doi: 10.1016/S0022-2836(02)00140-7 available online at http://www.idealibrary.com on IDEAL®   J. Mol. Biol. (2002) **318**, 911–932

JMB

# Homology, Pathway Distance and Chromosomal Localization of the Small Molecule Metabolism Enzymes in *Escherichia coli*

**Stuart C. G. Rison[1], Sarah A. Teichmann[1] and Janet M. Thornton[1,2]***

[1]*Department of Biochemistry and Molecular Biology University College London Darwin Building, Gower Street London WC1E 6BT, UK*

[2]*Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK*

**Corresponding author*

Here, we analyse *Escherichia coli* enzymes involved in small molecule metabolism (SMM). We introduce the concept of pathway distance as a measure of the number of distinct metabolic steps separating two SMM enzymes, and we consider protein homology (as determined by assigning enzymes to structural and sequence families) and gene interval (the number of genes separating two genes on the *E. coli* chromosome). The relationships between these three contexts (pathway distance, homology and chromosomal localisation) is investigated extensively. We make use of these relationships to suggest possible SMM evolution mechanisms.

Homology between enzyme pairs close in the SMM was higher than expected by chance but was still rare. When observed, homologues usually conserved their reaction mechanism and/or co-factor binding rather than shared substrate binding. The correlation between pathway distance and gene intervals was clear. Enzymes catalysing nearby SMM reactions were usually encoded by genes close by on the *E. coli* chromosome. We found many co-regulated blocks of three to four genes (usually non-homologous) encoding enzymes occurring within four metabolic steps of one another; nearly all of these blocks formed part of known or predicted operons.

The "inline reuse" of enzymes (i.e. the use of the same enzyme to catalyse two or more different steps of a metabolic pathway) is also discussed: of these enzymes, four were multifunctional (i.e. catalysed a different reaction in each instance), nine had multiple substrate specificity (i.e. catalysed the same reaction on different substrates in each instance) and one catalysed the same reaction on the same substrate but as part of two different complexes. We also identified 59 sets of isozymic proteins most commonly duplicated to function under different conditions, or with a different preferred substrate or minor substrate. In addition to transcriptional units, isozymes and inline reuse of enzymes provide mechanisms for controlling the SMM network.

Our data suggest that several pathway evolution mechanisms may occur in concert, although chemistry-driven duplication/recruitment is favoured. SMM exploits regulatory strategies involving chromosomal location, isozymes and the reuse of enzymes.

*Keywords:* homology; small molecule metabolism; pathway evolution; chromosomal localisation; operons

## Introduction

Metabolic pathways form highly regulated networks of enzymes and substrates. In the prokaryote model organism *Escherichia coli*, extensive published experimental work and a completed genome are available.[1] The extant pathways of *E. coli* small molecule metabolism (SMM) are therefore very well characterised and described in

Present addresses: S. A. Teichmann, Structural Studies Division, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; J. M. Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Abbreviations used: SMM, small molecule metabolism.

E-mail address of the corresponding author: thornton@ebi.ac.uk

databases such as EcoCyc,[2] KEGG[3] and WIT.[4] These resources allow us to analyse, at a global level, relationships between SMM enzymes by considering their evolutionary relationships, their position in pathways and the location of the genes encoding them on the *E. coli* chromosome. This will provide a better understanding of how the cell works and how it evolved.

## Pathway evolution

A number of theories have been advanced to explain the evolution of enzyme-catalysed metabolic networks from the "primordial soup". As early as 1945, Horowitz proposed the retrograde model of pathway evolution,[5] followed by Ycas and Jensen, who suggested a patchwork model.[6,7]

In the retrograde evolution model, pathways evolve "backwards" from a key metabolite. The model presupposes the existence of a chemical environment where both key metabolites and potential intermediates are available[5]; each time the environmental supplies of a key metabolite are used up, the organism recruits an enzyme capable of transforming an intermediate into this key metabolite. Every time the environmental reserves of the intermediate drop prohibitively, an enzyme is similarly recruited to catalyse the transformation of another metabolite into the intermediate, etc. In 1965, Horowitz restated his theory to take into account the discovery of operons.[8,9] At the time, the clustering of genes involved in known pathways into operons (e.g. leucine and tryptophan) along with a consideration of the probable origin of operons led him to suggest that operons would cluster genes with overlapping specificities, suggesting structural homology and common ancestry; enzymes within a pathway would tend to be recruited by duplications "within" a pathway.[9] In its strictest form, however, "the stepwise backwards route does not demand that the enzymes are evolutionarily related".[6]

Ycas proposed an alternative to the retrograde evolution theory,[6] which was later expanded and refined by Jensen.[7] In essence, both propose that pathways evolved from a system of broad-specificity enzymes. In this "patchwork evolution" model,[10] enzymes exhibit broad substrate specificities and catalyse classes of reactions.

Therefore, within this large network of possible interactions (including spontaneous non-enzymatic reactions), many paths, some synthesising key metabolites, may have existed, albeit at a very low level. Duplication of genes in such key-metabolite synthesising paths, followed by their specialisation, would account for extant pathways. Furthermore, fortuitous evolution of a novel enzyme-catalysed chemistry could bring into play environmental substrates previously unavailable to the metabolic network. This novel intermediate may then become a new precursor to a key metabolite, even if it is several enzymatic steps away.[7]

Retrograde evolution is generally thought to suppose a "substrate-driven" evolution as, for neighbouring enzymes in a pathway, the product of one enzyme will be the substrate for the next.[9,11,12] By contrast, patchwork evolution is thought to be "chemistry-driven", by recruitment and specialisation of broad substrate specificity enzymes capable of performing the required catalysis.[7,6,10]

## The structural and evolutionary anatomy of SMM pathways

We recently investigated the structural and evolutionary anatomy of SMM pathways in *E. coli*.[13] This investigation gave a comprehensive picture of the pattern of protein domain organisation both within *E. coli* metabolic genes and within and between different metabolic pathways. We found that half of the SMM genes encoded single-domain proteins, whilst the remaining half comprised two or more domains.

In this previous work,[13] we considered each pathway in the EcoCyc database as a separate entity. Comparing the distribution of domain family members within and across pathways, we observed that recruitment of domains across pathways is more common than recruitment within pathways. When considering domain families with more than one member, the majority of families had over twice as many members distributed across pathways as within pathways. Furthermore, pairs of consecutive enzymes exhibiting conservation of substrate binding with a change in catalytic mechanism, a pattern consistent with retrograde evolution,[9] were observed rarely. Rather, the patterns provided support for a non-local recruitment patchwork model of pathway evolution. Similar observations were made by Tsoka & Ouzounis.[14]

## Pathway distance

Here, we make use of a measure known as the pathway distance: the number of distinct metabolic steps separating two enzymes (see Figure 1). By metabolic step, we mean an enzyme-catalysed modification of one or more substrates into chemically distinct compounds. This concept is similar to that of reaction frames found in the EcoCyc database from which we extract our information on the *E. coli* metabolic pathways.[15] EcoCyc reaction frames are computational objects encapsulating an enzyme-catalysed substrate modification. The frame contains the reactants and products of the modification, and is associated with one or more pathways. The reaction frame is associated with one or more enzymes using a linking object. Conceptually, our metabolic steps are the product of this linking, i.e. the enzyme(s) catalysing the transition from reactant(s) to product(s). Indeed, in most cases, the number of "our" metabolic steps and the number of reaction frames in an

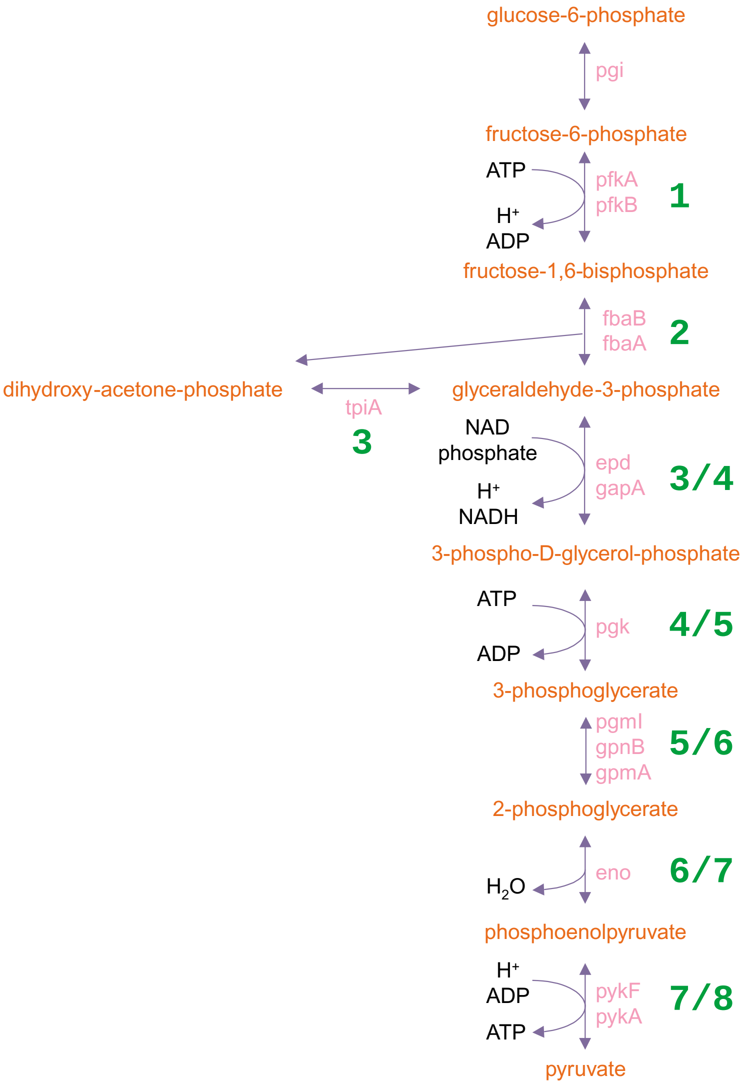

**Figure 1**. Pathway distance illustrated in glycolysis. The pathway shown is glycolysis as represented in the EcoCyc database.[2] Each enzyme-catalysed reaction (blue arrows) represents a metabolic step and therefore a unit of pathway distance. Enzymes catalysing the steps are shown in pink; major metabolites are listed in red; co-factors and minor metabolites are in black. For the enzymes pgi and pykF (located at the beginning and end of the pathway), the longest route (*via* tpiA) has a pathway distance of eight steps (traversing the metabolic steps catalysed by: (i) pfkA and pfkB, (ii) fbaA and fbaB, (iii) tpiA, (iv) epd and gapA, (v) pgk, (vi) pgmI, gpmA and gpmB, (vii) eno, and (viii) pykF). The minimal pathway distance between pgi and pykF is seven steps (avoiding tpiA). Pathway distances relative to pgi are in green; distances from glyceraldehyde-3-phosphate onwards show the minimal and maximal distances. 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 complex; 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.

EcoCyc pathway are identical and differ only when we merge two EcoCyc reaction frames into one (see Methods). Pathway distance has been used independently in a recent work, where it is called "metabolic distance".[16]

Using our measure of pathway distance, adjacent enzymes therefore have a pathway distance of 1. By extension, enzymes catalysing the same metabolic step in a pathway (for example, in Figure 1, pfkA and pfkB) can be thought of as having a pathway distance of 0.

## Gene and context

The analysis of aspects of the genome other than the predicted amino acid sequence of the proteins encoded by the genes has been described as the "context of a gene".[17] Here, we consider three contexts: the genome (i.e. the relative location of SMM enzyme genes on the *E. coli* chromosome); metabolism (i.e. the relative location of enzymes within the SMM network); and the evolutionary context.

Much work has already been done regarding the spatial organisation of genes in bacteria. Tamames *et al.*,[18] considering *Haemophilus influenzae* and *E. coli*, observed that functionally related genes (as classified within simplified scheme derived from GenProtEC's gene classification scheme[19,20]) were neighbours more often than functionally unrelated genes. A strong correlation between genomic clustering and function was detected when considering a large number of genomes†,[21] in particular, for genes in close proximity not just in one, but in many genomes. Recently, the concepts presented by Overbeek *et al.*†[21] were generalised and implemented within a function prediction algorithm that connects genes likely to share functional similarity (in particular, involvement in common metabolic pathways) by analysing orthology and genomic localisation of genes.[16] Such correlations are, however, strongly dependent on phylogenetic distance.[17,22,23]

† http://www.bioinfo.de/isb/1998/01/0009.

**Table 1.** An overview of collected data for analysis of SMM pathways in *E. coli*

| | |
|---|---|
| A. *Pathways* | |
| No. original EcoCyc pathways considered | 102 |
| No. final pathways | 82 |
| No. 82 final pathways composed of two or more of the original 102 EcoCyc pathways | 14 |
| B. *Reaction frames* | |
| No. frames in 82 analysed pathways | 619 |
| No. distinct frames | 581 |
| No. (%) of these 581 frames used more than once | 33 (5.68) |
| C. *Gene assignments* | |
| No. (%) 581 reaction frames with no known genes | 59 (10.15) |
| No. genes in 619 analysed reaction frames | 776 |
| No. distinct genes | 594 |
| No. (%) of these 594 distinct genes used more than once in the pathways | 79 (13.30) |
| D. *GenBank identifiers* | |
| No. (%) 594 distinct genes assigned a GenBank PID | 586 (98.75) |
| E. *Structural (CATH) and sequence domain families* | |
| No. CATH domain families | 220 |
| No. sequence families pre-linkage to structural families | 137 |
| No. sequence families post-linkage to structural families | 117 |
| No. families | 337 |
| F. *Domain family assignments* | |
| No. (%) 586 genes with known PID assigned to one or more CATH domain families | 382 (65.19) |
| No. (%) 586 genes with known PID assigned to one or more sequence domain families only | 98 (16.72) |
| No. (%) 586 genes with known PID assigned to one or more CATH and/or sequence domain families | 480 (81.91) |
| G. *Genomic locations* | |
| No. (%) 594 distinct genes with an identifiable chromosomal location | 584 (98.32) |

### Analysing pathways

SMM pathways have been analysed before.[13,14,24–26] Here, we extend our previous work[13]; patterns of domain distribution and recruitment within the *E. coli* SMM are explored further. We gain deeper insight by exploring a trinity of contexts (evolutionary relationships of genes, genomic location of genes and metabolic environment of enzymes) rather than considering each pathway as a "bag of genes". Furthermore, as much as possible, we analyse the SMM as a single network rather than a collection of arbitrarily defined pathways. Such a "stepwise" analysis allows us to detect hitherto unobserved patterns of recruitment as well as clarify the metabolic range of SMM gene clustering.

We also perform a large-scale analysis of isozymes (homologous enzymes participating in the same metabolic step) and the inline reuse of enzymes (i.e. the reuse of the same enzyme at different locations in the SMM), neither of which has been investigated before.

From these data, we identify certain properties of *E. coli* SMM and discuss their possible implications for the evolution of the SMM network and its regulation.

## Results

### Small molecule metabolism pathways

We obtained our SMM pathways from the EcoCyc database (see Methods).[2] In EcoCyc, data are stored in frames (objects) managed within a Frame knowledge representation system (FRS) known as OCELOT.[15] Frames have slots (attributes), which may be identifiers for instances of other frames. Thus, pathway frames have slots for reaction frames. Reaction frames list the substrates (reactants and products) and, using an intermediary object called an enzymatic reaction, link to the enzyme(s) that catalyse the reaction. From these frames (pathway, reaction and enzymatic reaction), we derived SMM enzymes and their connectivity. The reaction frames conceptualised our notion of a metabolic step defining, as they did, an enzyme-catalysed substrate modification.

Many of the pathways described separately in EcoCyc possess a high level of overlap, i.e. stretches of the same reaction frames are found in both. In our previous work,[13] genes found in reaction frames reused in different EcoCyc pathways were identified as virtual homologues. These virtual homologues reflect the "inter-connectedness" of the pathways, which can be considered more realistically as a network. Here, since we wanted to consider the whole network rather than traditionally defined separate pathways, we dealt with such duplication by merging pathways with three or more reaction frames in common. We began with 102 pathways, composed of 738 reaction frames; following iterative merges, our final dataset contained 82 pathways composed of 619 reaction frames. Of the original 102 EcoCyc pathways, 68 were left unchanged by the merging procedure, one was deleted (as it was found to be represented entirely in other pathways) and the remaining 33 were merged into 14 pathways, accounting for the 82 (68 + 14) final pathways. For example, our largest merged pathway was formed when merging the EcoCyc pathways GLYCOLYSIS/TCA/GLYOX-BYPASS, GLYCOLYSIS/E-D, ANARESP1-PWY, FERMENTATION-PWY, GLUCONEO-PWY and GLYCOL-GLYOXDEG-PWY (respectively, glycolysis/tricarboxylic acid cycle/glyoxylate bypass, glycolysis and Entner–Doudoroff, anaerobic respiration, fermentation, gluconeogenesis and glycol metabolism and degradation). The total number of frames in the six individual pathways was 81, the final number in the merged pathway was 42, illustrating the large overlap between individual pathways.

We briefly compared our final pathways to the 89 SMM pathways that we identified in the metabolic pathway section of the KEGG database.[3] The 14 pathways created from the merger of two or more of the original EcoCyc pathways tended to be similar in size, and sometimes substantially larger, than their KEGG equivalent; the majority of the other pathways were smaller than their KEGG equivalent. However, the KEGG pathways are composite pathways, combining reactions occurring in a number of different organisms into one representation.[3] When considering only the portions of the KEGG pathways predicted by KEGG curators to occur in *E. coli*, all our pathways appeared to be of a similar or larger size than their KEGG equivalent.

Our merging procedure ensured that no two pathways in our final dataset shared three or more consecutive reaction frames. It is worth noting that even if we had merged our pathways to completion (i.e. until no two distinct pathways in the final set shared a reaction frame) we would not have ended up with a single network representing all of *E. coli* SMM. This is because certain pathways are connected only by a common metabolite frame, rather than a reaction frame, and we considered only connectivity between reaction frames. A measure of the remaining level of overlap is that 33 of the 581 distinct reaction frames in the 82 final pathways are found in more than one pathway. Nevertheless, the merges represent a transition from the traditional representation of SMM as distinct pathways towards a network representation.[27,28]

Here, we deal with enzymes of the SMM. It was therefore necessary to assign enzymes to each reaction frame. Of the 581 reaction frames that we considered, 59 had no known genes associated with them, the remaining 522 reaction frames accounted for 594 distinct genes. These genes encode for all the SMM enzymes considered herein.

In Table 1 we summarise certain properties of our dataset, many of which are discussed below†.

## Structural annotation, sequence families and evolutionary relatedness

To investigate the relationship between pathway distance and evolutionary relatedness of SMM enzymes, it was necessary to describe the enzymes in terms of their structural domain composition. Evolutionary relatedness can be detected by pairwise sequence comparison but such methods fail to detect half of the relationships between sequences with identities ranging between 20% and 30%, a proportion that increases substantially when the level of identity drops even further.[29] Since a large number of *E. coli* SMM proteins have a level of sequence identity well below 40%,[13] many relationships would be undetected if we used only pairwise sequence comparison methods. However, structural similarities can detect homologies even for very distantly related proteins with low levels of sequence identity.[30] Therefore, if the structural make-up of *E. coli* SMM proteins can be determined, we can use the properties of structural relationships to determine evolutionary relationships. The "unit" of structure employed here is the structural domain; information on the domain structure and evolutionary relationships of the proteins of known atomic structure is available from the CATH database.[31,32] In CATH, structural domains in the Protein Data Bank (PDB)[33] are classified in a four-level hierarchical scheme. Domains predicted to share a common ancestor on the basis of sequence, structure and functional similarities are assigned the same CATH number and belong to the same superfamily (these superfamilies are subdivided into families on the basis of sequence identity, e.g. S95 sequence families contain members that are 95% or more sequence-identical). Two proteins containing a domain classified in the same CATH superfamily can be considered to be related evolutionarily, at least with respect to that domain.

We used the Gene3D database[34]‡ to assign 382 (65.1%) of the 586 *E. coli* SMM genes considered to one or more of 220 structural (CATH) families. Also, to find all possible evolutionary relationships, we used sequence comparison methods to analyse whole genes and gene regions of greater than 75 residues for which no structural assignments could be made. Using PSI-BLAST[35] and DIVCLUS,[36] as described in Methods, we identified a further 117 sequence domain families. An additional 98 enzymes with no structural assignments were classified into a sequence family, bringing the total number of enzymes assigned to at least one structural or sequence family to 480 (82%) (see Table 1).

These structural and sequence families are employed in this work as indicators of homology.

† Further information regarding the dataset may also be obtained from http://www.biochem.ucl.ac.uk/~rison/EcoliSMM/index.html

‡ http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/

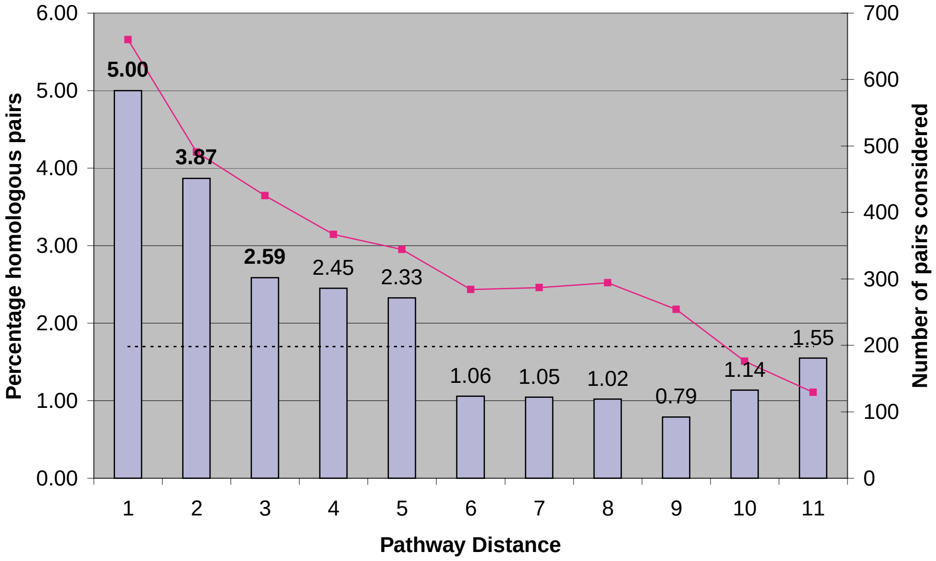

**Figure 2**. Homology and pathway distance. At each pathway distance ($x$-axis), the percentage of enzyme pairs at that distance sharing homology in at least one domain (histogram, primary $y$-axis) is plotted. Observed percentages found by simulation to be statistically significant are in bold type. Only pairs where a structural and/or sequence assignment has been made to both proteins are considered. The number of such pairs is shown (line plot, secondary $y$-axis). The broken line indicates the average percentage of homologous pairs expected if SMM enzymes were distributed randomly (~1.7%).

## Homology and pathway distance

We tallied all homologous pairs (which share at least one CATH or sequence domain) at each investigated pathway distance. The percentage of all positive pairs was then plotted for each pathway distance (Figure 2). Overall, we observed 95 recruitment events: homologous enzymes at 1–11 metabolic steps distance.

In order to ascertain the significance of these data, we calculated the probability ($p$-value) of observing these percentages by chance, as described in Methods. These $p$-values can be found in Table 2; they indicate that the observed percentage duplication for the conserved pathway distances is significantly different from random at only one, two, or three steps (significance cut-off: 0.075). At these distances, the observed level of duplication is significantly higher than expected by chance. Overall, homologous enzymes within the metabolic neighbourhood are rare, accounting for, at most, 5% of the observed instances. Beyond three steps, the likelihood of homology does not appear to be dependent on pathway distance. For each pair, we considered all the domains shared; the 95 homologous pairs accounted for 113 domains. For each of these domains, we classified the rationale for the duplication in one of the following categories: (i) chemistry conserved (where commonalities in the catalytic process dominate); (ii) substrate conserved (identical or similar substrates); and (iii) co-factor conserved pairs (shared co-factor or minor substrate binding domain).[37] Such distinctions are not always obvious to make; often, conservation of chemistry implies a common substrate moiety and, likewise, the nature of the

**Table 2.** The $p$-values for the observed percentages of homologous pairs

| Pathway distance | No. pairs | No. homologous pairs | $p$ |
|---|---|---|---|
| 1 | 660 | 33 | **0** |
| 2 | 491 | 16 | $\mathbf{5.0 \times 10^{-3}}$ |
| 3 | 425 | 11 | $\mathbf{6.2 \times 10^{-2}}$ |
| 4 | 367 | 9 | 0.1 |
| 5 | 344 | 8 | 0.23 |
| 6 | 284 | 3 | 0.71 |
| 7 | 287 | 3 | 0.72 |
| 8 | 294 | 3 | 0.74 |
| 9 | 254 | 2 | 0.93 |
| 10 | 176 | 2 | 0.58 |
| 11 | 129 | 2 | 0.37 |

For each pathway distance analysed, the number of pairs considered is listed (pairs were considered only if at least one domain assignment was available for each enzyme) and the number of homologous pairs is given. Statistically significant $p$-values (cut-off: 0.075) are in bold; the homology percentage observed at these distances is not the consequence of a chance distribution of enzymes.

**Table 3.** Domain conservation explanations

| Domain conservation explanation | No. (%) instances | Example |
|---|---|---|
| Chemistry conserved | 39 (34.5) | MetB and MetC (PLP-dependent aspartate aminotransferase like domain)[38] |
| Co-factor/minor substrate | 35 (31.0) | PurD and PurT (ATP-grasp fold)[13] |
| Substrate binding conserved | 6 (5.3) | TrpA and TrpC (TIM barrel)[39] |
| Unclassified | 33 (29.2) | |

The 113 instances of domain duplications are classified, where possible, into one of three categories: chemistry conserved (where conservation of chemistry is the most salient feature); co-factor/minor substrate binding conservation (e.g. the duplicated domain is nucleotide binding domain); and substrate binding conserved (where the duplicated domains bind identical or similar substrates but the homologous enzymes do not necessarily catalyse the same reaction). In 33 cases we were unable to classify the recruitment unambiguously.

chemistry is often linked to the co-factors used,[38] so not all instances of recruitment were classified. Furthermore, the sequence domain recruitments were not classified. The most common explanation for domain recruitment is conservation of the catalytic mechanism. This accounts for 39 of the 113 instances of domain duplication (34.5%). Conservation of co-factor binding comes second, accounting for 35 (31%) of the cases. The least common apparent cause of domain duplication is conservation of substrate binding, occurring in six instances (5.3%). We were unable to classify the remaining 33 domain duplications unambiguously (See Table 3).

## Homology and gene intervals

Similarly to pathway distance, we considered the relationship between gene intervals and gene homology. We can assign a genome position for 584 (98%) of the 594 distinct genes present in our pathways. Therefore, for the majority of enzyme pairs in our pathways, we can derive a gene interval (i.e. the number of genes on the *E. coli* genome separating the two genes encoding the enzymes in the pair). This generates a discrete distribution of gene intervals. A total of 4405 *E. coli* genes are identified in the Gene Table[1]† that we use for determining genomic locations, therefore, the largest gene interval possible is 2202 (since we consider only the minimal gene interval on the circular chromosome).

We binned gene pairs into five gene interval sets, i.e. the set of gene pairs separated by zero to five genes, by six to 50 genes, by 51 to 500 genes, by 501 to 1000 genes and by more than 1000 genes (see Methods). For each of these bins, we calculated the percentage of homologous pairs. We analysed 594 SMM genes in this work so, theoretically, there are a 176,121 possible gene pairs. However, only 584 genes had an identifiable genomic location (see Table 1) and we considered only pairs for which both genes had at least one structural/sequence family assignment. We could therefore plot the percentage of homologous pairs in the gene interval bins for a total of 124,750 pairs (Figure 3).

## Gene intervals and pathway distance

For a large number of the 176,121 possible SMM genes pairs, no pathway distance is available (i.e. the two genes are further apart than the 11 steps considered or they lie in two distinct pathways). Nevertheless, we have 3495 pairs for which both pathway distance and gene intervals are available; data for these pairs are plotted in Figure 4.

The scatter plot in Figure 4 shows no obvious pattern but binning revealed some trends. At each pathway distance, we grouped the enzyme pairs into gene-interval bins as described above. The relative contributions of the first three bins at various pathway distances is illustrated in Figure 5. There is an evident correlation between pathway distance and the proportion of genes with low gene intervals (zero to five genes) (see Figure 5(a)). To determine to what extent this correlation was due to clustering of metabolic genes into operon transcriptional units, we obtained a list of *E. coli* transcriptional units from the RegulonDB database[40] and used it to flag *E. coli* genes known or predicted to be part of operon structures. Figure 5 shows (b) the subset of all pairs in which both genes are in an operon and (c) where both genes are predicted not to be part of an operon. In the former case, the trend observed in Figure 5(a) is more marked, whilst in the latter case it disappears.

The plot in Figure 5 can be "reversed", considering the relative contributions of genes at a given pathway distance for each gene-interval bin, as shown in Figure 6.

## Homology within reaction frames

Here, isozymes are defined as homologous proteins that perform in the same catalytic step (reaction frame) in *E. coli* metabolism. We distinguish "complete isozymes", where the genes in question are detected to have identical domain make-up, from "partial isozymes", where the proteins have one or more, but not all, domains in common.

† http://www.genome.wisc.edu/pub/analysis/m52orfs.txt

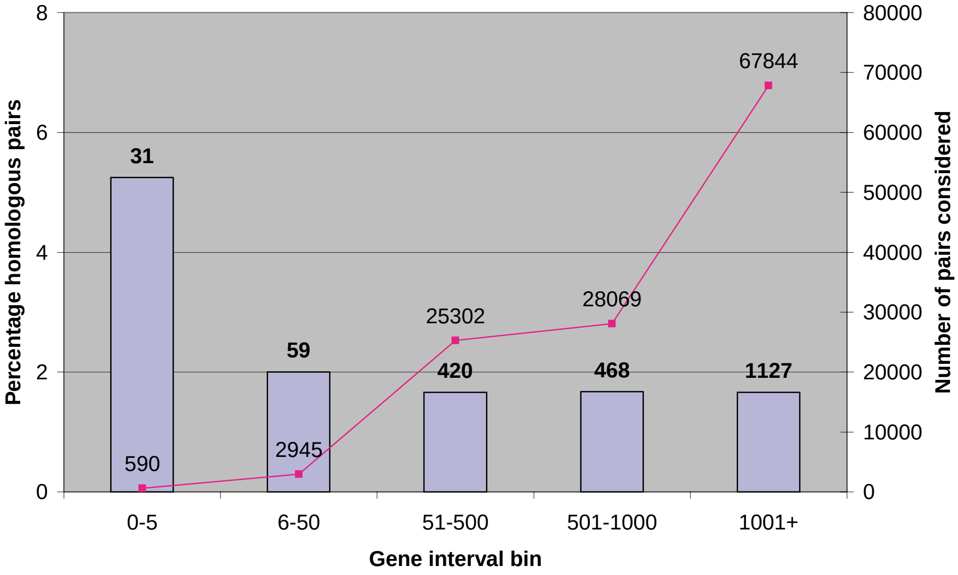

**Figure 3**. Gene intervals and homology. For each gene interval bin, the percentage of all pairs that are homologous are plotted (bars), the actual number of observed homologous pairs is given above each bar (bold type). The line plot shows the total number of pairs considered for each gene interval bin, the numbers of pairs are given.

Of the 339 possible pairs of proteins co-located within a reaction frame (e.g. a reaction frame containing enzymes A, B and C would have possible pairs A–B, A–C and B–C), 66 (19.5%) were complete isozymes (e.g. the aconitases AcnA and AcnB) and 29 (8.5%) were partial isozymes (e.g. the aspartate kinase LysC and the homologous bifunctional MetL and ThrA aspartate kinase/homoserine dehydrogenases, which have only the aspartokinase catalytic domain in common). For 244 (72%) of the protein pairs within a frame, no homology was detected.

One reaction frame may contain more than one set of homologues, because a reaction frame can contain more than one gene product. For example, a reaction frame could contain genes A, B, C and D. If A and B are homologues, and C and D are homologues, but no member of the first set is homologous with a member of the second set, then the reaction frame contains two distinct sets of homologues: AB and CD. Furthermore, as described above, isozymes can be complete or partial; even within one set of homologues, some members of the set may be complete homologues, whilst others may be only partial homologues. Finally, some completely homologous sets may contain proteins of varying sizes. For example, proteins A and B can be flagged as complete isozymes

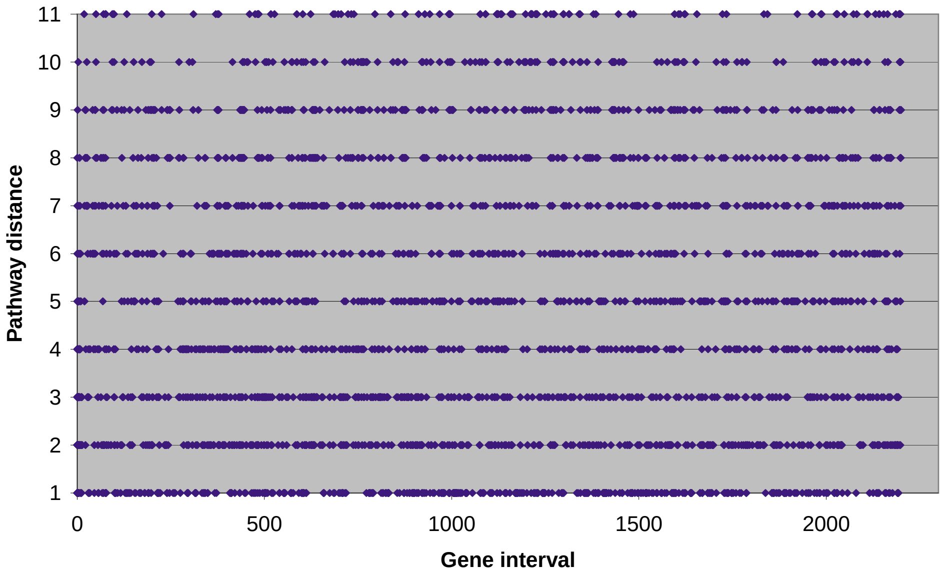

**Figure 4**. Gene interval and pathway distance. Gene intervals are plotted against pathway distance for the 3495 gene pairs where both these measures are available.

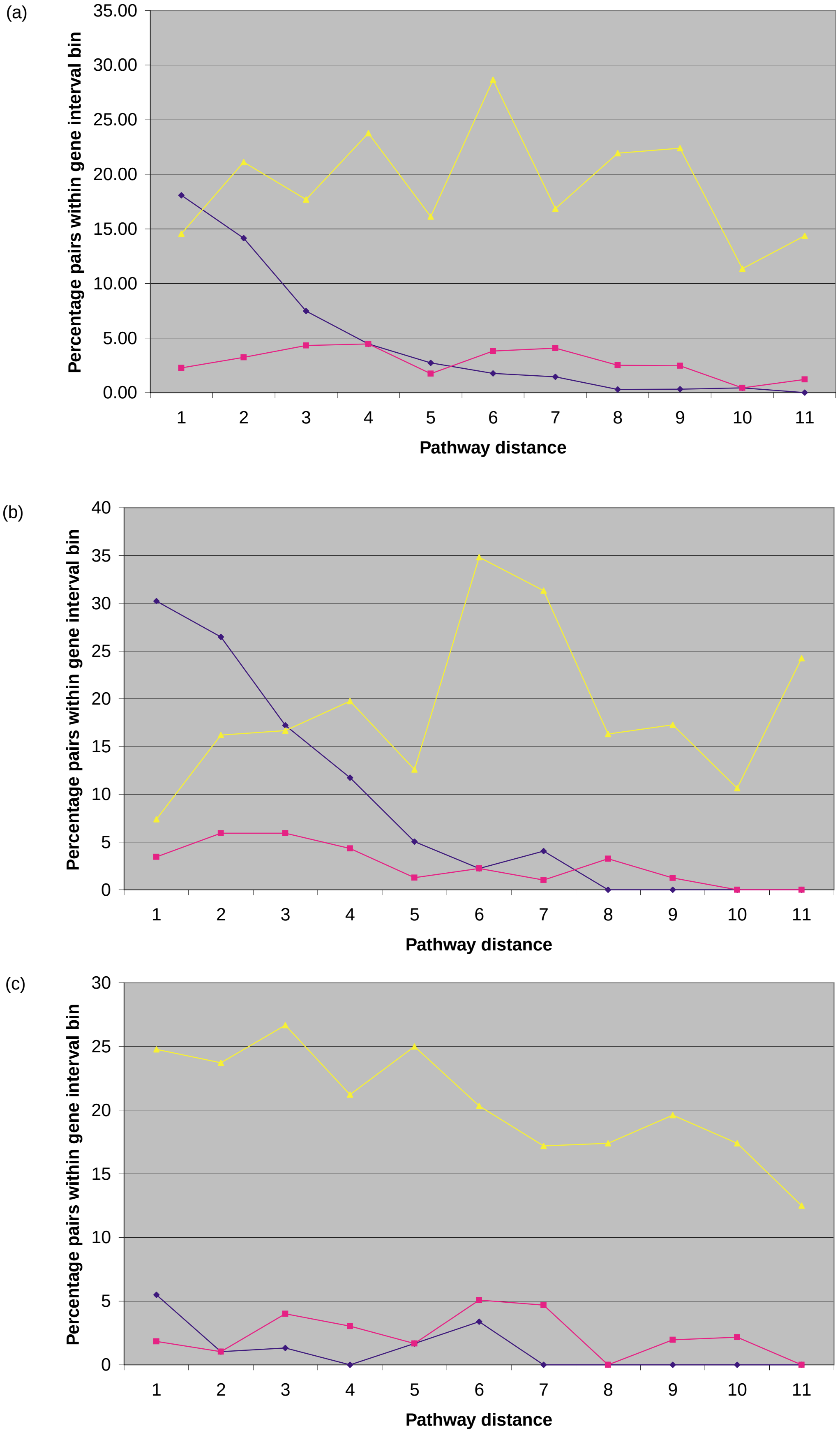

**Figure 5**. Pathway distance and gene intervals. At each pathway distance ($x$-axis), the percentage of enzyme pairs with a gene interval of zero to five genes (blue diamonds), six to 50 genes (pink square) and 51 to 500 genes (yellow triangle) is plotted for (a) all pairs, (b) pairs with both enzymes predicted to be in operons and (c) pairs with both enzymes predicted to be out of operons (operon prediction from Saldago *et al.*[39]).

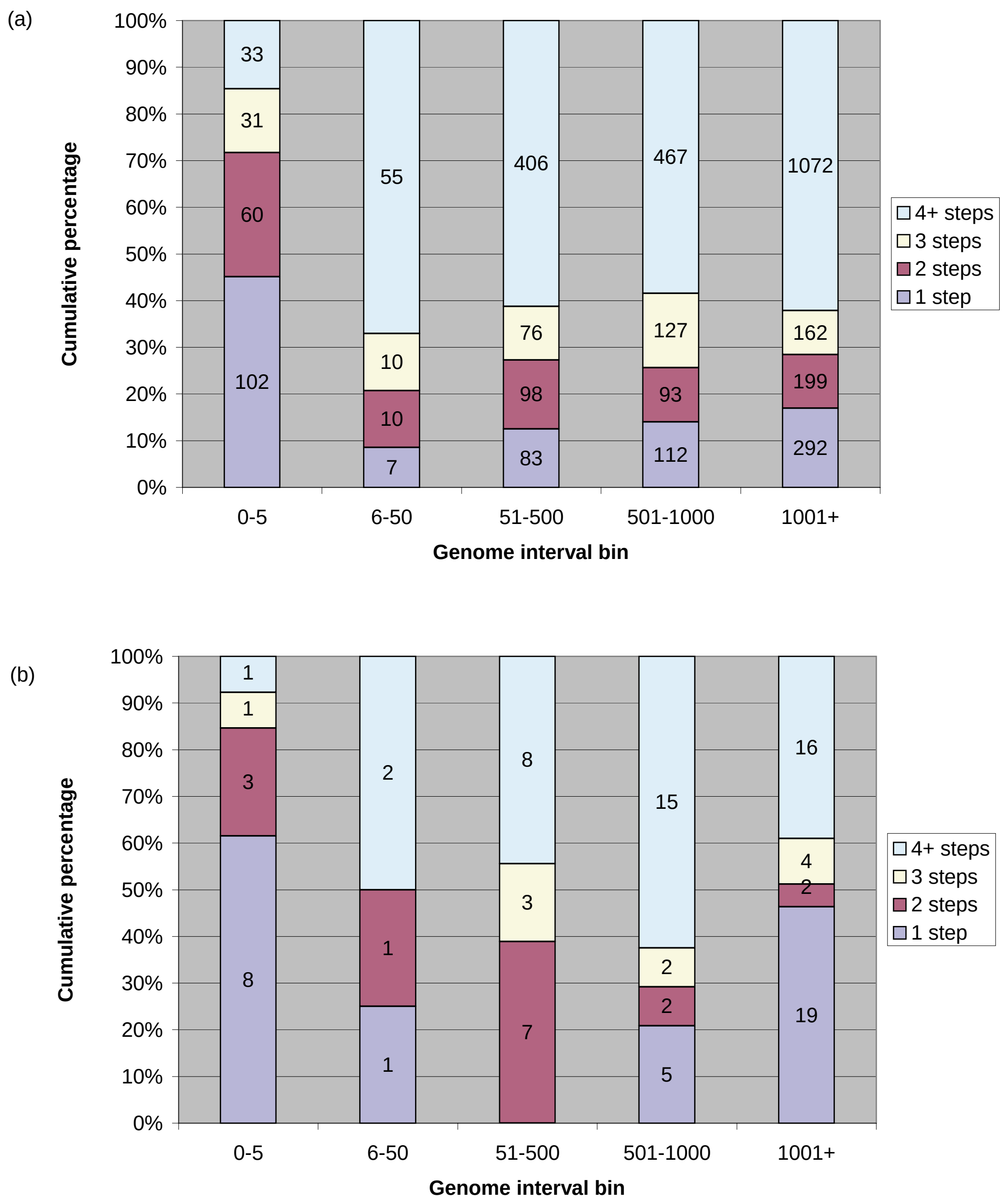

**Figure 6**. Gene intervals, pathway distance and homology. For each gene interval bin, the relative percentages of gene pairs with pathway distances one, two, three and four to 11 steps are plotted for (a) all pairs and (b) only pairs where the genes were found to be homologous. Numbers within bars represent the actual number of pairs observed.

by virtue of having the same domain make-up but protein B might be 50 or more residues longer than protein A, suggesting an unidentified additional domain in protein B.

The 95 (complete and partial) isozyme pairs cluster into 59 sets of homologous proteins (with five instances of reaction frames containing more than one set of distinct isozymes). Where possible, we assigned one or more rationales for the isozymes, identifying nine such reasons. The nine reasons are listed in Table 4 and the 59 sets of isozymes we identified are described in detail in Table 5. To illustrate the scenarios presented above, we look at a couple of examples selected from Table 5. Nitrase reductases NarG, NarZ and NapA are all homologous. NarG and NarZ are complete isozymes, having the same domain make-up, whilst NapA is a partial homologue to both NarG and NarZ, since NapA contains two domains not detected in NarG or NarZ. The homology for this set of proteins is therefore described as C{NarG, NarZ}/P in Table 5. NuoM, NuoN and NuoL are all subunits of NADH dehydrogenase 1. They have an identical domain make-up but NuoM is 509 residues long, NuoN is 425 residues long and NuoL is 613 residues long. The homology for this set of proteins is therefore described as C{104, 188} in Table 5; all members of the set are homologues and differences in size, relative to the largest protein (here NuoL) are given.

**Table 4.** The 9 rationales identified in the 59 set of isozymes described in Table 5

| Rationale | No. instances | Description |
|---|---|---|
| Substrate | 13 | Isozymes have different "preferred" substrates or have one substrate in common but differ in another (usually minor) substrate |
| Conditions | 11 | Isozymes are active under different external conditions (e.g. aerobic/anaerobic; growth media) |
| Different roles | 9 | Although theoretically both enzymes could catalyse the same metabolic step, one of them (usually through different substrate preference) is the effective enzyme with the other performing a similar, but distinct, role. This includes isozymes where one enzyme is anabolic and the other catabolic |
| Complex | 7 | Isozymes are active in the same complex or are constituents of separate (but functionally related) complexes |
| Regulation | 7 | Isozymes' activity regulated internally (e.g. constitutive/induced expression; different allosteric regulation) |
| Kinetics | 6 | Isozymes with different physico-chemical properties (e.g. optimal pH; $K_m$) |
| Localisation | 3 | Isozymes have different heterogeneous group |
| Different co-activity | 2 | Isozymes are both multifunctional and share only one activity in common. Either can perform the catalysis for this common reaction (e.g. metal ion) |
| Heterogeneous | 1 | Isozymes use a different heterogeneous group |
| Unknown | 14 | No clear rationale could be identified |

One set of isozymes can have more than one rationale associated with it, so the total number of rationales exceeds 59. For 14 sets of isozymes, we could assign no rationale.

Most commonly (13 cases), the isozymes had different preferred substrates or minor substrates. For example, AnsA and AnsB both catalyse transamination of aspartate to asparagine but AsnA uses $NH_3$ as the amine "donor" whilst AsnB uses glutamine; FabA and FabZ have different length preferred fatty-acid chain substrates. The isozymes were often active during "different conditions" (11 instances); for example, the fumarases FumA and FumB are active during aerobiosis and anaerobiosis, respectively. In nine cases, the isozymes have different roles, commonly one isozyme was catabolic and the other biosynthetic (e.g. Alr and DadX). In seven cases, the isozymes were part of the same enzymatic complex or constituents of separate (but functionally related) complexes. These cases were difficult to explain unambiguously, although between homologous complexes, homologous polypeptides often performed similar roles (e.g. see the formate dehydrogenases). Different regulation accounted for seven sets of isozymes. For example, the aldolases AroF, AroG and AroH are all subject to different feedback control. We observed different kinetics (six sets), alternative cellular localisations (three sets), different co-activity (two sets) and different heterogenous groups (one set).

## Inline reuse

Enzymes are sometimes used at two or more different metabolic steps within a pathway. Experimentally, this equates to an EcoCyc reaction frame used more than once in the SMM network. This is not the same as the virtual homologues investigated in our previous work,[13] which are a consequence of the arbitrary splitting of the SMM network into many pathways (see above). When we refer to inline reuse, we literally mean the same gene product is used more than once in the SMM network: one enzyme catalyses several distinct steps in different parts of the network. For example, the enzyme DeoD phosphorylates a number of different purine nucleosides during nucleotide metabolism. Between each of these phosphorylation steps one or more other enzymes modify the bases.

We can tally the occurrence of such reuses at each pathway distance. By definition, no inline reuse can occur at pathway distance 1 (enzymes catalysing successive steps in metabolic pathways), since we merge consecutive EcoCyc reaction frames catalysed by the same enzyme (see Methods). Each appearance of a reused enzyme within a network is therefore separated by one or more intervening metabolic steps, we call these intervening frames (IFs). The tally for reuses can be found in Table 6. We observe inline reuse at only pathway distances 2, 3 and 4. Details of the inline reuses are given in Table 7.

Of the 14 inline reused enzymes, one enzyme performs an identical (ID) reaction at each step and four are multifunctional enzymes (MF) that catalyse different reactions, mostly at separate active sites in separate domains. The vast majority of the enzymes reused (nine) perform the same chemistry but act on different substrates along the pathway. These enzymes have multiple-substrate specificity (MS).

The case of LpdA warrants special attention; this is a dihydrolipoyl dehydrogenase and is a subunit in both the pyruvate and α-ketoglutarate dehydrogenase complexes. Whilst the overall chemistries performed by these complexes is different, in both cases, the dihydrolipoyl dehydrogenase subunit re-oxidises dihydrolipoamide, a co-factor used in the reactions catalysed by the other subunits of the complexes. In a sense, LpdA therefore acts independently from the rest of the complex and performs the same chemistry on the same substrate at each point of recruitment.[40]

**Table 5.** Isozymes in *E. coli* SMM

| Homologous genes | Homology | Isozyme rationale | Enzymatic activity | Comment |
|---|---|---|---|---|
| 1. accA, accD | C | Complex | Acetyl-CoA carboxyltransferase | AccA and AccD form the α-2-β-2 complex of acetyl CoA carboxyltransferase |
| 1. entE, entF | P | Complex | Enterobactin synthase complex | EntE and EntF are part of the enterobactin synthase complex and have similar, but distinct, substrates |
| 1. hcaC, hcaE | P | Complex | Dioxygenase | HcaE forms the large α-subunit of 3-phenylpropionate dioxygenase. HcaC is a ferrodoxin |
| 1. nuoG, nuoI, nuoF; 2. nuoM, nuoN, nuoL | C{nuoI, nuoF}/P; C{104,188} | Complex | NADH dehydrogenase | A cluster of 13 genes encodes NADH dehydrogenase 1. NuoFGI are 4Fe−4S proteins. NuoLMN may be involved in proton translocation |
| 1. fdnG, fdoG, fdhF; 2. fdoH, fdnH; 3. fdnI, fdoI | C{fdnG, fdoG}/P; C; C | Complex conditions | Formate dehydrogenases | There are three formate dehydrogenases (FDHs) in *E. coli*: FDH-N, FDH-O and FDH-F. FDH-N is composed of FdnGHI; FDH-O of FdoGHI. The G, H, and I genes are, respectively, the active site subunits, electron transfer subunits and cytochrome subunit of FDH-N and FDH-O. FDH-N is used during nitrate respiration, FDH-O when shifting from aerobiosis to anaerobiosis. FdhF is linked to a hydrogenase complex and also contains an active site |
| 1. hyaB, hyfG, hycE, hybC; 2. hyfA, hyfH, hycF, hycB; 3. hyaA, hyfI, hycG, hybO; 4. hyfB, hyfD, hyfF | C{hyfG, hycE}/P; C{hyfA, hycB}; C{hyfH, hycF}/P; C{hyaA, hybO}; C{hyfI, hycG}/P; C{146,193} | Complex conditions | Hydrogenases | There are four hydrogenase complexes in *E. coli*, active under different conditions. HyaB, HyfG, HycE and HybC are all hydrogenase "large-subunits". HyfA, HyfH, HycF and HycB all have 4Fe−4S domains and HyaA, HyfI, HycG, HybO are all hydrogenase "small-subunits" |
| 1. narG, narZ, napA; 2. narH, narY; 3. narI, narV | C{narG, narZ}/P; C; C | Complex localisation | Nitrate reductases | Nitrate reductase A is composed of NarGHI, nitrate reductase Z of NarZYV and periplasmic nitrate reductase is composed of NapABCDFGH. NarG, NarZ and NapA contain the site of actual nitrate reduction. NarH and NarY are electron transfer subunits and NarI and NarV are the cytochrome *b*-like subunits |
| 1. aceB, glcB | C{190} | Conditions | Malate synthases | GlcB is most active in cells grown on glyoxylate. AceB is active in the glyoxylate bypass |
| 1. acnA, acnB | C | Conditions | Aconitases | AcnB is mainly catabolic, AcnA is a stabler maintenance enzyme |
| 1. fumA, fumB | C | Conditions | Fumarases | FumA is aerobic, FumB is anaerobic |
| 1. glpA, glpD | P | Conditions | Glycerol-3-phosphate dehydrogenases | GlpA forms part of the GlpAB catalytic dimer of glycerol-3-phosphate dehydrogenase. GlpD is aerobic, GlpAB is anaerobic |

(*continued*)

Table 5 Continued

| Homologous genes | Homology | Isozyme rationale | Enzymatic activity | Comment |
|---|---|---|---|---|
| 1. speC, speF | C | Conditions different roles regulation | Ornithine decarboxylases | SpeF is degradative and inducible, especially at low environmental pH. SpeC is biosynthetic and constitutively expressed |
| 1. sodA, sodB | C | Conditions heterogenous group | Superoxide dismutases | SodA complexes with manganese (Mn) and is aerobic. SodB complexes with iron (Fe) and is both aerobic and anaerobic |
| 1. aroK, aroL | C{66} | Conditions kinetics | Shikimate kinases | AroK has a higher $K_m$ than aroL. The enzymes are differently repressed by tyrosine and tryptophan |
| 1. treA, treF | C | Conditions localisation | Trehalases | TreA is periplasmic, TreF is cytoplasmic. TreA is active under conditions of high osmolarity |
| 1. cysK, cysM | C | Conditions substrate | Acetylserine lyases | CysK is acetylserine lyase A. CysM is acetylserine lyase B and can use thiosulphate instead of sulphide ($H_2S$). CysM is required for efficient cysteine biosynthesis during anaerobic growth |
| 1. pheA, tyrA | P | Different co-activity | Chorismate mutases | PheA acts as both a chorismate mutase and phrenate dehydratase whilst TyrA acts as a chorismate mutase and a phrenate dehydrogenase. Both are succeeded by TyrB which turns the product of the former into L-phenylalanine and the latter into L-tyrosine |
| 1. relA, spoT | C | Different co-activity | ppGpp synthases | RelA is a ppGpp synthase and a GTP pyrophosphokinase. SpoT is a ppGpp synthase and a ppGpp pyrophosphohydrolase |
| 1. tdcB, ilvA | C{185} | Different roles | Threonine dehydratases | IlvA is biosynthetic, TdcB is catabolic |
| 1. entC, menF | C | Different roles kinetics | Isochorismate synthases | EntC is the enterobactin synthesis-specific isochorismate synthase and catalyses a reversible reaction. MenF is the menaquinone synthesis-specific isochorismate synthase and catalyses an irreversible reaction |
| 1. alr, dadX | C | Different roles regulation | Alanine racemases | DadX (catabolic) is induced; Alr (biosynthetic) is constitutive |
| 1. sdaA, sdaB, sdhY | C{sdaA, sdaB}/P | Different roles regulation substrate | L-Serine/L-threonine deaminases | SdaA and SdaB are L-serine and L-threonine deaminases; SdhY is only an L-threonine deaminase |
| 1. argD, astC | C | Different roles substrate | Transaminases | AstC (catabolic) has a higher affinity for succinylornithine than for acetylorthinine. ArgD is anabolic |
| 1. epd, gapA | C | Different roles substrate | Dehydrogenases | GapA is the effective glyceraldehyde-phosphate dehydrogenase (GAPDH) with some possible erythrose-4-phosphate dehydrogenase (EPDH) activity. Epd is mainly involved in PLP biosynthesis as an EPDH but has low level GAPDH activity |

(*continued*)

Table 5 Continued

| Homologous genes | Homology | Isozyme rationale | Enzymatic activity | Comment |
|---|---|---|---|---|
| 1. gltA, prpC | C | Different roles substrate | Citrate synthases | PrpC is a methylcitrate synthase with only minor citrate synthase activity. GltA is the effective citrate synthase |
| 1. pflB, tdcE | C | Different roles substrate | Pyruvate/2-ketobutyrate formate lyases | PflB's principal substrate is pyruvate, TdcE's principle substrate is 2-ketobutyrate but both can use the other's main substrate |
| 1. cadA, ldcC | C | Kinetics regulation | Lysine decarboxylases | CadA is the most active decarboxylase. It is also more thermostable and has a low optimum pH LdcC is expressed weakly, less active and thermostable, but has a broad pH range with a higher optimum pH |
| 1. pykA, pykF | C | Kinetics regulation | Pyruvate kinases | PykF is remarkably stable. PykA shows only limited cooperativity among phosphoenolpyruvate binding sites |
| 1. gpt, hpt | C | Kinetics substrate | Phosphorybosyltransferases | Hypoxanthine is the main substrate for hpt, guanine the main substrate for gpt, but both enzymes can use the other's favoured substrate |
| 1. pdxK, pdxY | C | Kinetics substrate | Pyridoxine/pyridoxal kinases | There are two distinct activities: pyridoxal kinase (PL) and pyridoxine kinase (PN). PdxK, pyridoxal kinase, has high PN and moderate PL activity. PdxY, pyridoxal kinase 2, has low PN and high PL activity |
| 1. glpQ, ugpQ | C{111} | Localisation substrate | Glycerolphosphoryl phosphodiesterases | GlpQ is periplasmic. UgpQ is cytoplasmic. They act on different ranges of phosphodiesters |
| 1. aroF, aroG, aroH | C | Regulation | 2-Dehydro-3-deoxy-phosphoheptonate aldolases | These three aldolases have different feedback control and account for different percentages of aldolase activity: AroG (80%), AroF (20%) and AroH (1%) |
| 1. cls, ybhO | C{73} | Regulation substrate | Cardiolipin synthases | YbhO can use different substrates; however, it does not seem to have *in vivo* activity |
| 1. ansA, ansB | C | Substrate | Transaminases | Both AnsA and AnsB catalyse transamination of aspartate to asparagine. AsnA uses $NH_3$ as the amine donor whilst AsnB uses glutamine |
| 1. fabA, fabZ | C | Substrate | β-Hydroxyacyl-ACP dehydrolases | FabZ has broad substrate specificity acting on short to long fatty acid chains; FabA acts mainly on intermediate-length fatty acid chains |
| 1. fabB, fabF | C | Substrate | Acyltransferases | FabB is active in fatty acid elongation whilst FabF, used in membrane phospholipid synthesis, is not |
| 1. ackA, tdcD | C | Unknown | Acetate/propionate kinases | |
| 1. agaY, gatY | C | Unknown | Tagatose 1-6 bis-phosphate aldolases | |
| 1. aldA, aldB | C{63} | Unknown | Aldehydrogenases | AldB function predicted by homology to AldA |

(*continued*)

Table 5 Continued

| Homologous genes | Homology | Isozyme rationale | Enzymatic activity | Comment |
|---|---|---|---|---|
| 1. argF, argI | C | Unknown | Ornithine carbanoyltransferase | Trimers of identical and non-identical chains encoded by duplicate genes ArgI and ArgF produce active ornithine carbanoyltransferase. ArgI and ArgF are found at different loci |
| 1. ddlA, ddlB | C{58} | Unknown | D-Alanine-D-alanine ligases | |
| 1. garR, glxR | C | Unknown | Tartronate semialdehyde reductases | |
| 1. gntK, idnK | C | Unknown | Gluconokinases | |
| 1. gpmA, gpmB | C | Unknown | Phosphoglycerate mutases | |
| 1. ilvB, ilvI; 2. ilvN, ilvH | C; C{67} | Unknown | Acetohydroxybutanoate synthases (AHAS) | There are three AHAS complexes. AHAS I is composed of IlvB and IlvN, and AHAS III is composed of IlvI and IlvH. IlvB and IlvI are catalytic subunits; IlvN and IlvH are regulatory subunits |
| 1. metL, thrA, lysC | C{metL, thrA}/P | Unknown | Aspartate kinases/dehydrogenases | ThrA and MetL are both aspartate kinases and homoserine dehydrogenases. LysC acts as an aspartate kinase only |
| 1. rfbA, rffH | C | Unknown | dTDP-glucose pyrophosphorylases | RffH has not been characterised |
| 1. rfbB, rffG | C | Unknown | dTDP-glucose 4,6-dehydratases | RffG has not been characterised |
| 1. talA, talB | C | Unknown | Transaldolases | |
| 1. tktA, tktB | C | Unknown | Transketolases | TktA has major activity; TktB only minor activity |

Isozymes are homologous proteins found within the same reaction frame. We identified 59 such sets of isozymes. Sets of homologous genes are numbered. Where possible, one or more explanations for the presence of homologues within one frame are given. The "homology" of each set is also described. Sets flagged C are completely homologous (i.e. the same domains have been identified in all proteins in the set). Sets flagged P are partially homologous (i.e. they have one or more domains in common, but not all). Certain sets have mixed homologies, with some of the proteins in the set completely homologous, and others only partially homologous. In such cases, the completely homologous proteins are listed in curly-braces. Finally, some completely homologous sets have proteins of varying sizes, where size differences are greater than 50 residues (suggesting unidentified domain(s)), the size differences relative to the longest protein are listed within the curly-braces. See Results section for more details.

# Discussion

## Recruitment of homologous proteins from the metabolic neighbourhood is rare, but more likely at short distances

The pathway distance range considered herein (1–11 steps) corresponds, in essence, to the "within pathways" of our previous work.[13] Here, we show that homology within pathway distances 1–11 is essentially localised to the shortest of these distances, and that overall recruitment of homologous proteins is rare within this range. Even at pathway distance 2, the distance at which recruitment of homologous proteins is most likely, less than 5% of the possible enzyme pairs share one or more domains (see Figure 2). Of the 3711 enzymes pairs considered (i.e. all the pairs at distances 1–11), only 95 (2.56%) show homology. However, we know recruitment is a common feature in SMM pathways,[13] we can therefore conclude that much of the homology observed previously[13] is the consequence of recruitment from distances greater than 11 steps, from other pathways or indeed from non-SMM genes.

**Table 6.** Inline reuse in *E. coli* SMM

| Steps | No. inline recruitments |
|---|---|
| 1 | N/A |
| 2 | 11 |
| 3 | 1 |
| 4 | 3 |
| 5 + | None |

The number of inline reuses at each distance is listed. Inline reuses are observed only for distances 2, 3 and 4. By definition, there can be no inline reuse of enzymes side by side (pathway distance 1), as identical enzymes found in two adjoining EcoCyc reaction frames were merged into a single frame.

Nevertheless, we do observe 95 homologous pairs within pathways. These have a bias for short distances, with pathway distances 1, 2 and 3 accounting for two-thirds of the cases of homology (see Table 2). When homology does occur, our data show that it is most likely at the shortest pathway distances. Two patterns emerge: at a global level, recruitment events from the metabolic neighbourhood are rare. Recruitment does take place, but it

**Table 7.** Instances of inline reuse. For each instance, the gene recruited, the pathway in which the recruitment occurs, the number of intervening frames (IF) between the two occurrences of the recruited gene and the intervening genes are listed as well as some details (obtained from EcoCyc[2]) concerning the recruitment event (Genes square-bracketed together occur in the same reaction frame. The type of recruitment is also indicated; MF, multifunctional enzyme; MS, multiple substrate specificity; ID, identical reaction)

| Gene | Pathway | Intervening genes | No. IF | Recruitment type | Details |
|---|---|---|---|---|---|
| DgoA | Galactonate catabolism | DgoK | 1 | MF | DgoA is a multifunctional enzyme, it first catalyses the dehydration of D-galactonate, DgoK then phosphorylates the product and the product of the phosphorylation is then lysed by DgoA acting this time as an aldolase |
| metL/ thrA | Homoserine biosynthesis | Asd | 1 | MF | MetL is a bifunctional enzyme performing two non-consecutive reactions, first the phosphorylation of aspartate, then, after the dehydrogenase Asd, MetL oxidises L-aspartate-semialdehyde to homoserine. ThrA is an isozyme of MetL, similarly bifunctional and catalysing the same steps as described for metL |
| tktA/ tktB | Pentose phosphate pathway | [talA, talB] | 1 | MS | TktA catalyses the major transketolase activity in *E. coli.* In this pathway, it acts both on ribose-5-phosphate and xylulose-5-phosphate, producing the substrates for the next reaction catalysed by transaldolases talA and talB, which, in turn, produce one of the two substrates for the second transketolase reaction listed in the pentose phosphate pathway. TktB catalyses the minor transketolase activity in *E. coli*; it is an isozyme of TktA and performs reactions identical with those listed for tktA |
| RelA | ppGpp metabolism | GppA | 1 | MS | GTP pyrophosphokinase catalyses the synthesis of guanosine 5′-triphosphate 3′-diphosphate (pppGpp) as well as guanosine 3′,5′-bispyrophosphate (ppGpp) by transferring the pyrophosphoryl group from ATP to GTP or GDP, respectively. Phosphatase GppA catalyses the transition from pppGpp to ppGpp |
| deoD | Nucleotide metabolism | Add | 1 | MS | DeoD is a ubiquitous purine nucleoside phosphorylase multiply recruited within nucleotide metabolism. DeoD catalyses the generalised reaction purine nucleoside + orthophosphate = purine + α-D-ribose 1-phosphate. In this instance of reuse, DeoD phosphorylases |
| add | Nucleotide metabolism | DeoD | 1 | MS | Add (deoxyadenosine deaminase/adenosine deaminase) and DeoD mutually bracket one another (i.e. the chain deoD, add, deoD, add occurs in the nucleotide metabolism pathway). Functions of deoD and add are described above |
| hisB | Histidine biosynthesis | HisC | 1 | MF | HisB encodes a single polypeptide possessing the two enzyme activities: histidinol-P phosphatase and imidazoleglycerol phosphate dehydratase. The intervening enzyme, HisC, acts as a histidine phosphate aminotransferase |
| Ndk | Pyrimidine ribonucleotide/side metabolism | PyrG | 1 | MS | Ndk is a nucleoside diphosphate kinase with broad substrate specificity: the terminal phosphate group of a nucleoside-triphosphate is transferred to a nucleoside-diphosphate. In the first such reaction, UDP is phosphorylated to UTP. UTP is converted to CTP by the CTP synthase PyrG. In turn, CTP acts as the nucleoside-triphosphate donor to ADP |
| Udk | Pyrimidine ribonucleotide/side metabolism | Cdd | 1 | MS | Uridine kinase Udk phosphorylates both uridine and cytidine. Cytidine deaminase Cdd catalyses the conversion of cytidine to uridine |
| purB | Nucleotide metabolism | purH; purA | 2 | MS | Adenylosuccinate lyase PurB catalyses the removal of fumarate from 5′ phosphoribosyl-4-(*N*-succinocarboxamide)-5-aminoimidazole and from succinyl-AMP to form AICAR and AMP, respectively. PurH is a bifunctional AICAR transformylase and IMP cyclohydrolase. PurA is adenylosuccinate synthase. PurH and PurA convert AICAR to succinyl-AMP. (AICAR: aminoimidazole carboxamide ribonucleotide) |
| ubiG | Ubiquinone synthesis | ubiH; ubiE; ubiF | 3 | MS | UbiG catalyses both the O-methylation reactions involved in ubiquinone synthesis. These take place three metabolic steps apart. UbiH, UbiE and UbiF catalyse the intervening hydrolysis, methyltransferase and hydroxylase steps, respectively |

(*continued*)

Table 7 Continued

| Gene | Pathway | Intervening genes | No. IF | Recruitment type | Details |
|---|---|---|---|---|---|
| Ndk | Deoxy-pyrimidine nucleotide/side metabolism | dut; thyA; tmk | 3 | MS | Ndk's role in pyrimidine nucleotide/side metabolism is described above. It plays a similar role in deoxypyrimidine nucleotide/side metabolism, catalysing the transformation of dUDP to dUTP in one case and from dTDP to dTTP in the other. The intervening enzymes, a pyrophosphatase, a synthase and a kinase covert dUTP to dTDP via dUMP and dTMP |
| LpdA | Glycolysis and TCA | gltA; [acnA, acnB]; icdA | 3 | ID | LpdA is the dihydrolipoamide dehydrogenase subunit of the pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase complexes. As part of the first complex it is involved in the formation of acetyl-CoA from pyruvate and as part of the second, of succinyl-CoA from 2-oxoglutarate, but in both instances it perform the same chemistry on the same substrate.[40] These steps are connected by TCA enzymes citrate synthase (GltA), aconitases A and B (AcnA/B) and isocitrate dehydrogenase (IcdA) |

does so from the most suitable enzyme not the "nearest" enzyme.[13] At a local level, however, when considering only homology in the metabolic neighbourhood (pathway distances 1–11), instances are not distributed uniformly, rather, more occur at shorter distances. The simplest explanation for this observation would be that, for example, for two homologous enzymes A and B found one step apart, A was recruited from B or B was recruited from A (adjacent recruitment). In the absence of convincing phylogenetic information, however, we cannot exclude the possibility that either A or B was in fact recruited from a homologue at some distance in the SMM network. Assuming that we are observing instances of adjacent recruitment, the drop in observed homology with increasing distance is consistent with the retrograde model of evolution.[5,9]

Analysis of our data shows that substrate conservation is not the principal explanation for the observed recruitments in our pathway (See Table 3). Whilst the pattern of recruitment shown in Figure 2 is consistent with retrograde pathway evolution, the rarity of conservation of substrate binding is not. Furthermore, in the absence of phylogenetic data, we have no directional information to discriminate between forward or retrograde evolution.

## Nearby pathway enzymes are clustered in the genome

It is known that gene separation can be used as an indicator of shared function†[18,21] and physical interaction.[23] One possible conception of shared function is proximity in the SMM network so one might reasonably expect to observe a distinct trend when plotting pathway distance against gene interval, but the plot shows a range of gene intervals at each pathway distance (see Figure 4).

However, the process of binning the gene intervals reveals a clear trend: enzymes coded by nearby genes in the *E. coli* genome are more likely than distant ones to be close in a pathway (Figure 5). The correlation between pathway distance and gene interval when considering all pairs (Figure 5(a)) is strengthened when considering only "operon pairs" (Figure 5(b)) but disappears when considering "non-operon pairs" (Figure 5(c)); so it would appear that operons do account for this correlation. We considered the pattern observed for cumulative percentages at each pathway distance (data not shown). By pathway distance 4, over 90% of pairs observed with gene interval zero to five had already been encountered. By contrast, only by pathway distance 8 was a similar percentage of the pairs with gene interval 51–500 observed and, for larger bins, the pathway distance was 9 or greater. For SMM genes, we are observing an operon effect, but this is a short-range effect, essentially only clustering genes found at pathway distances of 4 or less.

We tested this theory by considering the 845 known and predicted operons obtained from RegulonDB.[39] Only 104 of these contained at least one pair of SMM genes (i.e. two or more of the 594 genes in our 82 SMM pathways). In 81 of these (78%), all SMM gene pairs with a known associated pathway distance were less than five metabolic steps apart, and in 72 cases (69%), all possible gene pairs were within five metabolic steps. That is, in nine of the cases, the operons included gene pairs for which no pathway distance was identified (i.e. genes in separate pathways, or at distances greater than 11 steps or containing non-SMM genes) but in 72 of the cases the operon was composed of only SMM genes within five steps of one another. The 81 operons obeying the "within five steps" rule account for 235 of the 594 SMM genes (40%). This increases to 58% when considering only the 402 SMM genes known or predicted to be in an operon.

Interestingly, a similar "plateau" at pathway distance 4 was observed by Kolesov *et al.*[16] and a

† http://www.bioinfo.de/isb/1998/01/0009.

median size of 3 "same-pathway" gene clusters was observed by Overbeek *et al.*,[21] with the latter considered an underestimate by the researchers.

The observation that, in prokaryotes, functionally related genes cluster and that these genes often participate in the same biosynthetic pathway is neither unexpected nor novel, and this clustering is generally accepted to be the consequence of the operon gene organisation of prokaryotes. However, this relationship has not previously been explored quantitatively for the whole SMM of an organism and verified on a set of known and predicted operons. By correlating gene interval and pathway distance we "measure" the range of the clustering. Analysis of the *E. coli* genome suggests an average operon size of three to four genes.[39] We conclude that, in general, for *E. coli* SMM enzymes, operons cluster blocks of three to four genes all within a short (four steps or less) pathway distance of one another. These operons are possibly co-regulated at a higher level in "uber-operons"[41] This observation constitutes an important rationale for the often-exploited use of genomic co-localisation in gene function prediction.

## Genome distance, pathway distance and homology

Following the observations that SMM genes nearby on the chromosome often code for enzymes nearby in the SMM network, and that enzymes nearby in the SMM are more likely to be homologous, we investigated the correlation of genome distance and homology (Figure 3). Of the 590 enzyme pairs with a gene interval of zero to five genes, 31 (5.25%) were homologous, whilst for the other bins considered, the proportion of homologous pairs was approximately 2%. We tested the significance of the percentage observed for the zero to five bin (data not shown). The observed increase in homology in this bin relative to others is not due to chance, nor to a sampling effect (due to the relatively small size of the bin). Genes close by in the genome are more likely to be homologous than genes further apart but homology is still rare. In other words, genes nearby on the genome are likely to be related functionally but not necessarily related evolutionarily.

The three contexts considered here (genome, metabolism and evolutionary relationship) are presented together in Figure 6. Three facts emerge from our investigation: (1) Enzymes close by in the SMM network are often encoded by genes close by in the genome (12% of pairs of proteins four or less metabolic steps apart are encoded by genes separated by, at most, five genes). (2) Enzymes close by in the SMM network are more likely to be homologous than distant ones (2.9% of pairs of proteins four or less metabolic steps apart are homologous compared to only 1.5% for pairs of proteins separated by more than four metabolic steps). (3) Genes close by in the genome are more likely to be homologous than distant ones (5.2% of pairs of genes separated by five or less genes are homologous compared to 1.7% of pairs of genes separated by more than five genes).

However, facts (2) and (3) must be mitigated; the number of relevant instances in both cases is low relative to the number of instances that do not exhibit homology, suggesting that these trends, although significant, do not apply to the majority of cases. Nevertheless, the simultaneous exploitation of three contexts is a novel development in the analysis of SMM networks. Even though facts (2) and (3) may have been expected, it remained to test them *in situ*. Indeed, the fact that they are rare events is in itself an interesting observation.

## Operons, inline-reuse, isozymes and regulation

Our data illustrate three regulatory mechanisms operating within *E. coli* SMM pathways: the use of operons; the inline-reuse of enzymes; and the use of isozymes. The first two act as co-regulatory mechanisms. Conversely, the latter mechanism allows organisms to "divide" control of metabolic steps between different sets of isozymes fine-tuned for different conditions.

Operons cluster functionally related genes. In the case of SMM genes, they ensure the coordinated presence of enzymes, as the absence of any one enzyme along a linear pathway would block it. Since SMM is a large network, it would not be feasible to place all SMM enzymes under the control of a single promoter. However, it is equally infeasible to have all SMM enzymes under individual control. Our observations suggest a compromise solution, the clustering of nearby (less than five metabolic steps) pathway genes in "blocks" of three to four genes.

Inline-reuse of proteins can be thought of as a form of co-regulation; expression of a single enzyme guarantees the catalysis of several steps. For multi-substrate (MS) reused enzymes, the catalysed steps are related by chemistry. The latter are classic examples of enzymes that have a broad specificity that have been utilised in the evolving cell.[7] In the case of multifunctional (MF) reuse, chemistries are different at each catalysed step but the fusing of two independently functional entities into one enzyme can be thought of as the ultimate co-regulation mechanism, a scenario known to occur commonly in *E. coli* SMM.[42]

Few of our isozymes are co-located within an operon structure and therefore within a short distance of one another (only five out of our 59 sets had all isozymes in a set within five genes of one another). We have previously found lateral gene transfer not to play a key role in this observation.[13] Adjacent genes would suggest a recent duplication event or strong evolutionary pressure to keep the genes nearby. It would appear that for our set of isozymes, the duplication events are not recent and evolutionary pressure has acted to separate the genes to allow segregation of transcriptional control and/or future specialisation of the isozymes.

SMM networks are ancient and have had a long time to be segregated and specialised. Nevertheless, a number of the instances of homology that we observe, in particular the isozymes with no clear rationale for duplication (e.g. argF and argI), could be awaiting functional and regulatory specialisation. It may be the case that nearby isozymes are more common in recently evolved pathways.[43]

## Conclusion

The data presented here add some support to the growing body of evidence suggesting patchwork evolution as the prevailing pathway evolution strategy:[13,14,24] recruitment from the metabolic neighbourhood (1–11 steps) is rare, as is conservation of substrate binding with a change in associated chemistry. Nevertheless, homology within the metabolic neighbourhood does occur, and when it does, it is more likely to occur at short pathway distances, including some well-known "retrograde-like" instances, suggesting multiple evolutionary mechanisms occurring in concert. We are observing catalytic constraints (i.e. the necessity to evolve a chemically efficient network for the production of small molecules), and we are observing extensive regulatory constraints (to ensure that the SMM is controlled efficiently to deal with changes in both intracellular and extracellular conditions). *E. coli*'s extant SMM pathways are the result of these pressures.

The picture is complex; further clarification may come from effective phylogenetic analysis of all SMM enzymes (as performed "manually" by Copley & Bork for TIM barrels[24]) and experimental and theoretical investigation of metabolic pathways in not one but many organisms.[44,45] Nevertheless, the interaction between the genome context, the metabolic context and the evolutionary context is certainly worth "mining" for information (e.g. see Kolesov *et al.*[16]). Such methods are effective because, as described here, there are exploitable relationships between all these contexts.

## Methods

### Generating the pathway dataset

The SMM pathways analysed in this work were obtained from the EcoCyc database.[2] Pathway data were downloaded and converted to a format suitable for easy parsing by a number of Perl scripts.[46] In keeping with the EcoCyc architecture, we downloaded data describing the pathway frames, and data describing reaction and enzyme-reaction frames,[15] and stored them locally using a relational database management system (postgreSQL). Some of the data were edited manually following update reports (Alida Pellegrini-Toole and Monica Riley, personal communication).

This architecture allowed us to calculate pathway distances for any two enzymes in a pathway, and to derive ancillary information for the enzymes (such as gene identifier, products, co-factors). In the EcoCyc database, certain pathways are represented both in isolation and as a subpathway of larger pathways. For example, glycolysis is represented on its own, as well as in combination with the tricarboxylic acid (TCA) cycle and glyoxylate bypass: glycolysis is considered a subpathway of the latter "combined" pathway and, conversely, the latter is a superpathway of glycolysis. To avoid partial pathway duplications, we downloaded only superpathways with no superpathways (i.e. superpathways not themselves a subpathway of an even larger superpathway) and "atomic" pathways (pathways with no superpathways or subpathways). Even then, the downloaded pathways exhibited some overlap; using a recursive procedure, we further merged the pathways such that no two pathways in our dataset overlapped by more than two EcoCyc reaction frames. Finally, we fused any set of two adjacent reaction frames catalysed by identical enzymes. In EcoCyc, these usually represent enzymes that generate an identifiable intermediate compound. For our purposes, however, we chose to think of the complete reaction from substrate(s) to final product(s) as a single metabolic step, regardless of the observable intermediates.

Our final dataset contained 82 pathways, containing 619 reaction frames. More information regarding the dataset can be found in Table 1.

### Gene identification and ancillary data

The dataset generated above described the reaction frames and their relationships. Reaction frames describe a metabolic transition in terms of the substrates, products, co-factors and enzyme(s) catalysing that step.[15,47] For all calculations involving the enzymes themselves, we needed to assign genes to reaction frames. Most of these assignments were obtained directly from EcoCyc, with some additional manual correction. In EcoCyc, genes are commonly described by their gene symbol (e.g. gapA or pgk) or their Blattner number (e.g. b3919 or b2926), but the Gene3D structural assignment procedure (see below) required GenBank protein identifiers (PIDs).[48] We converted gene symbols and Blattner numbers to GenBank identifiers using a conversion list obtained from GenProtEC,[19] which was edited manually following update reports (Margrethe Serres, personal communication).

### Genomic location and gene intervals

Genes were assigned a chromosomal location by consulting the Gene Table† for *E. coli*[1] using the GenBank identifiers described above. We derived a gene order with genes ordered, irrespective of their strand, on the basis of their boundaries (i.e. starting at position 1 on the circular chromosome and numbering genes by scanning clockwise for boundaries, regardless of whether the boundary was a start codon, as would be the case for genes on the (+) strand or as stop codon, as would be the case for genes on the (−) strand); the ranking obtained was nearly identical with the Blattner

† http://www.genome.wisc.edu/pub/analysis/m52orfs.txt

numbering. The gene interval is a measure of the number of genes separating two genes as derived from the aforementioned ordering (e.g. a gene interval of zero for genes sides by side, of one for two genes separated by a third gene, of two for genes separated by two other genes, etc.)

### Assignments to structural and sequence families

We used the Gene3D database to obtain structural assignments, where possible, for enzymes in the *E. coli* SMM. Full details of the methodology have been described†[34] but, briefly, the method was as follows: (1) PSI-BLAST[35] profiles were generated for non-identical sequences in CATH v1.7, filtered at 95% sequence identity (S95 representatives). (2) *E. coli* (GenBank) SMM genes were scanned against the S95 profiles using IMPALA[49]; matches were considered only when the profile match covered 50% or more of the S95 representative sequence. (3) The assignments were finalised for each gene using clean-up scripts that resolved assignment clashes and fixed domain boundaries.

To expand this repertoire of evolutionary relationships, we considered both gene segments encoding 75 or more residues for which no structural assignment was made (suggesting a undetected domain) and sequences wholly unassigned. These were used as query sequences in PSI-BLAST searches against the *E. coli* genes incorporated within NRDB100 non-redundant nucleotide database obtained from GenBank[48] (maximum 20 iterations or convergence; *e*-value cut-off for inclusion in next iteration 0.0005). The results were clustered into sequence families using the DIVCLUS package.[36] Query sequences connected to a structural (CATH) family by virtue of an intermediate sequence were assigned to that family. The remaining clusters for "sequence" families represent evolutionary relationships undetected using the IMPALA strategy (e.g. because the structural domain equivalent to the sequence family was not present in CATH v1.7).

We identified 138 sequence families, 21 of which could be associated with a structural family by virtue of one or more intermediate sequences,[50] leaving 117 sequence families. Of the 382 *E. coli* SMM enzymes assigned to one or more of the structural families, a further 98 enzymes were classified within a sequence family, giving an overall evolutionary relationship coverage of approximately 82%. These observations are summarised in Table 1.

### Calculating pathway distances

For each of the pathways analysed, we defined source and sink metabolites, and identified all possible reaction frames' traversals between these using a depth-first search (DFS) algorithm.[51] Cycles were "snipped" arbitrarily and reaction direction was not taken into account. From the traversals, pairs of reaction frames at user-defined distances (i.e. specified number of steps) were extracted. Duplicate pairs (i.e. same reaction frames at the same pathway distance but reached via alternative routes) were eliminated. However, we did not eliminate identical reaction frames pairs found at different pathway distances.

### Estimating *p*-values

We performed an all *versus* all comparison of SMM enzymes with at least one structural or sequence assignment and flagged all pairs sharing at least one sequence or structural domain. We then picked randomly, with no replacement, a number of pairs equal to that considered for each distance (e.g. there are 660 valid pairs observed at pathway distance 2). For each set of pairs picked, we calculated the percentage of positive pairs (i.e. having at least one domain in common). We repeated the picking process 500,000 times to derive the average random percentage of positive pairs, its standard deviation and the *p*-value for the experimental percentages.

## Acknowledgments

We thank Peter Karp for access to, and help with, the EcoCyc database; Monica Riley, Alida Pellegrini-Toole and Margrethe Serres for access to, and help with, GenProtEC and for generating gene identifier conversion tables; Daniel Buchan, David Lee and Frances Pearl for help with Gene3D and CATH; Julio Collado-Vides and Gabriel Moreno-Hagelsieb for access to RegulonDB and kindly providing easily parseable *E. coli* transcriptional unit data to us; and Gail Bartlett and Annabel Todd for many useful enzyme-related discussions. S.C.G.R. was funded by GlaxoSmithKline. S.A.T. had a Beit Memorial Fellowship. This is a publication from the Bloomsbury BBSRC Structural Biology Centre.

## References

1. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
2. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M. & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucl. Acids Res.* **28**, 56–59.
3. Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30.
4. Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Jr. & Kyrpides, N. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucl. Acids Res.* **28**, 123–125.
5. Horowitz, N. H. (1945). On the evolution of biochemical synthesis. *Proc. Natl Acad. Sci.* **31**, 153–157.
6. Ycas, M. (1974). On earlier states of the biochemical system. *J. Theor. Biol.* **44**, 145–160.
7. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
8. Jacob, F. & Monod, J. (1961). On the regulation of gene activity. *Cold Spring Harbor Symp. Quant. Biol.* **26**, 193–211.
9. Horowitz, N. H. (1965). The evolution of biochemical synthesis: retrospect and prospect. In *Evolving Genes and Proteins* (Bryson, V. & Vogel, H., eds), pp. 15–23, Academic Press, New York.
10. Lazcano, A. & Miller, S. L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell*, **85**, 793–798.

† http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/

11. Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D. & Kozarich, J. W. (1993). On the origin of enzymatic species. *Trends Biochem. Sci.* **18**, 372–376.
12. Gerlt, J. A. & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246.
13. Teichmann, S. A., Rison, S. C. G., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001). The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693–708.
14. Tsoka, S. & Ouzounis, C. A. (2001). Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. *Genome Res.* **11**, 1503–1510.
15. Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.
16. Kolesov, G., Mewes, H. W. & Frishman, D. (2001). SNAPping up functionally related genes based on context information: a colinearity-free approach. *J. Mol. Biol.* **311**, 639–656.
17. Huynen, M. A. & Snel, B. (2000). Gene and context: integrative approaches to genome analysis. *Advan. Protein Chem.* **54**, 345–379.
18. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
19. Riley, M. (1998). Genes and proteins of *Escherichia coli* K-12. *Nucl. Acids Res.* **26**, 54.
20. Rison, S. C. G., Hodgman, T. C. & Thornton, J. M. (2000). Comparison of functional annotation schemes for genomes. *Funct. Integr. Genom.* **1**, 56–69.
21. 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**, 2896–2901.
22. Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol.* **2**, research0020.1–research0020.11.
23. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328.
24. Copley, R. R. & Bork, P. (2000). Homology among $(\beta\alpha)^8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627–641.
25. Saqi, M. A. S. & Sternberg, M. J. E. (2001). A structural census of metabolic networks for E. coli. *J. Mol. Biol.* **313**, 1195–1206.
26. Rison, S. C. G. & Thornton, J. M. (2002). Pathway evolution structurally speaking. *Curr. Opin. Struct. Biol.* In the press.
27. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
28. Küffner, R., Zimmer, R. & Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825–836.
29. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
30. Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380–387.
31. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH: a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
32. Pearl, F. M., Martin, N., Bray, J. E., Buchan, D. W., Harrison, A. P., Lee, D. *et al.* (2001). A rapid classification protocol for the CATH domain database to support structural genomics. *Nucl. Acids Res.* **29**, 223–227.
33. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
34. Buchan, D. W. A., Shepherd, A. J., Lee, D., Pearl, F., Rison, S. C. G., Thornton, J. M. & Orengo, C. A. (2002). Gene 3D: structural assignment for whole genes and genomes in the CATH domain structure database. *Genome Res.* **12**, 503–514.
35. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
36. Park, J. & Teichmann, S. A. (1998). DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, **14**, 144–150.
37. Nahum, L. A. & Riley, M. (2001). Divergence of function in sequence-related groups of *Escherichia coli* proteins. *Genome Res.* **11**, 1375–1381.
38. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
39. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F. *et al.* (2001). RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucl. Acids Res.* **29**, 72–74.
40. Voet, D. & Voet, J. G. (1995). *Biochemistry*, 2nd edit., Wiley, New York.
41. Lathe, W. C., III, Snel, B. & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25**, 474–479.
42. Tsoka, S. & Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet.* **26**, 141–142.
43. Copley, S. D. (2000). Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem. Sci.* **25**, 261–265.
44. Dandekar, T., Schuster, S., Snel, B., Huynen, M. & Bork, P. (1999). Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**, 115–124.
45. Forst, C. V. & Schulten, K. (2001). Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* **52**, 471–489.
46. Wall, L., Christiansen, T. & Schwartz, R. L. (1996). *Programming Perl*, 2nd edit., O'Reilly, Sebastopol, CA.
47. Ouzounis, C. A. & Karp, P. D. (2000). Global properties of the metabolic map of *Escherichia coli*. *Genome Res.* **10**, 568–576.
48. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2000). GenBank. *Nucl. Acids Res.* **28**, 15–18.
49. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

50. Teichmann, S. A., Chothia, C., Church, G. M. & Park, J. (2000). Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics*, **16**, 117–124.

51. Orwant, J., Hietaniemi, J. & Macdonald, J. (1999). *Mastering Algorithms with Perl*, 1st edit., O'Reilly, Sebastopol, CA.

***Edited by G. von Heijne***

*(Received 30 November 2001; received in revised form 19 February 2002; accepted 19 February 2002)*