Available online at www.sciencedirect.com
SCIENCE DIRECT®
ACADEMIC PRESS
Metabolic Engineering 5 (2003) 211–219
METABOLIC ENGINEERING
http://www.elsevier.com/locate/ymben

Analysis of metabolic networks using a pathway distance metric through linear programming

Evangelos Simeonidis,[a,1] Stuart C.G. Rison,[b,1,2] Janet M. Thornton,[b,c,3] I. David L. Bogle,[a] and Lazaros G. Papageorgiou[a,*]

[a] *Department of Chemical Engineering, Centre for Process Systems Engineering, UCL (University College London), London, WC1E 7JE, UK*
[b] *Department of Biochemistry and Molecular Biology, UCL (University College London), London, WC1E 6BT, UK*
[c] *Department of Crystallography, Birkbeck College, London WC1E 7HX, UK*

Received 27 December 2002; accepted 11 June 2003

Abstract

The solution of the shortest path problem in biochemical systems constitutes an important step for studies of their evolution. In this paper, a linear programming (LP) algorithm for calculating minimal pathway distances in metabolic networks is studied. Minimal pathway distances are identified as the smallest number of metabolic steps separating two enzymes in metabolic pathways. The algorithm deals effectively with circularity and reaction directionality. The applicability of the algorithm is illustrated by calculating the minimal pathway distances for *Escherichia coli* small molecule metabolism enzymes, and then considering their correlations with genome distance (distance separating two genes on a chromosome) and enzyme function (as characterised by enzyme commission number). The results illustrate the effectiveness of the LP model. In addition, the data confirm that propinquity of genes on the genome implies similarity in function (as determined by co-involvement in the same region of the metabolic network), but suggest that no correlation exists between pathway distance and enzyme function. These findings offer insight into the probable mechanism of pathway evolution.

Keywords: Pathway distance; Shortest path algorithm; Linear programming; Metabolic pathways; Genome distance; Enzyme function

1. Introduction

Metabolism is a complex network of enzymes, substrates and co-factors. For some model organisms, such as *Escherichia coli*, these networks are well characterised (Karp et al., 2002), making them ideal specimens for the study of metabolic systems.

Much work has already been done on modelling metabolism (Edwards and Palsson, 2000a) and analysing the possible mechanisms of pathway evolution (Teichmann et al., 2001; Rison et al., 2002; Rison and Thornton, 2002). The wealth of currently available data can be used in the creation of models that may also be applied for the simulation and optimisation of biochemical systems. Optimisation techniques have already been used in studies to meet objectives such as flux maximisation, optimal growth and studying the effect of gene deletions or additions to network robustness (Varma and Palsson, 1993; Regan et al., 1993; Pramanik and Keasling, 1997; Schilling et al., 1999; Edwards and Palsson, 2000b; Burgard and Maranas, 2001).

Lately, there has been an increasing interest in metabolic pathways as an indicator of "connectivity" between genes (Marcotte et al., 1999; Kolesov et al., 2001; Rison et al., 2002). The pathway distance metric can serve as such a measured descriptor of the relationship between two enzymes in the metabolic network. Minimal pathway distances are identified as the smallest number of metabolic steps separating two enzymes: the shortest path from one point in the network to another.

Metrics based on the application of shortest path algorithms in biochemical systems have been considered

*Corresponding author. Tel.: +44-20-7679-2563; fax: +44-20-7383-2348.
E-mail address: l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou).
[1] Equally contributed to this work.
[2] Current address: Royal Veterinary College, Pathology and Infectious Diseases, Royal College Street, London NW1 0TU, UK.
[3] Current address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 ISD, UK.

1096-7176/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1096-7176(03)00043-0

before. Graph-oriented representations of metabolism have been used to reconstruct metabolic pathways (Arita, 2000). The large-scale organisation of cellular networks has been addressed with a systematic comparative mathematical analysis based on a shortest path algorithm that examines the properties of the metabolic networks of different organisms (Jeong et al., 2000). A quantitative basis for identifying a set of central metabolites defining the core of metabolism by calculating the shortest distances between substrates has also been established (Fell and Wagner, 2000).

There are two main models for the evolution of metabolic pathways: 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 (capable of catalysing a variety of metabolic reactions); this suggests that metabolically close enzymes are no more likely to be functionally and evolutionarily similar to the 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).

Recently, the biochemical properties of the *E. coli* small molecule metabolism (SMM) genes and enzymes were investigated using a simple but inefficient graph depth-first-traversal algorithm (Rison et al., 2002). The work demonstrated that propinquity of SMM genes on the *E. coli* chromosome was matched by propinquity of the encoded proteins in the metabolic network. Patterns of enzyme homologies and conservation of catalytic chemistry between homologues were suggestive of a patchwork model of pathway evolution, as opposed to the retrograde model of evolution (Rison et al., 2002; Rison and Thornton, 2002). A network approach was also used to study the evolution of enzymes in metabolism (Alves et al., 2002). Interestingly, the authors find that neighbouring enzymes (less than 3 steps apart) in the reaction network are more likely to be homologous than distant enzymes (more than 3 steps apart). The work also suggests that blocks of similar catalysis have evolved in metabolism.

The paper is structured as follows. First, the generation of the pathway dataset is discussed. The mathematical programming formulation of an algorithm designed to calculate minimal pathway distances based on linear programming (LP) techniques (Lawler, 1976; Cormen et al., 2001) is then described. The model is applied to the *E. coli* metabolism, and the correlations of minimal pathway distance with genome distance (i.e., the number of base pairs separating two SMM genes on the *E. coli* chromosome), and enzyme function (as described by Enzyme Commission (EC) number (Enzyme Nomenclature, 1992)) are investigated. Both the LP method itself, and the biological implications of the analysis results are then discussed.

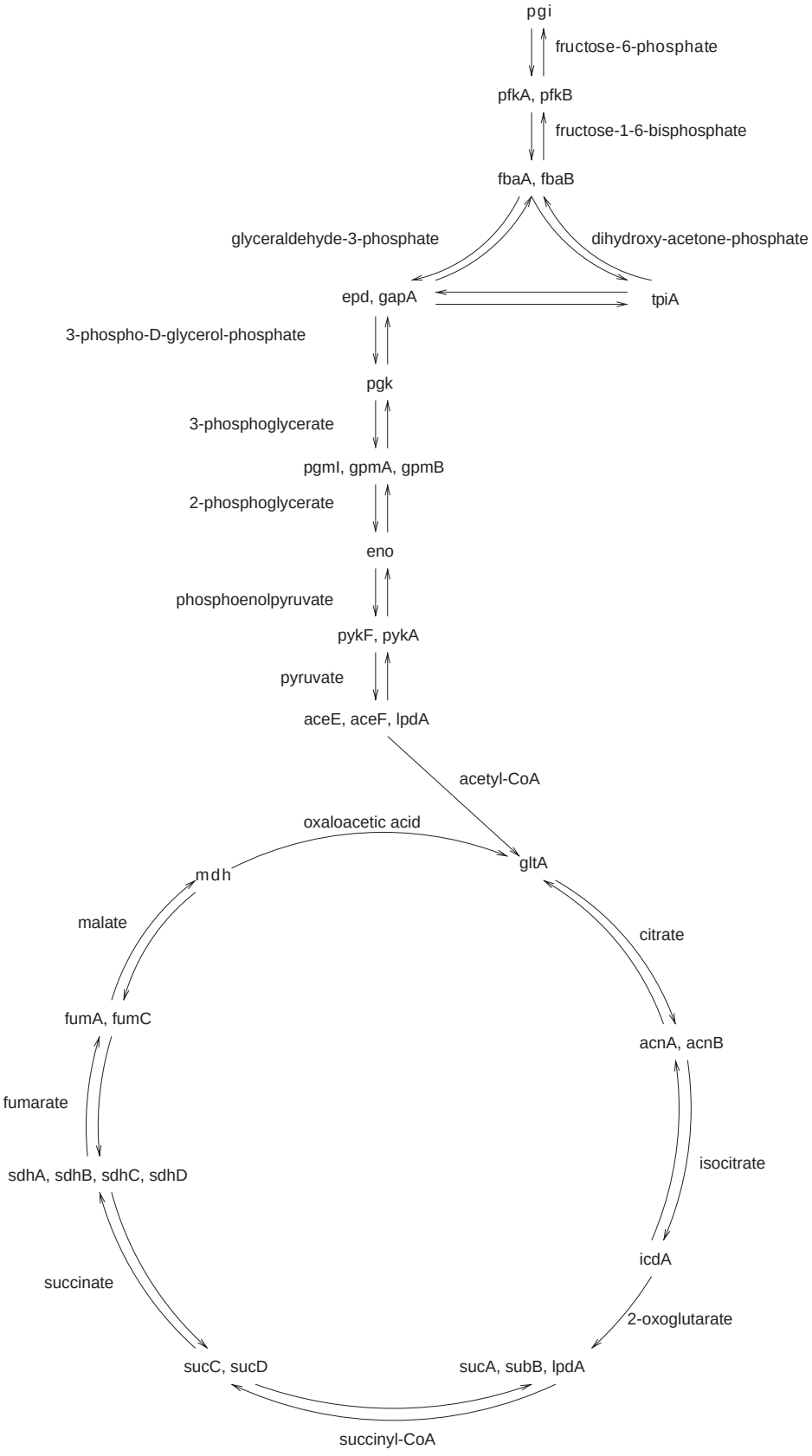
2. Methods

2.1. Generating the pathway dataset

Often, the metabolic network is subdivided into individual pathways, as commonly depicted in biochemistry textbooks (e.g., Glycolysis, TCA, fatty-acid biosynthesis) (Voet and Voet, 1995). However, whilst each individual pathway can be considered a separate entity, and distinction can be made between inter- and intra-pathway properties (Teichmann et al., 2001), metabolism is a complex and complete network. Thus, the division of metabolism into distinct pathways is arbitrary (Gerrard et al., 2001). A possible way to deal with this issue is to ignore these divisions, and instead consider metabolism as a single network. Herein, such a network approach, similar to that of Alves et al. (2002), was adopted. When individual pathways are mentioned in the text, this is done in order to simplify the discussion; the analyses presented were performed on the whole network, not on a "per pathway" basis.

The SMM network used was obtained from the EcoCyc database (Karp et al., 2002). Even though metabolite-centric representations of metabolic networks are the most common (Michal, 1998), in this work a protein-centric representation was adapted instead. As illustrated in Fig. 1, the enzymes are considered as the nodes of the graph, and the substrates are the edges (Gerrard et al., 2001).

Fig. 1. A protein-centric (Gerrard et al., 2001) view of glycolysis and the tricarboxylic acid (TCA) cycle (adapted from EcoCyc; http://www.ecocyc.org/). Enzymes are the nodes, substrates label the edges, and only key metabolites are shown. The arrows can be read as "produces a substrate for" (full discussion in text). pgi: phosphoglucose isomerase; pfkA and pfkB: 6-phosphofructokinase-1 and 2; fbaB and fbaA: fructose bisphosphate aldolase class I and II; tpiA: triose phosphate isomerase; epd: glyceraldehyde-3-phosphate dehydrogenase 2; gapA: glyceraldehyde-3-phosphate dehydrogenase-A; pgk: phosphoglycerate kinase; gpmA and gpmB: phosphoglycerate mutase 1 and 2; pgmI: phosphoglycerate mutase, co-factor independent; eno: enolase; pykF and pykA: pyruvate kinase I and II; aceE, aceF and lpdA: pyruvate dehydrogenase multienzyme complex; gltA: citrate synthase; acnA and acnB: aconitase A and B; icdA: isocitrate dehydrogenase; subA, sucB and lpdA: 2-oxoglutarate dehydrogenase complex; sucC and sucD: succinyl-CoA synthase complex; sdhA, sdhB, sdhC and sdhD: succinate dehydrogenase complex; fumA and fumC: fumarase A and fumarase C; mdh: malate dehydrogenase.


pgi
fructose-6-phosphate
pfkA, pfkB
fructose-1-6-bisphosphate
fbaA, fbaB
glyceraldehyde-3-phosphate
dihydroxy-acetone-phosphate
epd, gapA
tpiA
3-phospho-D-glycerol-phosphate
pgk
3-phosphoglycerate
pgmI, gpmA, gpmB
2-phosphoglycerate
eno
phosphoenolpyruvate
pykF, pykA
pyruvate
aceE, aceF, lpdA
acetyl-CoA
oxaloacetic acid
gltA
mdh
malate
citrate
fumA, fumC
acnA, acnB
fumarate
isocitrate
sdhA, sdhB, sdhC, sdhD
icdA
succinate
2-oxoglutarate
sucC, sucD
sucA, subB, lpdA
succinyl-CoA

In Fig. 1, enolase (the gene product of *eno*) produces substrate "*phosphoenolpyruvate*" for PykF and PykA. Likewise, PykF and PykA produce "*phosphoenolpyruvate*" for Eno when catalysing the reverse direction reaction. Malate dehydrogenase (the gene product of *mdh*) produces substrate "*oxaloacetic acid*" for GltA, but GltA does not produce "*oxaloacetic acid*" for Mdh. The minimal pathway distance from GltA to Mdh is therefore 1 if directionality is not taken into account (all edges are assumed to be bi-directional), but 7 if directionality is considered (clockwise around the tricarboxylic acid (TCA) cycle).

2.2. Genome distance

Genes encoding the SMM enzymes investigated were assigned a chromosomal location by consulting the Gene Table for *E. coli* (http://www.genome.wisc.edu/pub/analysis/m52orfs.txt; Blattner et al., 1997). These were used to derive genome distances for gene pairs, i.e., the smallest distance in base pairs (bp) separating the two genes on the chromosome. Since the *E. coli* chromosome is ~4.6 Mbp and only the smallest genome distance is considered, two genes can, at most, be separated by ~2.3 Mbp. In this paper, pairs are sorted into bins containing genes separated by: less than 100, 101–1000, 1001–10,000, 10 001–100,000, 100,001–1,000,000 and more than 1,000,000 bp. The choice of bin sizes has a biological rationale. The first of these bins accounts for genes likely to belong to the same operon (Salgado et al., 2000), the second bin size approximates to the average size of a prokaryotic gene (Casjens, 1998). Subsequent bins were simply enlarged by an order of magnitude.

2.3. Function similarity

Enzymes in the dataset were assigned an EC number by reference to the GenProtEC database (Riley, 1998), and following communications from the database curators (Monica Riley and Margrethe Serres, pers. comm.). EC numbers classify reactions within a hierarchical four-level scheme (e.g., the reaction catalysed by the enzyme *glyceraldehyde-3-phosphate dehydrogenase* has EC number 1.2.1.12) (Enzyme Nomenclature, 1992). The level to which EC numbers assigned to two enzymes are identical can therefore be used as a measure of the similarity of the function they perform (Martin et al., 1998; Todd et al., 2001). Enzymes assigned identical EC numbers perform the same biochemical function, enzymes with only the first EC level in common share only very generalised functional similarity (e.g., both *oxidoreductases*). Finally, enzymes assigned completely different EC numbers often share little or no functional commonalities. Therefore, in this paper, the number of matching EC levels (none, 1, 2, 3 or 4) is used as the functional similarity metric.

3. Algorithm

Linear programming is an extensively used optimisation technique, ranked as a significant scientific advance of the mid-20th century. The numerous applications involve the allocation of limited resources to competing activities in the optimal way (Williams, 1999; Cormen et al., 2001). These types of problems arise in varying situations, ranging from graphs and network flows to plant management (e.g., manufacturing and transportation of goods) and economics. The most prominent method for solving LP problems is the simplex method (Dantzig, 1963).

The recognition of the shortest possible directed path from a specified source node to some other node of a weighted, directed graph is known as a shortest path problem. A variety of combinatorial problems can be formulated and solved as shortest path problems. In addition, a number of more complex problems can be solved by procedures, which call upon shortest path algorithms (Lawler, 1976).

An LP model (Lawler, 1976; Cormen et al., 2001) applied to metabolic networks is suggested, capable of finding in a single pass the minimal pathway distances (shortest path lengths) of all enzymes in a network that are reachable from a source enzyme (i^*). First, the notation used in the mathematical model is given:

Indices: i,j = enzymes.

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 enzyme to enzyme i.

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

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

subject to

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

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

$$D_i \geqslant 0. \tag{4}$$

Constraints (2) incorporate pathway information related to reaction connectivity, circularity and reaction directionality, facilitated by the use of parameter L_{ij} (for reversible reactions $L_{ij}=L_{ji}=1$; however, for irreversible reactions $L_{ij}=1$ and $L_{ji}=0$). Constraint (3) assigns the initial value of zero to enzyme i^* to denote it as the source enzyme, while constraint (4) requires all D_i variables take positive values.

Finally, unbounded solutions can be avoided by adding

$$D_i \leqslant T \quad \forall i, \tag{5}$$

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 enzyme with enzyme 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 (refinement of the basic simplex method; Dantzig, 1963) for solving LP problems such as the one in hand. Finally, post-processing calculations were incorporated in the algorithm to derive correlations of minimal pathway distance with genome distance and function similarity.

4. Results and discussion

4.1. SMM dataset

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 dataset was kindly provided by the curators of the EcoCyc database (Karp et al., 2002). Pathway distances obtained by the solution of the algorithm ranged from 1 to 26. After a certain pathway distance, the results cease to be informative because: (i) they do not deviate substantially from that found at the previous pathway distance and/or (ii) they are based on such a small number of pairs that their validity is questionable. Therefore, in all plots, only pathway distances up to 15 are considered.

4.2. Pathway distance and genome distance

The minimal pathway distances for all gene pairs in the SMM network were calculated (Table 1).

For the established pairs, the bp separation of the genes encoding the enzymes in the *E. coli* genome was determined. For example, the enzymes glyceraldehyde-3-phosphate dehydrogenase 2 and phosphoglycerate kinase (respectively epd and pgk in Fig. 1) have a pathway distance of 1, and are encoded by genes separated by only 50 bp. The pair therefore falls into the first pathway distance bin. However, the enzymes phosphoglycerate kinase and phosphoglycerate mutase 1 (respectively pgk and gpmA in Fig. 1), which also have a pathway distance of 1, are encoded by genes separated by 2,282,661 bp. The percentages of gene pairs in the first four genome distance bins are plotted against pathway distance in Fig. 2.

There is a clear correlation between pathway distance and genome distance. As pathway distance increases, the percentage of genes separated by short genome distances drops. For pathway distances of 1, 2, 3, and 4 steps, gene pairs separated by at most 10,000 bp (i.e., bins 0–100, 101–1000, and 1001–10,000 bp) account for 19.51%, 13.9%, 3.63% and 1.72%, respectively, of the pairs analysed (Fig. 2). For the other three distance bins (101,000–1,000,000 and 1,000,001 bp and above, which are not plotted here), no clear trend is evident.

A statistical measure is applied to demonstrate that the results of the analysis are not due to chance. We are using the standard normal deviate, or Z-score, which measures the distance of a value from the mean of a distribution in standard deviation units. For the needs of this analysis, the mean and standard deviation used are those of randomised networks. Fig. 3 presents the Z-score results calculated for the SMM network.

Random interconnected networks were created by arbitrarily pairing the enzymes of the *E. coli* SMM, making sure that the same number of pairs was created for each distance as for the original *E. coli* network (i.e.,

Table 1
Number of gene pairs in the six genome distance bins for each pathway distance

Genome distance bins (bp)	Pathway distance														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0–100	52	45	8	2	3	2	2	0	0	0	0	0	0	0	0
101–1000	12	21	4	2	1	1	1	1	1	1	1	0	0	1	1
1001–10,000	48	63	31	21	17	12	7	7	3	7	3	1	2	2	2
10,001–100,000	25	35	41	55	54	63	59	71	78	93	59	45	43	33	46
100,001–1,000,000	174	311	463	557	645	679	705	777	856	701	516	457	379	304	274
1,000,001–10,000,000	263	453	638	816	1015	1081	1062	1125	1108	1061	766	586	550	507	424
Total	574	928	1185	1453	1735	1838	1836	1981	2046	1863	1345	1089	974	847	747

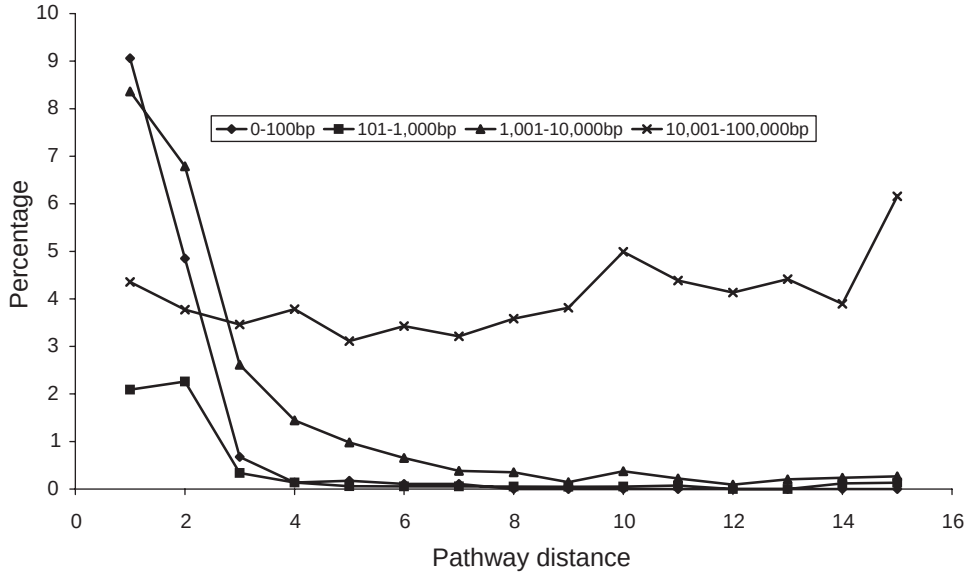


Fig. 2. Pathway distance and genome distance. At each pathway distance (x-axis), the percentage of enzyme pairs within various genome distance bins is plotted.

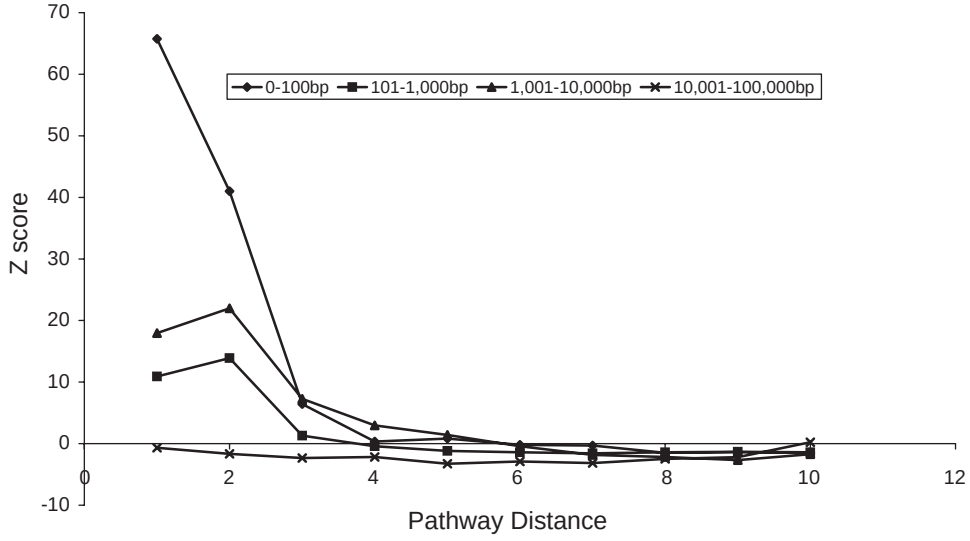


Fig. 3. Z-score. At each pathway distance (x-axis), the Z-score of the number of enzyme pairs within various genome distance bins is plotted. The mean and standard deviation used for the estimation of the Z-score are those of randomised networks.

the connectivity of all the random networks was the same as for the *E. coli* SMM network): 574 enzyme pairs at pathway distance 1; 928 pairs at distance 2; 1185 pairs at distance 3; etc. Then, a mean and a standard deviation of the number of pairs in each genome distance bin was calculated, by averaging over the pairs produced for 100 random networks. The distance in standard deviation units of the mean of the distribution from the number of pairs of the protein-centric network existing in each bin and each pathway distance was calculated: there are 52 pairs with a pathway distance of 1 in the 0–100 bp bin (X_{0-100}), but only 0.75 pairs appear on average in the same bin for randomised networks ($\bar{X}_{0-100}$). The standard deviation for this bin for randomised networks (σ_{0-100}) is 0.78. Therefore, at a pathway distance of 1 and for the genome distance bin 0–100 bp:

$$Z_{0-100} = \frac{X_{0-100} - \bar{X}_{0-100}}{\sigma_{0-100}} = \frac{52 - 0.75}{0.78} = 65.75.$$

The Z-scores indicate how far and in what direction each item deviates from the random mean, expressed in standard deviation units. Z-score values greater than 3 are usually considered to be significant. As observed in Fig. 3, our results for the first three bins and the first four pathway distances deviate the most from the

random estimations. After that, the network approaches a more or less random behaviour in the distribution of its enzyme pairs.

The patterns observed in Fig. 2 indicate that SMM genes are "metabolically clustered" on the genome. Furthermore, the relatively high percentage of metabolic-gene pairs found within 100 bp (a very short distance in an ~4.6 Mbp 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). The observation that short genome distances are often observed for functionally related genes has been made before (Tamames et al., 1997; Overbeek et al., 1999; Rison et al., 2002). Here, we show this observation holds true using co-participation in a metabolic pathway as an indication of shared function and "measuring" this relationship using our pathway and metabolic distance metrics.

An intriguing feature of these results is that the main "contributor" to the trend shown in Fig. 2 are the genes within 0–100 bp of one another. The next chromosomal distance bin, 101–1000 bp, is nearly always the rarest. A possible explanation for this comes from assuming an average gene length of approximately 1000 bp; a length thought to be uniform in bacterial genomes (Casjens, 1998). Since the 101–1000 bp just reaches the average length of a gene, it represents an "impossible distance": two genes will either be contiguous (and hence separated by 100 bp or less), or separated by at least one gene (so separated by at least 1000 bp)—thus avoiding the 101–1000 bp bin.

4.3. Pathway distance and function similarity

EC numbers were used as an indicator of shared function. The EC numbers assigned to each enzyme were compared, and the level of EC number conservation was determined. The results are plotted in Fig. 4.

No obvious correlation between EC number and pathway distance could be established. Furthermore, the data show that conservation of EC number is relatively rare at all distances (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 (as defined by an identical first EC number) approximately once out of 4. Furthermore, this percentage is relatively constant at all 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. This contrasts with the recent work of Alves et al. (2002) who, when analysing the metabolic networks of 12 organisms derived from the metabolite-centric KEGG database (Kanehisa et al., 2002), concluded there was often a clustering effect of enzymes belonging to the same class (i.e., sharing the same first EC number) in metabolic networks. In Alves' work, although levels of function conservation in enzyme less than 3 steps apart are significantly higher than that in enzyme pairs more than 3 steps apart regardless of homology, the correlation is substantially more pronounced when

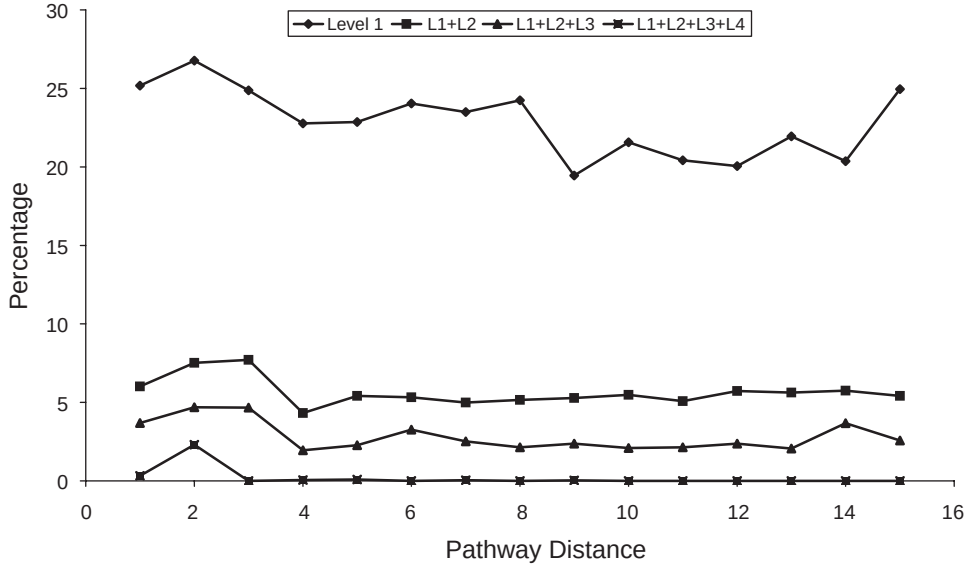


Fig. 4. Pathway distance and function similarity. At each pathway distance (*x*-axis), the percentage of enzyme pairs with all (L1 + L2 + L3 + L4), 3 or more (L1 + L2 + L3), 2 or more (L1 + L2) or 1 or more (Level 1) EC levels matching is plotted. The L1 + L2 + L3 + L4 is a subset of the L1 + L2 + L3 set (which in turn is a subset of L1 + L2, etc.).

considering homologous pairs. In our work, we consider all pairs regardless of homology. It is hard to directly compare the two studies since they use different databases, and the Alves' study exploits pathway distance indirectly (comparing conservation of chemistry in pairs less than 3 steps apart and pairs 3 or more steps apart).

5. Concluding remarks

This work has two salient conclusions: (i) the LP technique is a fast and effective method of analysing certain properties of metabolic networks; (ii) in our work, pathway distance and genome distance correlate, but pathway distance and enzyme function do not, which offers insight into the likely model of pathway evolution.

The algorithm that has been presented here is a single-source shortest path algorithm formulated as an LP model. The algorithm is characterised by its simplicity and deals efficiently with network circularity (i.e., cycles within metabolic pathways). All the computational experiments were performed on an IBM RS6000 workstation. In the case of the study of correlations between minimal pathway distance and genome distance, the analysis required 127 s for the solution of 540 LPs. In the case of the study of correlations between minimal pathway distance and enzyme function, the experiment required 124 s for the solution of 507 LPs. It should be noted that these CPU times include pre- and post-processing of the data, a fairly time-consuming part of the process.

Minimal pathway distances between *E. coli* SMM enzymes have been studied using the algorithm. In our dataset, human intervention has dealt with the issue of promiscuous compounds such as ATP, NAD(P) or water, which if unaccounted for give the representation undesired properties (Alves et al., 2002).

The correlations between minimal pathway distance and genome distance and enzyme function have been investigated. As expected, pathway distance correlated with genome distance with a higher probability of proximity on the genome for genes encoding enzymes involved in nearby metabolic reactions. However, pathway distance 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., 2002; 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 (Jensen, 1976).

Acknowledgments

We thank Sarah Teichmann, Gail Bartlett, and Sophia Tsoka for useful discussions; Peter Karp and Pedro Romero for kindly generating the SMM dataset derived from the EcoCyc database, and for many useful discussions; Monica Riley, Margrethe Serres and Alida Pellegrini-Toole for access to the GenProtEC database and help with our dataset. ES is financially supported by EPSRC (Award No. 00319001) and the Centre for Process Systems Engineering. SCGR was financially supported by GlaxoSmithKline.

References

Alves, R., Chaleil, R.A., Sternberg, M.J.E., 2002. Evolution of enzymes in metabolism: a network perspective. J. Mol. Biol. 310, 311–325.

Arita, M., 2000. Metabolic reconstruction using shortest paths. Simulat. Pract. Theory 8 (1–2), 109–125.

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., Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277 (5331), 1453–1474.

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

Burgard, A.P., 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.

Casjens, S., 1998. The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet. 32, 339–377.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2001. Introduction to Algorithms. The MIT Press, Cambridge, MA.

Dantzig, G.B., 1963. Linear Programming and Extensions. Princeton University Press, Princeton, NJ.

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

Edwards, J.S., Palsson, B.O., 2000b. Robustness analysis of the *Escherichia coli* metabolic network. Biotechnol. Prog. 16 (6), 927–939.

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

Fell, D.A., Wagner, A., 2000. The small world of metabolism. Nat. Biotechnol. 18 (11), 1121–1122.

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

Horowitz, N.H., 1945. On the evolution of biochemical syntheses. Proc. 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., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L., 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.

Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30 (1), 42–46.

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

Kolesov, G., Mewes, H.W., Frishman, D., 2001. SNAPing up functionally related genes based on context information: a colinearity-free approach. J. Mol. Biol. 311 (4), 639–656.

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

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D., 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402 (6757), 83–86.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C., Thornton, J.M., 1998. Protein folds and functions. Structure 6 (7), 875–884.

Michal, G., 1998. Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology. Wiley, London, UK.

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

Pramanik, J., 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., Dunnill, P., 1993. Simulation and optimization of metabolic pathways. Comput. Chem. Eng. 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., Thornton, J.M., 2002. Pathway evolution, structurally speaking. Curr. Opin. Struct. Biol. 12, 374–382.

Rison, S.C.G., Teichmann, S.A., 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.

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

Schilling, C.H., Edwards, J.S., Palsson, B.O., 1999. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. 15 (3), 288–295.

Tamames, J., Casari, G., Ouzounis, C.A., 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., Chotia, C., 2001. The evolution and structural anatomy of the small molecule metabolic pathways of *Escherichia coli*. J. Mol. Biol. 311 (4), 693–708.

Todd, A.E., Orengo, C.A., Thornton, J.M., 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307 (4), 1113–1143.

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

Varma, A., Palsson, B.O., 1993. Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. J. Theor. Biol. 165 (4), 503–522.

Voet, D., Voet, J.G., 1995. Biochemistry, 2nd Edition. Wiley, New York, NY.

Williams, H.P., 1999. Model Building in Mathematical Programming. Wiley, New York, NY.