph G is defined as an induced subgraph of G on  $Q \subseteq V$ , i.e. if  $E(Q) = \{(v,w) \in E | v,w \in Q\}$ . A clique with the size of being maximum in G is called maximum clique. The number of vertices of a maximum cliques of G is expressed is expressed with  $\omega(G)$  (Cardinality of the clique),

**Definition 17:** (Maximal Common Subgraph (MCS)): Let graphs  $G_1 = (V_1, E_1, L)$  and  $G_2 = (V_2, E_2, L)$  be subgraph of graph G. A common subgraph S, i.e. subsets  $E_1^{'} \subseteq E_1$  and  $E_2^{'} \subseteq E_2$  of subgraphs  $G_1^{'} = (V_1, E_1^{'}, L)$  and  $G_2^{'} = (V_2, E_2^{'}, L)$  are isomorphic. Cardinality of the maximal common subgraph, i.e., |E'|.

In order to find a fast and optimized MCSA the problem was reduced to finding maximum clique in the association graph (AG) (Hattori, Okuno et al. 2003).

<sup>\*</sup> The MCS library was coded by Markus Leber, CUBIC, Koeln, Germany as a part of his PhD dissertation.