Analysis of Biochemical Networks Using Linear Programming

Evangelos Simeonidis[1], Lewis Dartnell[2], I. David L. Bogle[1,2] and Lazaros G. Papageorgiou[1,*]

[1]*Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), London, WC1E 7JE, UK*

[2]*CoMPLEX (Centre for Mathematics and Physics in the Life Sciences and Experimental Biology), UCL (University College London), London NW1 2HE, UK*

ABSTRACT

The application of mathematical programming methodologies to biochemical systems is demonstrated with the presentation of a linear programming (LP) algorithm for calculating minimal pathway distances in biochemical networks. Minimal pathway distances are identified as the smallest number of steps separating two nodes in the network. Two case studies are examined: 1) the minimal distances for *Escherichia coli* Small Molecule Metabolism (SMM) enzymes are calculated and their correlations with genome distance and enzyme function are considered; 2) a study of the p53 cell cycle and apoptosis control network is performed in order to assess the survivability of the network to both random node failures and a directed assault, by studying the modification of the network's diameter for successive protein knockouts. The results verify the applicability of the algorithm to problems of biochemical nature.

Keywords: *linear programming; shortest path algorithm; pathway distance; metabolic pathways; genome distance; p53 cell cycle and apoptosis control network; network robustness, tumour inducing viruses*

1. INTRODUCTION

The methodologies of mathematical programming and optimisation, developed and matured successfully within the Process Systems Engineering community, have not yet been transferred extensively to studies of biochemical nature. Biological information can form the basis for the development of quantitative computer-aided methods that will address problems of biology. Mathematical programming techniques have been used in the past in diverse biological studies, *e.g.* studies on metabolic networks (Regan *et al.*, 1993; Pramanik and Keasling, 1997; Edwards and Palsson, 2000; Burgard and Maranas, 2001), microarray analysis (Wolkenhauer, 2002) or protein structure (Backofen and Will, 2003; Klepeis and Floudas, 2003).

[*] To whom correspondence should be addressed: l.papageorgiou@ucl.ac.uk

Here, we demonstrate the application of linear programming (LP) to biological systems, with the presentation of an LP algorithm for calculating minimal distances in biochemical networks (Simeonidis *et al.*, 2003). Minimal distances are identified as the smallest number of steps separating two nodes in the network. Graph-oriented approaches have been employed before for the study of biological networks, such as metabolic pathways (Arita, 2000; Jeong *et al*., 2000; Fell and Wagner, 2000). The applicability of the studied algorithm is demonstrated with two case studies:

- To facilitate studies of evolution, the minimal distances for *Escherichia coli* Small Molecule Metabolism (SMM) enzymes are calculated and their correlations with genome distance (distance separating two genes on a chromosome) and enzyme function (as characterised by their Enzyme Commission (EC) number) are considered.
- Furthermore, a study of the p53 cell cycle and apoptosis control network is performed, which assesses the survivability of the network to both random node failures and a directed assault, by studying the alteration of the network's diameter (defined as the average of all pathway distances among all pairs of nodes in the network) for successive protein knockouts.

The rest of the paper is structured as follows: the mathematical programming formulation of an algorithm designed to calculate minimal pathway distances based on Linear Programming (LP) techniques is described. Then the model is applied to two case studies: first, the minimal pathway distances for the *E. coli* metabolism are calculated, and their correlations with genome distance and enzyme function are investigated. Second, the robustness of the p53 protein interaction network is studied. Finally, we discuss our conclusions for the LP method, and the biological implications of the results.

2. ALGORITHM

The shortest path problem consists of the identification of the shortest possible path from a source node of a network, to some other node in the network. Here, an LP model (Simeonidis *et al.*, 2003) applied to biological networks is suggested, capable of finding in a single pass the minimal distances (shortest path lengths) of all nodes in a network that are reachable from a source node (i^*). First, the notation used in the mathematical model is given:

Indices

i, j = nodes

Parameters

$L_{ij} = 1$ if there is an edge (link) from i to j; 0 otherwise

Positive continuous variables

D_i = distance from the i^* source node to node i

For each source node (i*) in the network, the algorithm finds the minimal distances to all other nodes by solving the following LP optimisation model:

$$\text{maximise} \sum_i D_i \tag{1}$$

subject to

$$D_j \le D_i + 1 \qquad \forall (i,j): L_{ij} = 1 \tag{2}$$

$$D_{i^*} = 0 \tag{3}$$

$$D_i \ge 0 \tag{4}$$

Constraints (2) incorporate network information related to connectivity, circularity and directionality, facilitated by the use of parameter L_{ij} (for two-way connections $L_{ij} = L_{ji} = 1$, however for one-way connections $L_{ij} = 1$ and $L_{ji} = 0$). Constraint (3) assigns the initial value of zero to node i^* to denote it as the source node, while constraints (4) require all D_i variables take positive values.

Finally, unbounded solutions can be avoided by adding:

$$D_i \le T \qquad \forall i \tag{5}$$

where T is an appropriately large number. It should be noted that if D_i equals T at the final solution then it can be concluded that there is *no* path connecting the i^* source node with node i in the network under consideration. This feature of the algorithm is particularly useful to identify cases where the connectivity of part of the network is missing.

The algorithm was implemented within the General Algebraic Modeling System (GAMS) software (Brooke *et al*., 1998), using the CPLEX 6.5 LP solver. All the computational experiments were performed on an IBM RS6000 workstation.

3. ILLUSTRATIVE EXAMPLES AND DISCUSSION

3.1. E. coli metabolism

In this work metabolism is considered as a single network. The SMM network used was obtained from the EcoCyc database (Karp *et al*., 2002). A protein-centric representation was adapted, *i.e.* the enzymes are considered as the nodes of the graph, and the substrates are the edges (Gerrard *et al.*, 2001). Genes encoding the investigated SMM enzymes were assigned a chromosomal location by consulting the Gene Table for *E. coli* (Blattner *et al.*, 1997). These were used to derive genome distances for gene pairs. Pairs were sorted into bins containing genes separated by less than 100bp, 101-1,000bp, 1,001-10,000bp, 10,001-100,000bp, 100,001-1,000,000 and more than 1,000,000bp.

Enzymes in the dataset were also assigned an EC number by reference to the GenProtEC database (Riley, 1998). EC numbers classify reactions within a hierarchical 4-level scheme (Enzyme Nomenclature, 1992). The number of matching EC levels (none, 1, 2, 3 or 4) is used as the functional similarity metric. The SMM dataset was composed of 599 enzyme pairs and 391 distinct metabolites. For 540 distinct enzymes a chromosomal localisation was identified, and 507 enzymes were assigned an EC number.

The objective of this case study is to draw evidence for the evolution of metabolism. Two main evolutionary models have been proposed: the patchwork model and the retrograde model (Rison and Thornton, 2002). The patchwork model proposes that metabolic pathways evolve by *ad hoc* recruitment of broad-specificity enzymes; this suggest that metabolically-close enzymes are no more likely to be functionally and evolutionarily similar than distant ones (Jensen, 1976). The retrograde model proposes that enzymes are recruited in a direction reverse to the metabolic "flow" from the preceding enzyme in the pathway; this suggests that nearby enzymes are likely to be evolutionarily related, and share some functionality (Horowitz, 1945).

3.1.1. Pathway distance and genome distance

First, the minimal pathway distances for all gene pairs in the SMM network were calculated. For the established pairs, the base pair separation of the genes encoding the enzymes in the *E. coli* genome was determined. The percentages of gene pairs were plotted against pathway distance in Figure 1.

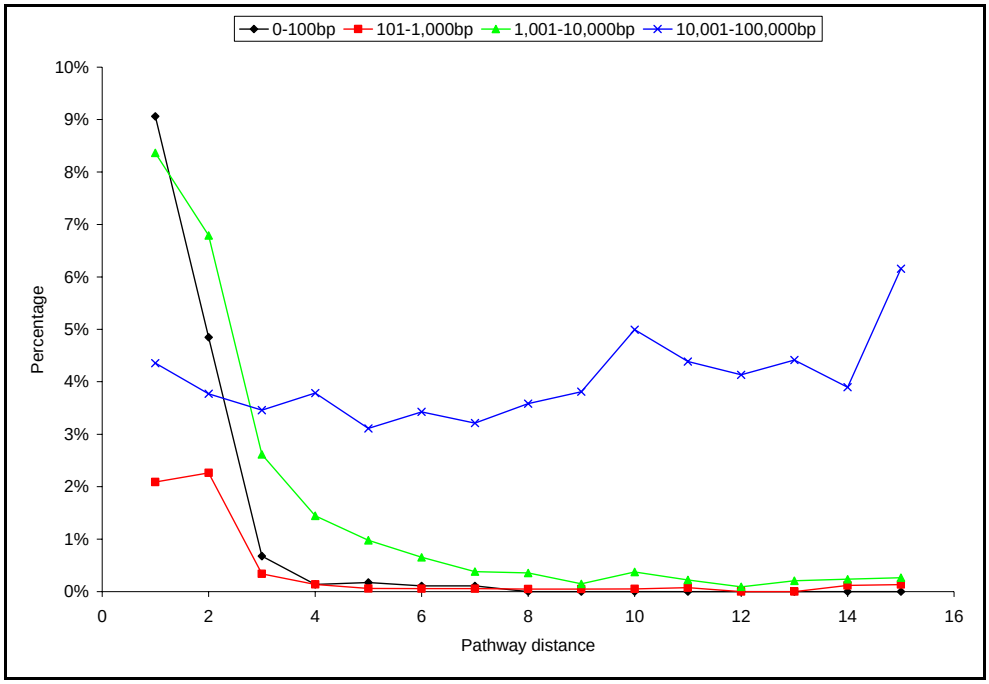

***Figure 1**: Pathway distance and genome distance. At each pathway distance (x-axis), the percentage of enzyme pairs within various genome distance bins is plotted.*

The correlation between pathway distance and genome distance is clear. For the first three distance bins, as pathway distance increases, the percentage of genes separated by short genome distances drops. For distances of 10,001 to 100,000 base pairs there is no clear trend observed (the same is true for 100,001bp-1,000,000bp and 1,000,001bp and above, but these bins are not plotted in Figure 1). Figure 1 indicates that SMM genes are metabolically clustered on the genome. The relatively high percentage of metabolic-gene pairs found within 100bp, which is a very short distance in a 4.6Mbp long chromosome, suggests that this clustering is the consequence of prokaryotic operon structures in which co-regulated genes are rarely separated by longer distances (Salgado *et al.*, 2000). This observation has been made before (Tamames *et al.*, 1997; Overbeek *et al.*, 1999; Rison *et al.*, 2002). Here, we show that it holds true using co-participation in a metabolic pathway as an indication of shared function and measuring this relationship with our pathway distance metrics.

3.1.2. Pathway distance and function similarity

The EC numbers assigned to each enzyme were compared, and the level of EC number conservation was determined. The results are plotted in Figure 2.

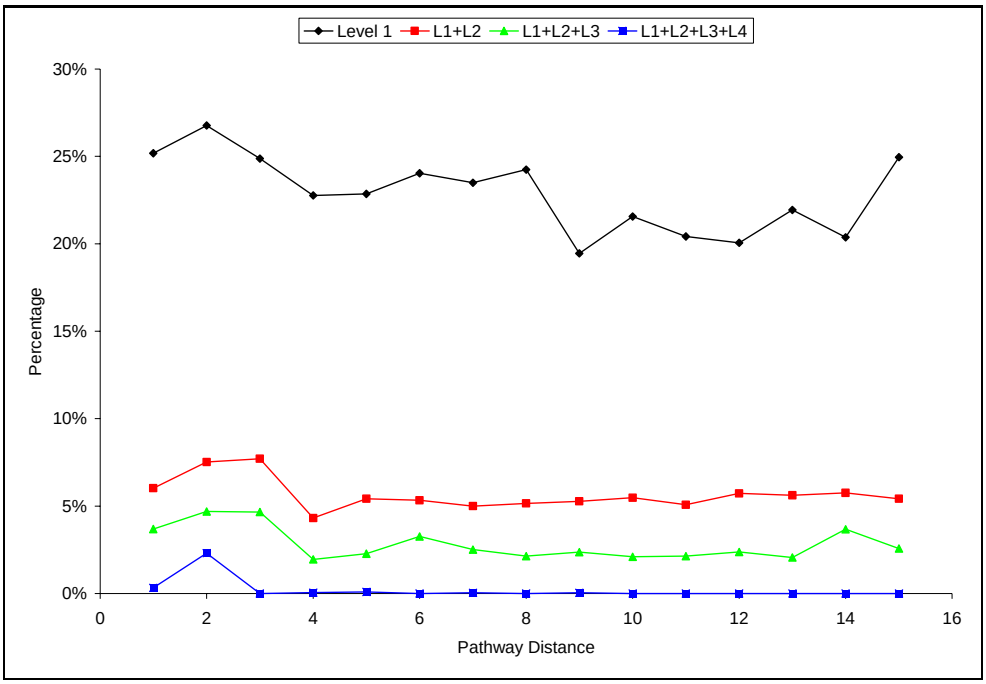

FIG. 2*: Pathway distance and function similarity. At each pathway distance, the percentage of enzyme pairs with all (L1+L2+L3+L4), 3 or more (L1+L2+L3), 2 or more (L1+L2) or 1 or more (Level1) matching EC levels is plotted. The L1+L2+L3+L4 is a subset of the L1+L2+L3 set (which in turn is a subset of L1+L2, etc.).*

No obvious correlation between EC number and pathway distance can be observed. Furthermore, the data suggests that conservation of EC number is relatively rare at all distances, as the percentage of enzyme pairs with at least two EC levels is always under 8%. Even at short pathway distances, enzyme pairs only catalyse the same type of reaction approximately once out of 4 cases. This percentage remains relatively constant over most distances, suggesting no particular bias for EC

number conservation at shorter distances. It is known that the relationship between EC numbers and pathways is complex, with pathways requiring a number of enzyme types to perform their task (Tsoka and Ouzounis, 2001). These data would suggest that enzymatic chemistries are varied along the substrate conversion routes.

3.2. Robustness of the p53 protein interaction network

In multicellular organisms (metazoans) cellular proliferation is tightly regulated. Cells proliferate and accumulate in a co-ordinated way during growth or repair, and undergo programmed cell death (apoptosis) when genetically-damaged, virally-infected or the developmental program requires it. In response to stress a metazoan cell must decide between continued (or resumed) progression through the cell cycle or initiation of apoptosis. This decision is mediated by a protein-interaction network, at the centre of which lies the p53 protein. p53 is found only in metazoan cells, and combines protein interaction domains, regulatory domains and a sequence specific DNA recognition domain that allow the integration of intra- and intercellular signals with gene transcription (Kohn, 1999). Failure of the apoptotic control system, leading to either unregulated proliferation or unnecessary apoptosis, is causative of both tumorigenesis and developmental diseases. The importance of the p53 response network in the prevention of cancer is striking, and mutations reducing p53 activity are present in over 50% of human tumours (Haupt *et al.*, 2003).

3.2.1. Model of the p53 network

Data relating to the interconnections of the known component proteins of the p53 network (and some non-peptide molecules such as dsDNA) were extracted from the molecular interaction map presented in Kohn (1999), in order to create a computer model of the network containing 104 nodes and 226 bidirectional connections. A log plot of the connectivity (total number of a node's connections) against the distribution shows a power-law relationship - the defining feature of a 'scale-free' network architecture (Barabási *et al.*, 2004). The network has no characteristic degree of connectivity: the vast majority of nodes have only a few connections, but there are several hubs that are very highly connected. In recent years, a great number of organic networks have been shown to be scale-free, including the Internet (Hawoong *et al.*, 1999), social interactions (Albert *et al.*, 2000), neural networks (Strogatz, 2001), ecological food webs (Strogatz, 2001), metabolism (Jeong *et al.*, 2000), protein-protein interactions (Jeong *et al.*, 2001), and gene transcription regulation networks (Barabási *et al.*, 2004).

The LP algorithm presented in section 2 was used to calculate the average path length (APL) of each protein - that is, the mean of the shortest paths to all other nodes. The APL provides an informative metric of a node's centrality within the network and has been calculated for substrates

in *E. coli* core metabolism (Fell and Wagner, 2000). The ten most central proteins - the hubs - in the p53 network were calculated as: p53 (APL=1.92), CDK2 (2.09), Cyclin A (2.20), CDK1 (2.29), MDM2 (2.29), DP (2.34), pRb (2.35), PCNA (2.36), and RPA (2.38), with the least crucial protein being DNA ligase III (4.55). The diameter of a network is defined as the mean of all path lengths, and this measure of navigability is used as a proxy for the functional health of the network (Albert *et al.,* 1999).

3.2.2. Network model robustness

Network robustness is analysed by cumulatively knocking out nodes (*i.e.* removing all their connections) and studying the increase in diameter as the network degenerates. Nodes can be knocked out in one of two attack modes: randomly, or in a directed attack by preferentially targeting the hubs. It has been shown that a scale-free network is relatively immune to random node failure, but extremely vulnerable to a targeted onslaught (Barabási *et al.,* 2004).

Here, the survivability of the p53 network in the face of both a directed and random attack against its nodes is examined. Proteins are progressively knocked-out in a specified order, either random permutation or rank order of APL, with the diameter recalculated with the LP algorithm at each step. The random attack is repeated 100 times, and the diameter at each step averaged across all runs. If a knockout isolates a node from the rest of the network the path lengths are set equal to the arbitrarily large number *T*, given the value of 100 in this study. Figure 3 shows the plot of network diameter over the first 30 knockouts.

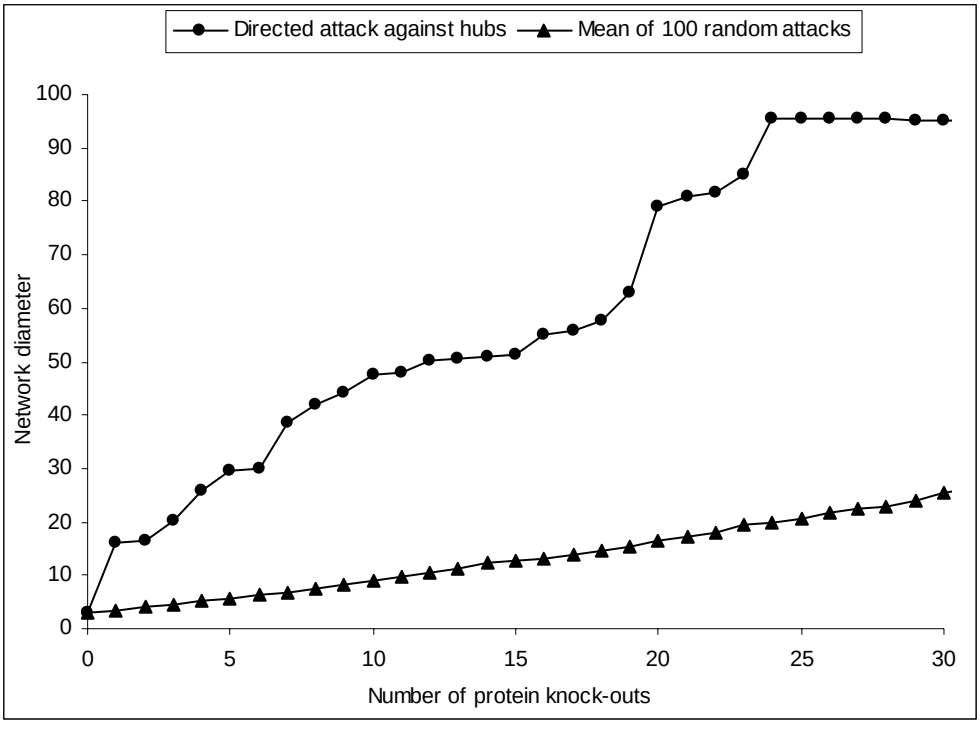

FIG. 3*: Degeneration of p53 network diameter when nodes are knocked out in either a random pattern, or in a directed attack against the hubs.*

Diameter deteriorates very slowly under a random attack - the architecture of the p53 network provides an inherent robustness against such a scheme. Hub nodes are uncommon so they are rarely hit and most of the protein knockouts have negligible impact on the global integrity of the network. This reliance on highly-connected nodes, however, renders the network vulnerable to a directed attack. Network communication fails rapidly under the onslaught (loss of only p53 results in a five-fold increase in diameter, from 3.1 to 16.1). A similar result was obtained on simulated attacks on the Internet, which was found to be robust to random server failures, but vulnerable to the activities of hackers deliberately targeting the hubs so as to wreak maximal havoc (Albert *et al.*, 2000).

3.2.3. Biological Hackers

Mutational damage to biological networks is essentially random, but in fact there exists a targeted threat against the p53 network, operating not at a genetic level but against the translated proteins. DNA tumour-inducing viruses (TIVs) increase their replication rate and survival with an armoury of proteins that suppress the normal apoptotic infection response and short-circuit the cell cycle into continually synthesizing viral DNA. TIVs, including adenovirus, human cytomegalovirus, human papillomavirus, and simian virus 40, all selectively inhibit similar proteins in the infected host cell (Levine, 1992; Burgert *et al.*, 2002; Banks *et al.*, 2003), causing an average diameter increase to about 23. It is no coincidence that the most common targets, pRb and p53, are also two of the most central hubs. The TIV directed strikes are very effective at disrupting communication within the p53 network, but do not increase the diameter so much that the network shatters and function fails completely. The tumour inducing viruses thus behave like biological hackers – targeting their attack against some of the p53 network hubs and so exploiting the inherent weakness of its architecture.

4. CONCLUDING REMARKS

The applicability of mathematical programming techniques in the analysis of biochemical networks has been demonstrated with the presentation of a fast and effective LP algorithm characterised by its ability to deal efficiently with network circularity and bidirectionality. Despite its simplicity, the LP algorithm exhibits a first step towards building optimisation-based tools for studying biological networks, which are characterised by their complex and dynamic nature.

The algorithm has been applied in the study of the correlations between minimal pathway distances of the *E. coli* SMM enzymes and genome distance, and between *E. coli* minimal pathway distances and enzyme function. As expected, genes encoding enzymes involved in nearby metabolic reactions were more likely to be in close proximity on the genome. However, pathway distances did not correlate with enzyme function (as described by assigning EC numbers to SMM enzymes). These data, in conjunction with the result of previous analyses incorporating work concerning sequence

and structural similarity of SMM enzymes (Teichmann *et al*., 2001; Rison *et al*. 2002), suggest a patchwork model of pathway evolution: the lack of obvious correlation between pathway distance and EC numbers is consistent with the *ad hoc* recruitment of enzymes where required within the metabolism of an organism.

The robustness of the p53 protein interaction network has also been investigated. A non-weighted, bidirectional model was used, which represents a first step towards building a biologically realistic representation of the p53 network. The network is proven robust to random knockouts of its proteins, which signifies resilience against mutational perturbation. However, the reliance on highly-connected nodes makes the network vulnerable to the loss of its hubs. Evolution has produced organisms that exploit this very weakness in order to disrupt the p53 network for their own ends: tumour inducing viruses target specific proteins, and this study has identified these same proteins as the network hubs.

REFERENCES

Albert, R., Jeong, H. and Barabási, A.L. (1999). The diameter of the World Wide Web. *Nature*, **401,** 130-131.

Albert, R., Jeong, H. and Barabási, A.L. (2000). Error and attack tolerance of complex networks. *Nature*, **406** (6794), 378-382.

Arita, M. (2000). Metabolic reconstruction using shortest paths. *Simulat. Pract. and Theory*, **8**, 109-125.

Backofen, R. and Will, S. (2003). A constraint-based approach to structure prediction for simplified protein models that outperforms other existing methods. *Lect. Notes Comput. Sc.* **2916**, 49-71.

Banks, L., Pim, D. and Thomas, M. (2003). Viruses and the 26S proteasome: hacking into destruction. *Trends Biochem. Sci.*, **28** (8), 452-459.

Barabási, A.L. and Oltvai, Z.N. (2004). Network Biology: Understanding the cell's functional organisation. *Nat. Rev. Genet.*, **5** (2), 101-113.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277** (5331), 1453-1474.

Brooke, A., Kendrick, D., Meeraus, A., and Raman, R. (1998). "GAMS: A User's Guide," GAMS Development Corporation, Washington.

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

Burgert, H.G., Ruzsics, Z., Obermeier, S., Hilgendorf, A., Windheim, M. and Elsing, A. (2002). Subversion of host defence mechanisms by adenoviruses. *Curr. Top. Microbiol. Immunol.*, **269**, 273-318.

Casjens, S. (1998). The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.*, **32**, 339-77.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein C. (2001). "Introduction to algorithms," The MIT Press, Cambridge, Massachusetts.

Edwards, J.S., and Palsson, B.O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, **97** (10), 5528-5533.

Enzyme Nomenclature (1992). "Recommendations of the Nomenclature Committee of the International Union of Biochemistry (NC-IUB)," Academic Press, San Diego, CA.

Fell, D.A., and Wagner, A. (2000). The small world of metabolism. *Nature Biotechnol.*, **18** (11), 1121-1122.

Gerrard, J.A., Sparrow, A.D., and Wells, J.A. (2001). Metabolic databases – what next? *Trends. Biochem. Sci.*, **26** (2), 137-140.

Haupt, S., Berger, M., Goldberg, Z. and Haupt, Y. (2003). Apoptosis - the p53 network, *J. Cell Sci*., **116**, 4077-4085.

Horowitz, N. H. (1945). On the Evolution of Biochemical Syntheses. *Pro. Natl. Acad. Sci.*, **31**, 153-157.

Jensen, R. A. (1976). Enzyme recruitment in the evolution of new function. *Annu. Rev. Microbiol.*, **30**, 409-425.

Jeong, H., Mason, S.P., Barabási, A.L. and Oltvai, Z.N. (2001). Lethality and Centrality in protein networks. *Nature*, **411** (6833), 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.L. (2000). The large-scale organization of metabolic networks. *Nature*, **407** (6804), 651-654.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. (2002). The EcoCyc database. *Nucleic Acids Res.*, **30** (1), 56-58.

Klepeis, J.L. and Floudas, C.A. (2003). ASTRO-FOLD: a combinatorial and global optimization framework for *ab initio* prediction of three-dimensional structures of proteins from the amino acid sequence, *J. Biophys. J.*, **85** (4), 2110-2146.

Kohn, K.W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, **10** (8), 2703-2734.

Lawler, E.L. (1976). "Combinatorial Optimization: Networks and Matroids," Holt, Rinehart and Winston, New York, NY.

Levine, A.J. (1992). The DNA Tumor Viruses. In "Viruses". New York, Scientific American Library, 87-111.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, **96** (6), 2896-2901.

Pramanik, J., and Keasling, J.D. (1997). Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependant biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.*, **56** (4), 398-421.

Regan, L., Bogle, I.D.L., and Dunnill, P. (1993). Simulation and optimization of metabolic pathways. *Computers & Chemical Engineering*, **17** (5-6), 627-637.

Riley, M (1998). Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.*, **26** (1), 54.

Rison, S.C.G., Teichmann, S.A., and Thornton, J.M. (2002). Homology, Pathway distance and Chromosomal localisation of Small Molecule Metabolism enzymes in *Escherichia coli*. *J. Mol. Biol.*, **318** (3), 911-932.

Rison, S.C.G., and Thornton, J.M. (2002). Pathway Evolution, Structurally Speaking. *Curr. Opin. Struct. Biol.*, **12**, 374-382.

Salgado, H., Moreno-Hagelsieb, G. Smith, T.F., and Collado-Vides, J. (2000). Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA*, **97** (12), 6652-6657.

Simeonidis, E., Rison, S.C.G., Thornton, J.M., Bogle, I.D.L. and Papageorgiou L.G. (2003). Analysis of metabolic networks using a pathway distance metric through linear programming. *Metab. Eng.* **5** (3), 211-219.

Strogatz, S.H. (2001). Exploring complex networks. *Nature*, **410** (6825), 268-276.

Tamames, J., Casari, G., Ouzounis, C.A., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44** (1), 66-73.

Teichmann, S.A., Rison, S.C.G., Thornton, J.M., Riley, M., Gough, J., and Chotia, C. (2001). The evolution and structural anatomy of the small molecule metabolic pathways of *Escherichia coli*. *J. Mol. Biol.*, **311** (4), 693-708.

Tsoka, S., and Ouzounis, C.A. (2001). Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. *Genome Res.*, **11** (9), 1503-1510.

Wolkenhauer, O. (2002). Mathematical modelling in the post-genome era: Understanding genome expression and regulation - A system theoretic approach, *Bio Systems*, **65** (1), 1-18.