doi:10.1006/jmbi.2001.4912 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **311**, 693–708

JMB

The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in *Escherichia coli*

Sarah A. Teichmann[1]*, Stuart C. G. Rison[1], Janet M. Thornton[1,2] Monica Riley[3], Julian Gough[4] and Cyrus Chothia[4]

[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*

[3]*Josephine Bay Paul Centre for Comparative Molecular Biology and Evolution, 7 MBL St. Woods Hole, MA 02543-1015, USA*

[4]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 1TQ, UK*

**Corresponding author*

The 106 small molecule metabolic (SMM) pathways in *Escherichia coli* are formed by the protein products of 581 genes. We can define 722 domains, nearly all of which are homologous to proteins of known structure, that form all or part of 510 of these proteins. This information allows us to answer general questions on the structural anatomy of the SMM pathway proteins and to trace family relationships and recruitment events within and across pathways. Half the gene products contain a single domain and half are formed by combinations of between two and six domains. The 722 domains belong to one of 213 families that have between one and 51 members. Family members usually conserve their catalytic or cofactor binding properties; substrate recognition is rarely conserved. Of the 213 families, members of only a quarter occur in isolation, i.e. they form single-domain proteins. Most members of the other families combine with domains from just one or two other families and a few more versatile families can combine with several different partners.

Excluding isoenzymes, more than twice as many homologues are distributed across pathways as within pathways. However, serial recruitment, with two consecutive enzymes both being recruited to another pathway, is rare and recruitment of three consecutive enzymes is not observed. Only eight of the 106 pathways have a high number of homologues. Homology between consecutive pairs of enzymes with conservation of the main substrate-binding site but change in catalytic mechanism (which would support a simple model of retrograde pathway evolution) occurs only six times in the whole set of enzymes. Most of the domains that form SMM pathways have homologues in non-SMM pathways. Taken together, these results imply a pervasive ''mosaic'' model for the formation of protein repertoires and pathways.

Keywords: gene duplications; metabolic pathways; protein families; domain architecture; hidden Markov models

Introduction

The idea of chromosomal and gene duplications as a general source for new genes was proposed more than 60 years ago.[1,2] Horowitz[3] proposed a retrograde model of pathway evolution, in which enzymes evolve backwards from the protein that produces the final product. Subsequently, he suggested that this evolution occurred through gene duplications of the proteins within a pathway.[4] Jensen[5] showed that enzyme recruitment across pathways could occur by duplicated enzymes conserving their catalytic functions but evolving different substrate specificities. The changes in sequence and structures that produce proteins with different specificities were seen in atomic detail in the first protein structures.[6] The early protein structures also showed how different combinations of duplicated domains have produced enzymes with different activities.[7] More recent studies have described how mutations in

Abbreviations used: SMMP, small molecule metabolic pathways; EC, enzyme classification; HMM, hidden Markov model; PDBD, domain definitions in the structural classification of proteins database; SCOP, structural classification of proteins database; SMM, small molecule metabolic.

E-mail address of the corresponding author: sat@mrc-lmb.cam.ac.uk

0022-2836/01/040693–16 $35.00/0

active-site residues produce new catalytic properties for enzymes and, hence, the formation of new pathways.[8–11]

Until now, investigations of pathways have been limited to particular protein families, a single pathway, or few pathways. To begin to answer general questions about how a large set of related pathways are structured and have evolved, we have analysed all the pathways involved in the small molecule metabolism of *Escherichia coli*. This bacterium is a free-living organism and, therefore, has a set of the small molecule metabolic pathways sufficient for independent life. Similar, if not identical, sets of pathways are believed to exist in all free-living bacteria and eukaryotes. The very extensive experimental work that has been carried out on *E. coli*, including the determination of its genome sequence, means that our knowledge of these pathways is probably close to complete.

Using sequence and structural information, we have obtained a detailed picture of the evolutionary relationships and recombinations of domains in 510 of the 581 enzymes that form the small molecule metabolic pathways (SMMP) in *E. coli*. With these data, we can answer general questions on the structural anatomy and evolution of the SMMP proteins, and have organised the text as follows: after an introduction to the *E. coli* SMMP and the methods used here, we describe: (i) the domain structure of the SMMP proteins; (ii) the number and size of the families to which these domains belong; and (iii) the extent to which different types of domains combine to form multidomain proteins.

Taken together, these descriptions form what can be called the structural anatomy of the SMMP. We then go on to analyse and discuss:
(i) distribution of family members within and across pathways;
(ii) the types of features that can be conserved in protein families;
(iii) the nature of the homologues that are found within pathways;
(iv) the nature of the homologues that have been recruited across pathways; and
(v) the extent to which the families that form the SMMP are unique to these pathways.

These results have implications for the evolution of the pathways and these are discussed in the final part of this work.

E. coli small molecule metabolic pathways

The SMMP in *E. coli* are described in the EcoCyc database.[12] In EcoCyc, the pathways are placed in one of 16 categories in a ''Taxonomy of Pathways'' on the basis of the type of small molecules the pathway synthesises or degrades, or on the basis of its general function. Examples of such classes include amino acid biosynthesis and energy metabolism. In all, there are 106 pathways or superpathways that vary in size from one to 37 genes; three-quarters of the pathways contain between two and ten genes.

The SMMP proteins are formed from the products of 581 genes. There are 12 whose sequence is unknown at present: they have been identified only from their genetic or biochemical characteristics, but their activity has not been linked to an *E. coli* gene or protein sequence. Thus, the number of different SMMP proteins for which sequences are available and are used here is 569.

The SMMP is not just a collection of individual pathways but a metabolic network. The description of the SMMP in terms of separate pathways means that the enzymes that occur at nodes in the complex network appear to be used repeatedly in different pathways. In the case of the *E. coli* SMMP: 427 proteins are active in just one pathway; 96 proteins are active in two pathways; 32 in three pathways; 12 in four pathways; one in five, and one in six pathways. When considering properties of pathways as such, the proteins that are active in more than one pathway can be seen as having ''virtual homologues'' (see below).

Determining the domain structure and homology of *E. coli* SMMP proteins

During the course of evolution, new proteins have been produced by gene duplication, divergence and, in many cases, recombination. Thus, to begin to understand the evolution of the SMMP proteins we need to know if they consist of one domain, or combinations of two or more domains, and the evolutionary relationships of these domains. Ideally, the evolutionary relationships and domain structure of proteins would be found by a direct comparison of their sequences. However, related proteins can diverge to such an extent that simple comparisons of their sequences fail to detect their relationships. Pairwise sequence comparison methods such as BLAST, FASTA and SSEARCH detect only one-half of the relationships that occur in sequences with identities of 20-30 % and, for proteins with lower identities, the proportion is much smaller.[13] The large majority of the different SMMP proteins have sequence identities well below 40 %. This means that, if a significant proportion of them are related to one another, it will be discovered only by the use of information that goes beyond that given by the simple pairwise comparison of their sequences.

There are two sources of additional information that can help overcome these limitations, at least in part. First, the sequence comparison methods that use multiple sequences, whilst still failing to find all distant relationships, are three times more effective than pairwise methods for proteins whose sequence identities are less than 30 %.[14] Second, on a different level, if the structures of the proteins being compared are known, both their domain structure and evolutionary relationships, even when distant, can usually be detected using the combination of structural, functional and sequence information. This means that, if the proteins in the *E. coli* metabolic pathways can be shown to be

homologous to proteins of known structure, we can infer both their domain structure and evolutionary relationships from what is known about the domain structure and evolutionary relationships of the homologues.

Information on the domain structure and evolutionary relationships of the proteins of known atomic structure is available from the structural classification of proteins (SCOP) database.[15] In this database, the unit of classification is the structural, functional and evolutionary unit of proteins: the domain. Small proteins, and most medium-sized proteins, consist of a single domain. The domains that form large proteins are classified individually in SCOP if there is evidence from known protein structures that they are evolutionary units that can undergo independent duplication and recombination.

The evolutionary relationships in SCOP are described on two levels: family and superfamily. The family level brings together domains whose sequence similarities imply an evolutionary relationship. The superfamily level brings together families whose structural and functional features imply an evolutionary relationship even though their sequence identities are low.[15] In the work described here, the distinction between these two levels is not significant and throughout we refer to both as just ''family'' relationships.

As described in Methods, using hidden Markov models (HMMs) of SCOP domains and structural information, we identified the nature and the evolutionary relationships of 695 domains in 487 SMMP proteins. Four-fifths of these proteins are completely, or nearly completely, covered by these assignments and one-fifth are partially matched in that they also have an unassigned region of 75 or more residues. In addition, the sequence matches made between 27 domains from (i) the unmatched regions in four proteins partially matched by structural information and (ii) 23 other proteins, clustered them into 11 families (see Methods below for more details of these calculations).

Putting together the 487 *E. coli* proteins whose domains are defined on the basis of structural information, and the 23 whose domains are defined by sequence comparisons, means that we have information on the evolutionary relationships for 510 of the 581 different proteins that form the small molecule metabolic pathways in *E. coli*, i.e. 88 % of the total number†. In terms of pathways, 71% of the 106 pathways have a structural assignment for at least four-fifths of the enzymes, and 44% have a structural assignment for every single enzyme.

A summary of the most important numbers and results is given in Table 1.

† A detailed description of the enzymes discussed here is given at http://www.biochem.ucl.ac.uk/ ~ sat/ec_metpath.html

Table 1. Summary of protein and pathway data

Number of metabolic pathways	106
Number of proteins	581
Number of proteins of known sequence	569
Number of proteins with assigned domains	510
Structural domains	695 in 202 families
Sequence domains	27 in 11 families

Domain structure of *E. coli* SMMP proteins

The matches made to SMMP proteins by the HMMs and the sequence comparisons give either the exact number of domains of which the proteins are composed, or allow an estimate of this number. As shown in Table 2A, of the 510 matched sequences, there are 271 where a single domain very largely covers the whole of the sequence, i.e. it leaves less than 75 residues unmatched at the N or C termini. In most of these cases, the unmatched sections are much shorter than 75 residues. There are another 128 SMMP proteins that are fully covered by two, three, four, five or six domains.

The remaining 111 matched proteins are partly covered by between one and four domains; i.e. they have an unmatched region of 75 residues or more indicating the presence of one, or, in the case of much longer regions, more unmatched domains (Table 2A). A rough estimate of the number of domains in the unmatched sequences can be made knowing that the average size of a domain in the SCOP database[15] is 175 residues, and assuming that unmatched regions of 75-260 residues corresponds to one domain; one of 260-440 correspond to two domains, etc. Using the same procedure, a rough estimate can also be made of the number of domains in the 59 SMMP proteins of known sequence without any assigned domains. The results of this calculation are also given in Table 2A and show that overall, close to half of SMMP proteins contain one domain, a third contain two domains and one-sixth contain three to six domains.

Protein families that form the *E. coli* SMMP

Families of protein domains

As described in Methods, the domains we have identified in the SMMP proteins can be clustered into families on the basis of their evolutionary relationships. We found 695 domains with structural information, which belong to one of 202 different families. The 27 domains clustered by sequence comparisons give another 11 different families. Thus, in total, the 722 domains we have identified in the 510 SMMP proteins come from 213 different families. The average size of these families is $722/213 = 3.4$.

The sizes of individual families have an exponential character: there are few large families and many small families (Table 2B). The total member-

Table 2. Domains and families

A. *The number of domains in SMMP proteins*

Number of domains (n)	Number of sequences completely matched by n domains	Number of sequences partly matched by n domains	Partially matched sequences: estimated number with n domains	Unmatched sequences: estimated number with n domains	Total: all proteins
1	271	77	0	41	312
2	96	26	55	14	165
3	28	5	36	3	67
4	2	3	8	0	10
5	1	-	9	1	11
6	1	-	3	-	4
Total no. proteins	399	111	111	59	569

B. *Number and size of protein familes*

Family size (n)	Number of families of size n	Family size (n)	Number of families of size n
1	74	10	1
2	61	11	3
3	24	12	2
4	17	13	1
5	4	14	1
6	9	19	2
7	5	20	1
8	3	21	1
9	3	53	1

ship of the largest 33 families, which have between six and 51 members, is slightly larger than the total membership of the other 180 families, which have between one and five members. Full descriptions of each family and its members can be found at the accompanying web site.

Domain combinations in E. coli SMMP proteins

The previous section has shown that close to a half of the SMMP proteins are made of combinations of domains. Here, we describe the extent to which different families that form these proteins make different kinds of combinations. We calculated, for each family, the number of other families from which its members draws combination partners (Table 3A). There are 57 families whose members are always in isolation, i.e. they form only one-domain proteins. Members of another 141 families may occur in isolation but most members of these families form combinations with other domains of known identity or with homologues of themselves. There are also 15 families that occur in proteins in which they are adjacent to domains of unknown character, i.e. regions that have not been assigned sequence or structural domains.

Of the families that combine with other domains, the large majority combine with domains from only one or two other families, but a few families are more versatile (Table 3A). The family that makes the largest number of different kinds of combinations, that of the Rossmann NAD-binding domains, combines with domains from 12 different families. The next 11 largest families combine with partners from between three and six different families, see Table 3B†.

If domains from two families combine, they do so in the same N to C, orientation in 99 of the 103 different types of pairwise combinations. There are only four exceptions to the rule: three exceptions involve three proteins that combine in an ABA or ABAB fashion, and hence the families A and B occur next to each other in both ways within one sequence. The fourth exception is aconitate hydrase I and aconitate hydrase B, where the C-terminal domain in aconitate hydrase I is at the N terminus of aconitate hydrase B as shown in Figure 5(a).

Families of whole proteins

The previous two sections have been concerned with the families formed by domains and the combinations that the domains make with each other. Of course, most new proteins are produced without the involvement of recombination: just by the simple duplication of whole proteins that have one or more domains. We examine the set of 399 SMMP proteins completely matched by structure or sequence domains to determine the extent to which they have arisen by the simple duplication whole proteins. There are: 265 single-domain proteins that belong to 59 families; 55 two-domain proteins in 17 families; and 16 three-domain proteins in six families.

Thus, 336 proteins form 82 whole protein families. These 336 are 84% of the completely matched sequences and 254 (=336 − 82) of them were produced by simple gene duplications.

Table 3. Domain combinations

A. *The number of different families from which partners are drawn*

The number n of different families from which partners are drawn	Number of families whose members can be linked to n different families
0	57
1	103
2	26
3	7
4	3
6	1
12	1
Adjacent regions of unknown domain type	15
Total no. families	213

B. *Protein families that have many partner families*

Protein family	Family size	Number of partner families	Families N or C-terminal to the protein
1. NAD(P)-binding Rossmann domain	51	12	N:7 C:5
2. Glutathione synthetase ATP-binding domain	11	6	N:4 C:2
3. Thiamine diphosphate-binding fold (THDP-binding)	21	4	N:1 C:2 N + C:1
4. Regulatory domain in the amino acid metabolism	10	4	N:2 C:2
5. Class I glutamine amidotransferases (GAT)	8	4	N:2 C:2
6. β-Galactosidase/glucuronidase domain	3	3	N:1 C:1 N + C:1
7. FAD/NAP(P)-binding domain	19	3	N:3
8. Copper amine oxidase, domains 1 and 2	2	3	N:1 C:2
9. 4Fe-4S ferredoxins	5	3	N:2 C:1
10. Glycosyltransferases	7	3	N:1 C:1 N + C:1
11. N-terminal nucleophile amidohydrolases	4	3	N:3
12. Cobalamin (vitamin B12)-binding domain	2	3	N:1 C:2

The distribution of family members within and across pathways

The distribution of individual domains

Families with more than one member can have homologues in different pathways, within the same pathway or a combination of both (see Figure 1).

Description of the distribution of homologues is complicated by the quarter of SMMP proteins (146 out of 581) that are active in two to six pathways; see above and Table 4. This means that a family may function in different pathways not just through the use of different homologues but also through the multiple use of a particular member. Thus, a protein that functions in n pathways can be seen has having $(n-1)$ ''virtual homologues''. The sequence matching calculations described above identified 201 domains in 131 of the 142 proteins that function in more than one pathway. If we count the number of times these domains are used in different pathways we find that the 201 domains have 304 virtual homologues (see Table 4).

To take these virtual homologues into account in describing the distribution of domains within and between pathways, we add the virtual homologues to the total number of true domains to give a total of $304 + 722 = 1026$ effective domains. They come from 213 families and this means that

Table 4. Distribution of the domains in the SMMP proteins that function in more than one pathway

Number of pathways (n)	Number of proteins active in n pathways	Number of these proteins with assigned domains	Number of domains in the matched proteins	Virtual homologs: number of domains in the other $(n-1)$ different pathways
2	96	81	131	131
3	32	30	43	86
4	12	12	23	69
5	1	1	2	8
6	1	1	2	10
Total	142	131	201	304

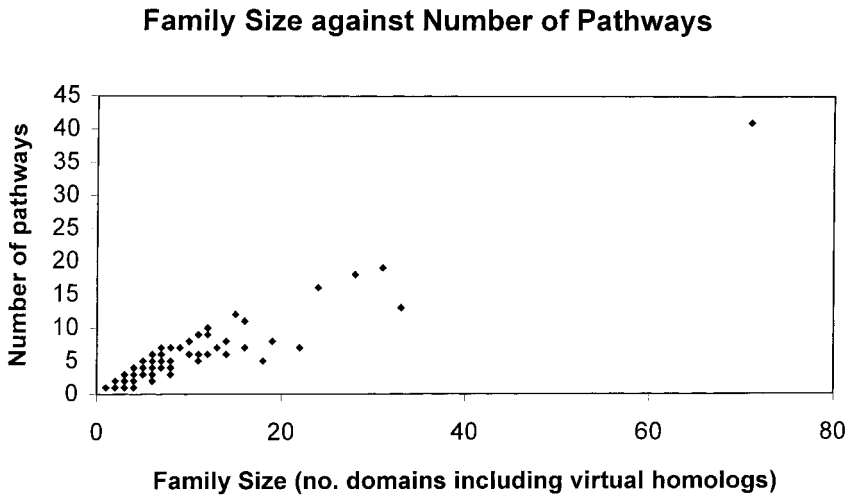

Figure 1. This graph shows the size of a family in number of domains (double-counting proteins that occur in multiple pathways) against the number of pathways in which the domains of a family are involved. Sixty-nine small families are present in one pathway, with 47 families of size one, 20 families of size two, one of size three and one of size four in one pathway. All the other 144 families are present in more than one pathway, underlining the importance of recruitment.

$1026 - 213 = 813$ are either true homologues or virtual homologues of the other 213.

Examination of where the 813 homologues occur shows that 506 are in different pathways and the remaining 307 are in the same pathway as another member of the family or a virtual homologue. Thus, in the families that have more than one member, most members are involved in recruitment across pathways. This can also be seen in the data shown in Figure 1, where the size of the domain families, including virtual homologues, is plotted against the number of pathways in which they occur: 144 families, with 932 domains, are represented in more than one pathway and most of them (123) have between 50 and 100 % of their members are in different pathways. For example, if a family of four domains has two members in one pathway and one in each of two other pathways, 50 % of the family will be in different pathways.

The members of 69 families are limited to one pathway. All these families are small: 67 have one or two members, and altogether they have 94 members. Of the homologous pairs of enzymes within pathways, over half are isozymes.

The distribution of the members of whole protein families

In the previous section we discussed the distribution of members of domain families. These individual domains exist in isolation, in one-domain proteins, or combined with partners to form proteins with two or more domains. We have described above the families formed by just the duplication of whole proteins. Here, we describe the distribution of members of these families within and between pathways. The proteins that are completely matched by multiple domains are either two or three-domain proteins. The 17 families of two-domain proteins have 13 homologues within pathways and 25 across pathway. Out of the 16 three-domain proteins that belong to six three-domain families, there are four proteins that have homologues within pathways and six across pathways. The families of multi-domain proteins with the same domain architecture therefore exhibit the same trend as individual domains in having more homologues across than within pathways.

Types of conservation within families

To be able to discuss the role of family members within and across metabolic pathways, we need to define the functional roles they perform. The proteins usefully produced by duplication and divergence nearly always retain some functional aspect(s) of their precursors and modify or change others. This means that protein families can be classified in terms of the functional features that they conserve. We can define a number of different types of conservation: (1) conservation of chemistry, which occurs when they retain the same or a closely related catalytic mechanism; (2) conservation of a binding site for a main substrate; (3) conservation of a binding site for a cofactor or minor substrate.

The nature of the conservation that occurs in different families can be determined in many instances by considering the Enzyme Classification ommission[16] (EC) number of the reaction catalysed and inspecting the substrates and products, and their positions in the pathways. This information, as well as information on complexes and isozymes, is contained in the EcoCyc database. If at least the first two EC numbers are conserved for the reactions catalysed by a pair of enzymes (assuming both have been assigned an EC number), we classify the duplication as conserving chemistry. We can

make this assumption because we are looking at homologues that belong to the same family in all cases, as opposed to considering only EC number when comparing proteins that are not necessarily homologous. There are exceptions to the connection between conservation of chemistry and EC number in homologous families, so that if EC numbers are not conserved, a more detailed inspection of the reactions and the proteins is required: sometimes chemistry is conserved while EC numbers are not.[17,32] If substrates or cofactors are similar, the two enzymes are classified as conserving their main substrate-binding site or a cofactor or minor substrate-binding site. If two enzymes belong to the same family and catalyse the same reaction at the same point in the same pathway, they are considered to be isozymes.

Homologues within pathways

Conservation of function within pathways

There are 56 families that have more than one protein in the same pathway (excluding isozymes). Of these families, 17 conserve their chemistry, 27 modify their chemistry and 12 conserve a cofactor or minor substrate-binding site. Some of these families also conserve the main substrate-binding site and these are discussed in the next section. All but 11 of the 56 families also have homologues in different pathways.

Pairs of homologues that form consecutive steps in pathways

One might expect duplications within pathways to be present in enzymes in consecutive reactions, or one step apart, because the substrates of such enzymes are usually similar.[3,4] In fact, of 445 consecutive pairs of enzymes that occur in the SMMP, there are only 26 (6%) pairs that contain domains from the same family. The SMM pathways contain 340 triplets of consecutive enzymes. Amongst these there are 37 (11%) cases where first and third enzyme have an homologous domain, as depicted in Figure 2(a). Thus, homologous domains in proteins carrying out consecutive reactions are not only uncommon, they are less common than in enzymes one catalytic step apart.

Triplets of reactions in which all three enzymes have at least one domain from the same family are very rare, with only two out of 340 triplets exhibiting this feature. Indeed, the two triplets are actually one quadruplet of consecutive ligases in the peptidoglycan biosynthesis pathway (see below). There is also a triplet of $(\alpha/\beta)_8$ barrels in three consecutive enzymes in tryptophan biosynthesis, but the first two barrels are part of the same protein, and this case will be discussed below.

These results show that gene duplications that conserve substrate-binding properties yet diverge in catalytic mechanism have played a very minor role in the formation of consecutive steps in the SMMP and that, instead, recruitment usually takes

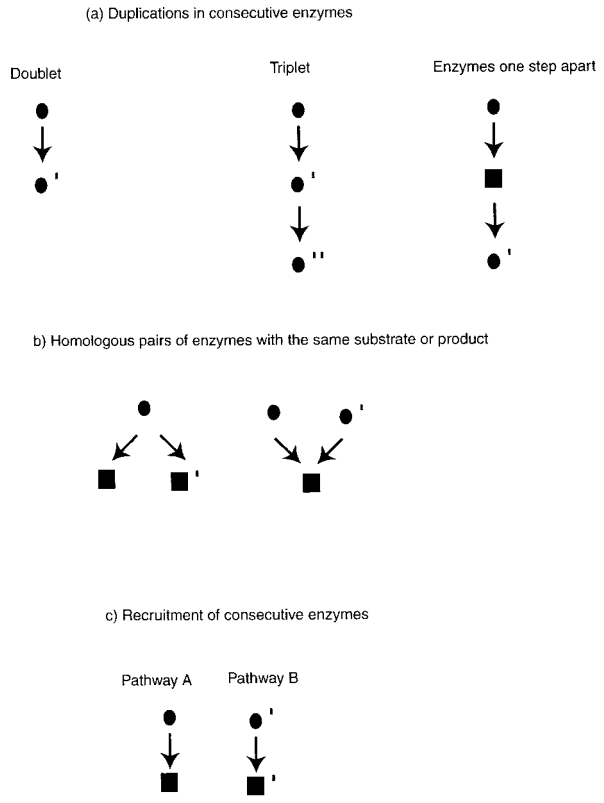

Figure 2. Duplications within and across pathways. The shapes represent enzymes, and enzymes of the same shape, distinguished by apostrophes, are homologous. (a) Duplications in consecutive enzymes: the examples show duplications in enzymes adjacent or enzymes separated by an intermediate enzyme in a metabolic pathway. (b) Homologous pairs of enzymes (''parallel'' enzymes) with the same substrate or product. (c) Serial recruitment of enzymes: homologous doublets are shown.

place from other pathways, or from enzymes in the same pathway, based on other criteria, such as the chemistry of catalysis.

Of the 26 consecutive pairs of enzymes with one domain from the same family, 13 conserve a cofactor or minor substrate-binding site and eight conserve a catalytic mechanism. The other five conserve the ligand-binding site of the main substrate and change the catalytic mechanism (Table 5). There are two pairs of enzymes where this type of conservation also occurs although they are not consecutive enzymes. Two of these examples occur within one pathway. This means that conservation of the main substrate-binding site with change in catalytic mechanism in enzymes close in a pathway occurs in only five of the 106 pathways.

Two of the six are well-known examples: the $(\alpha\beta)_8$ barrels trpC and trpA in tryptophan biosynthesis,[18] and hisA and hisF in histidine biosynthesis.[19] trpC is a bifunctional enzyme consisting of two $(\alpha\beta)_8$ barrels, one of which is *N*-(5′ phosphoribosyl)anthranilate isomerase and the other indole-3-glycerolphosphate synthase. trpA is the α-subunit of tryptophan synthase. The two genes are part of the trp operon and are one gene apart on the *E. coli* chromosome. hisA and hisF are also adjacent on the *E. coli* chromosome and one may well be a direct duplicate of the other.

The four other cases listed in Table 5 are more complex than the two described above. In three of the four cases, the first EC number is conserved, so the reactions are not as different as those catalysed by trpC/trpA or hisA/hisF. Also, none of the genes are close to each other on the *E. coli* chromosome. In two of the four cases, the enzymes are not consecutive, but either ''parallel'' or one step apart. In fermentation, the pyruvate kinase isozymes pykA and pykF act on phosphoenolpyruvate in an EC class 2 reaction, as does the phosphoenolpyruvate carboxylase, ppc, in an EC class 4 reaction. These ''parallel'' enzymes belong to the same family of $(\alpha\beta)_8$ barrels. There are three phosphoribosyltransferase enzymes in histidine, purine and pyrimidine biosynthesis that are related: amidophosphoribosyl transferase (purF) and orotate phosphoribosyltransferase (pyrE) both act on the substrate PRPP and follow the enzyme phosphoribosylpyrophosphate synthase (prsA). In deoxypyrimidine nucleotide/nucleoside metabolism, dCTP deaminase (dcd) is followed by the related dUTP pyrophosphatase (dut). Finally, there are two members of the inosine monophosphate dehydrogenase $(\alpha\beta)_8$ barrel family in nucleotide metabolism: IMP dehydrogenase (guaB) and GMP reductase (guaC), which are one step apart, as the GMP synthase (guaA) reaction is in between.

Homologous pairs of enzymes with the same substrate or product

When pairs of enzymes catalyze reactions that produce the same or similar products, so that they are both succeeded by the same enzyme, the pair of enzymes share a homologous domain in nine out of 56 such cases, but in two of the cases only one out of many domains is shared. An example of this type of scenario is given in Figure 3: fucA and rhaD in fucose and rhamnose catabolism. The seven cases are described in Table 6A.

Conversely, there are also cases where pairs of different enzymes have the same or similar substrates but produce different products (Figure 2(b)). The two enzymes can either carry out related reactions in these cases or quite different reactions. There are eight out of 48 such cases as described in Table 6B. Most of the pairs of enzymes described here are in the same EcoCyc pathways, but some, such as lyxK and rhaB are not in the same EcoCyc pathway, although they are preceded by the same enzyme.

Table 5. Conservation of the main substrate-binding site with change in reaction catalysed within a pathway

Superfamily and pathway	Enzymes	Comments
Phosphoenolpyruvate/pyruvate $(\alpha\beta)_8$ barrels in fermentation	pykF/pykA, ppc	The pykA/pykF and ppc both have phosphoenolpyruvate as their substrate and belong to different EC classes
Ribulose-phosphate binding $(\alpha\beta)_8$ barrels in tryptophan biosynthesis[a]	trpA, trpC	Consecutive enzymes in different EC classes
P-binding α/β barrels in histidine, purine and pyrimidine biosynthesis	hisA, hisF	Consecutive enzymes in different EC classes
Phosphoribosyltransferases (PRTases) in histidine, purine and pyrimidine biosynthesis	prsA, purF and prsA, pyrE	Consecutive pairs of enzymes in EC class 2 (transferases)
dUTPase domains in deoxypyrimidine nucleotide/nucleoside metabolism	dcd, dut	Consecutive enzymes in EC class 3 (hydrolases)
Inosine monophosphate dehydrogenase $(\alpha\beta)_8$ barrels in nucleotide metabolism	guaB, guaC	Enzymes one step apart in EC class 1 (oxidoreductases)

These examples are the only detected cases of enzymes that belong to the same family and share a similar binding site for the main substrate within a pathway, but change their reaction chemistry. Therefore, this type of conservation is much rarer than change in substrate specificity with conservation of chemistry in metabolic pathways.

[a] The P-binding α/β barrels are a diverse family of α/β barrels that are likely to be related, as they share a phosphate-binding site in the loop between beta-strand 7 and alpha-helix 7 and the N terminus of an additional helix 8′.

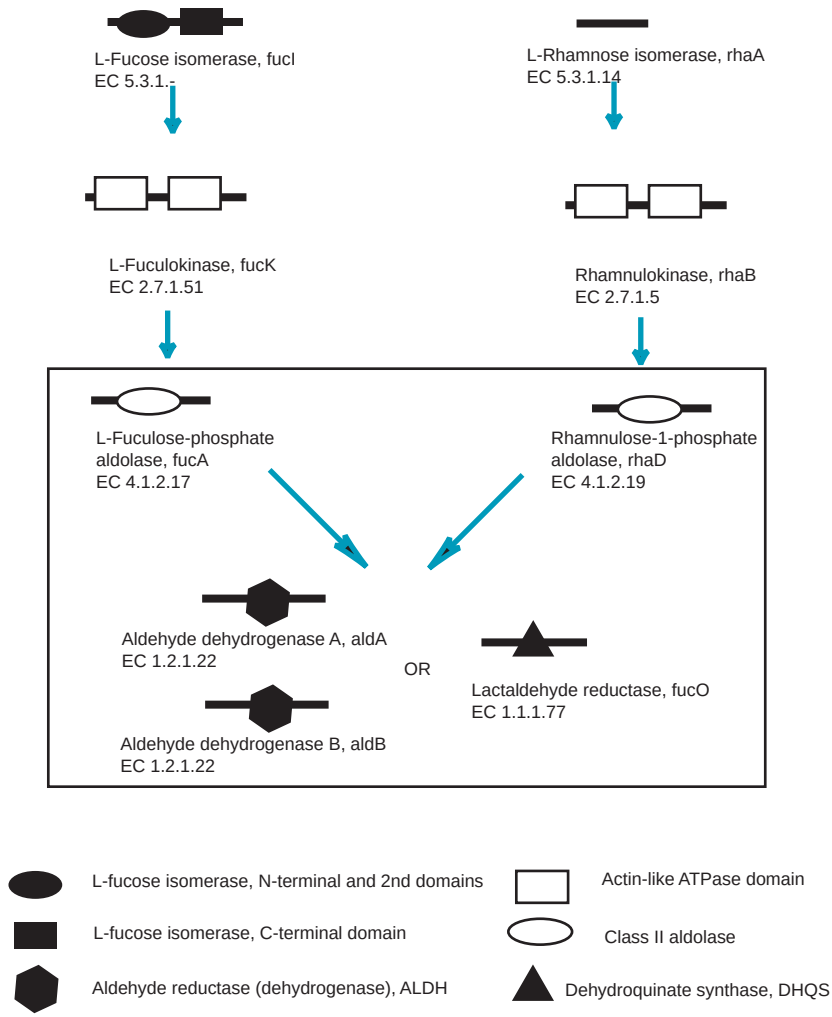

Figure 3. Fucose and rhamnose catabolism. In fucose and rhamnose catabolism, a superpathway in EcoCyc that consists of the fucose catabolism and rhamnose catabolism subpathways, there is an example of serial recruitment and an example of ''parallel'' enzymes, which are boxed. Serial recruitment has occurred because fucK (L-fuculokinase) is homologous to rhaB (rhamnulokinase), and fucA (L-fuculose-phosphate aldolase) is homologous to rhaD (rhamnulose-1-phosphate aldolase). fucA and rhaD have the same product, and are both followed by aldA/aldB or fucO, and are thus ''parallel'' enzymes.

For consecutive enzymes that are expected to have some similarity in their substrate-binding sites, we saw that there was little bias towards homology. In cases where enzymes have (near) identical substrates or products, the fraction of homologous enzymes is somewhat higher: 13 % and 17 % for the two scenarios described above.

Pathways with a high proportion of homologues

The existence of homologues within pathways is of special interest as it might be expected because of the similarities of substrates in consecutive steps in pathways and it can potentially support the Horowitz[3] model of retrograde evolutionary of pathways.

Taking together all the duplications within pathways (but excluding isozymes and internal duplications), we can establish by a simple statistical test whether there are particular pathways where the levels of duplication are unusually high. In this test, the domains in our set, amplified to include virtual homologues, are distributed across the 106 pathways at random 10,000 times, and the observed duplication levels are compared to the random distribution. Thus an expectation value can be calculated for the observed duplication level for each pathway size measured in number of domains.

Eleven out of 106 pathways have probability of 1 % or less for the high level of duplication in the pathway to occur by chance and these are listed in Table 7. Three of the 11 sets of enzymes are labelled as pathways, but in fact they are simply lists of similar single redox reactions for several electron donors or acceptors in aerobic and anaerobic respiration. Thus the occurrence of many iron-sulphur or nickel-iron centres in these three is to be expected. In the other eight pathways, the reasons for the high proportion of homologues are varied.

Table 6. Homologous pairs of enzymes that act on the same substrate or produce the same product

A. *Homologous pairs that produce the same product*

Homologous pair	Subsequent enzyme(s)	Homologous domains
Different substrates, similar reactions, same product		
malP, glgP	malZ	β-Glucosyltransferase & glycogen phosphorylase
fucA, rhaD	aldA/aldB or fucO	Class II aldolase
firD, mltD	pfkA/pfkB	NAD(P)-binding Rossmann fold domain
Same substrates, different cofactors, similar reactions, same product		
maeA, maeB	ppsA[a]	Amino acid dehydrogenase domain and NAD(P)-binding Rossmann domain
Different substrates, similar reactions, different products: subsequent enzyme uses several substrates simultaneously or has multiple substrate specificity		
thiD, thiM	thiE	Ribokinase-like domain
tmk, cmk	ndk	P-loop nucleotide triphosphate hydrolase domain
serA, pdxB	serC	Formate/glycerate catalytic domain and NAD(P)-binding Rossmann domain

B. *Homologous pairs that act on the same substrate*

Preceding enzyme(s)	Homologous pair	Homologous domains
Same substrate, similar reactions		
aroC	trpDE, pabAB	trpE/pabA: anthranilate synthase, aminodeoxyisochorismate synthase/lyase subunit trpD/pabB: class I glutamine amidotransferases
Same main substrate, different cofactor/minor substrate, similar reactions		
glcB/aceB	maeA, maeB	Amino acid dehydrogenase domain and NAD(P)-binding Rossmann domain
carAB	pyrlB, argF and argI	Two aspartate/ornithine carbamoyltransferase domain
Same substrate, different reactions		
eno[a]	pykA and pykF, ppc	Phosphoenolpyruvate/pyruvate domain
purF, pyrE	prsA[a]	Phosphoribosyltransferase domain
Preceeding enzyme with multiple substrate specificity: different substrates, similar reactions		
uxaC	uxaB, uxuB	NAD(P)-binding Rossmann domain
deoD	gpt, apt	P-loop nucleotide triphosphate hydrolase domain
rhaD	lyxK, rhaB	Two actin-like ATPase domains

[a] See Table 4.

We describe the nature of the duplications in these pathways below.

Phosphatidic acid and phospholipid biosynthesis. There is one sequence family with two members, plsB and plsC, which are both acyltransferases, and one family (phospholipase D/nuclease superfamily in SCOP) with two members, cls and pssA, which are transferases of a similar type. pssA and cls are also significantly similar at the sequence level. In summary, this pathway contains two families, both related at the sequence level, in which the catalytic mechanism is conserved.

Colanic acid biosynthesis. The high level of duplication is primarily due to a family of four Rossmann domain proteins which are all related at the sequence level: galE, ugd, fcl and gmd. In addition, there is a pair of sequence-related proteins that both belong to the nucleotide-diphospho-sugar transferase SCOP superfamily. That many of the relationships in the two pathways discussed so far are detectable at the sequence level suggests either that these duplications are recent, or that the enzymes are subject to constraints on their structure and sequence that prevent further divergence.

Nucleotide metabolism. Since most of the reactions in this pathway involve transfers of phosphate groups, there are three members of the nucleotide triphosphate hydrolyse superfamily, adk, gmk and purA, as well as three members of the purine and uridine phosphorylase superfamily, amn, xapA, deoD. In addition, there are two NAD(P)-linked oxidoreductases, guaB and guaC, and three phosphoribosyltransferases, hpt, gpt and apt. In all of these four families in nucleotide metabolism, the type of conservation is conservation of chemistry.

3-Deoxy-D-manno-octulosonate, peptidoglycan and lipid A-precursor biosynthesis. The four consecutive ligases murC, murD, murE and murF all have catalytic domains whose homology is apparent at the sequence level, as well as a glutamate ligase domain. These four consecutive enzymes have similarities in their catalytic mechanism and also act on similar substrates. There are three acetyl/acyltransferases, glmU, lpxA and lpxD, which share a trimeric β-helix domain. glmU also has a sequence domain in common with kdsB, the CPM-KDO synthetase, which probably harbours its *N*-acetylglucosamine-1-phosphate uridyltransferase activity. A further sequence family with conservation of chemistry encompasses lpxB, dktA and murG, which are all transferases with similar

Table 7. Pathways with high levels of duplication

Pathway	*P* value	Duplication level	No. of domains in pathway	No. of superfamilies in pathway
Nucleotide metabolism	0	0.35	17	11
Histidine, purine and pyrimidine biosynthesis	0	0.35	42	27
KDO, peptidoglycan and lipid-A precursor biosynthesis	0	0.36	30	19
Phosphatidic acid and phospholipid biosynthesis	0	0.4	5	3
Aerobic respiration, electron donors reaction list	0	0.41	24	14
Anaerobic respiration, electron donors reaction list	0	0.55	27	12
Anaerobic respiration, electron acceptors reaction list	0	0.58	31	13
Gluconeogenesis	0.01	0.30	20	14
Colanic acid biosynthesis	0.01	0.30	13	8
Glycogen catabolism	0.01	0.33	6	4
Polyisoprenoid biosynthesis	0.01	0.50	2	1

P values were calculated as described in the text, and duplication levels for each pathway were calculated using the expression: (total number of domains in pathway − no. families in pathway)/total number of domains.

functions. The murC, murD, murE and murF and glmU, lpxA and lpxD relationships are all detectable at the sequence level, implying that this pathway could have evolved to a large extent by duplication within itself, and possibly more recently than most of the other pathways.

Polyisoprenoid biosynthesis. Two of the three enzymes in this pathway, ispA and ispB, are related at the sequence level as well as the structural level, sharing a terpenoid synthase domain.

Glycogen catabolism. Two pairs of the six enzymes in glycogen catabolism share similarity

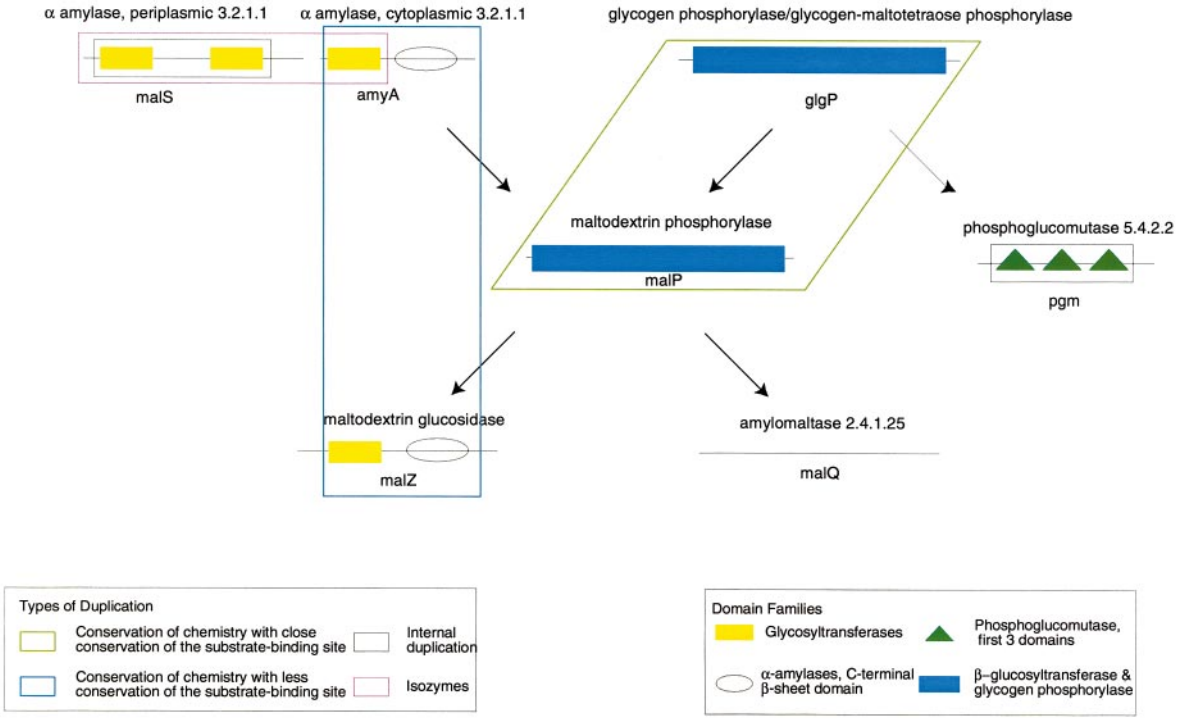

Figure 4. Glycogen catabolism. The glycogen catabolism pathway contains two duplications with conservation of catalytic mechanism. One is in consecutive enzymes (glgP/malP), so with a close conservation of substrate-binding site as well, while the other duplication occurs for enzymes one step apart (amyA/malZ), with less conservation of the substrate-binding site. There are also internal duplications, where the same type of domain occurs several times in one polypeptide sequence (malS and pgm) and isozymes (malS and amyA).

detectable at the sequence and structural level: malS and malZ both have a glycosyltransferase domain, and glgP and malP are both phosphorylases that have a β-glucosyltransferase and glycogen phosphorylase domain, as shown in Figure 4.

Gluconeogenesis. In gluconeogenesis, there are four proteins with Rossmann domains, so the type of conservation is of a cofactor-binding domain. In addition, there is a sequence family of two types of malic enzymes, one $NADP^+$-linked (maeB) and one NAD-linked (sfcA).

Histidine, purine and pyrimidine biosynthesis. There are two examples of conservation of the substrate-binding site with change in catalytic mechanism: phosphoribosylpyrophosphate synthase (prsA) and amidophosphoribosyl transferase (purF), which are consecutive enzymes and are also related to pyrE, and hisA and hisF, both consecutive enzymes whose relationship is detectable at the sequence level, but which carry out different types of reactions. There are several families with conservation of a minor substrate-binding site: there are four proteins that have glutathione synthetase ATP-binding domains and biotin carboxylase N-terminal domains (carB, purK, purT, purD) and four proteins with class I glutamine amidotransferase domains (purL, hisA, carA, guaA), of which two (guaA and carA) are sequence-related. In addition, the six-domain protein carB shares a methylglyoxal synthase-like domain with purH.

From this description of the eight pathways with a significantly high level of homologues within pathways, it is obvious that there are many examples of conservation of chemistry, which occur when there is a requirement for the same type of catalysis several times within a pathway. Conservation of cofactor-binding domains is also common. But homologues that conserve the main substrate-binding site and change the catalytic mechanism are extremely rare.

There are two types of duplications within pathways that have not been discussed extensively here; duplications within polypeptide chains and isozymes. These types of duplications are almost as common as conservation of cofactor or chemistry, but cannot take place across pathways by definition, and so have not been considered in detail here.

Homologues in different pathways

Conservation of function across pathways

There are 114 families that have different members in more than one pathway excluding virtual homologues. In 40 of these families there is fairly close conservation of catalytic mechanism with the first two or more EC numbers being shared in all members of the family. Another 13 families conserve their cofactor-binding function.

The remaining 61 families are mostly enzymes that are variable in the types of reactions they catalyse and substrates they act on; they conserve only the first or no EC number. Despite this variability in the EC numbers of the family members, major aspects of the catalytic mechanisms are known to be conserved in some families and this probably occurs in many of these families.[17] Also, for a number of enzyme families, previous studies have described the actual changes in molecular structure that modify a few crucial features of the active site to create a different, though related, catalytic activity.[8–11,17,20] This means, of course, that these modified homologues have different EC numbers.

The extent of serial recruitment of proteins across pathways

Having established above that recruitment of individual domains and domain combinations across pathways is very widespread, we now consider the extent to which consecutive proteins have recruited from one pathway to another. Serial recruitment would be expected if, for instance a chunk of a chromosome, such as an operon, were duplicated, and the duplicated enzymes were all recruited to form a new pathway. We can start to investigate this by looking at whether there are homologous doublets or triplets of consecutive enzymes in different pathways, as shown in Figure 2(c).

As mentioned above, there are 445 different sets of two consecutive reactions involving enzymes of known sequence in our set of 106 different pathways. In some cases, a reaction may be catalysed by one or more polypeptide chains. Also, a pair of reactions may be carried out by different regions of one polypeptide chain.

We consider two pairs of consecutive enzymes to be homologous, if both the first and second reactions in each pair have at least one homologous domain in the enzyme(s) catalysing that reaction. (The proteins involved must be non-identical.) If we exclude the pairs that share only one of the common nucleotide-binding domains such as Rossmann domains, P-loop nucleotide triphosphate hydrolases and PLP-dependent transferases, and also cases where only one domain is shared between multi-domain proteins with different domain architectures, there are only a very small number of genuine candidates for serial recruitment, which we describe below and in Figures 3 and 5(a) and (b).

If the two pairs of enzymes that are homologous are also both close to each other on the *E. coli* chromosome, this is additional evidence for serial recruitment by duplication of a chunk of the chromosome. We were able to find only two such examples, and one of these is shown in Figure 3. The kinase and aldolase pairs (fucK and fucA as well as rhaB and rhaD) in fucose and rhamnose catabolism (Figure 3) are each one gene apart on the *E. coli* chromosome, and are homologous and

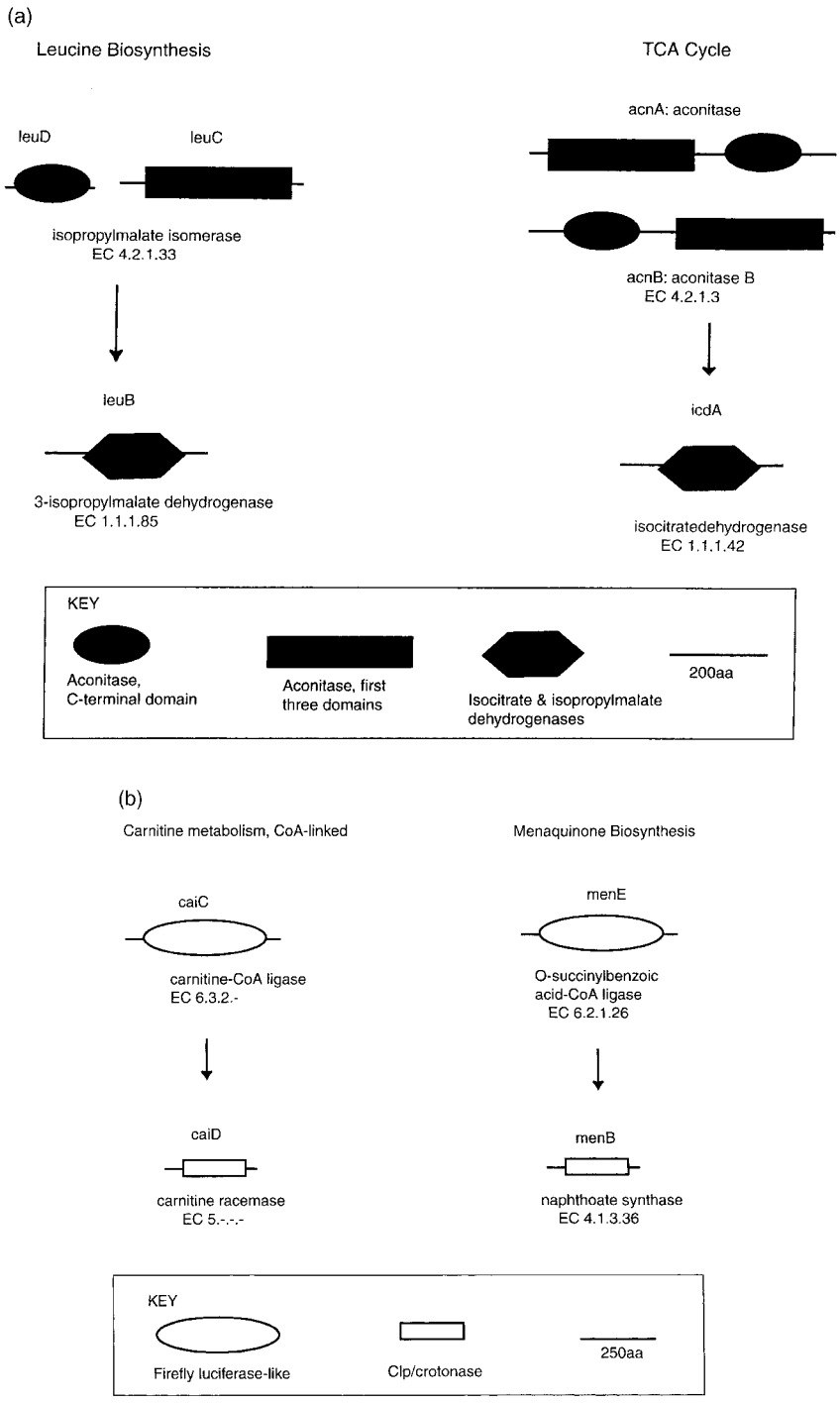

Figure 5. Homologous pairs of consecutive enzymes. These two examples are cases in which the two first and two second enzymes in different pathways have homologous domains. In both cases, the homology is detectable using sequence and structural domains. In leucine biosynthesis, leuC and leuD are subunit of one enzyme, while acnA and acnB in the TCA cycle are isozymes.

located in the same way as araB and araD in arabinose catabolism (not shown). The second example of this are the two pairs of enzymes in Figure 5(b): the enzymes in carnitine metabolism and menaquinone biosynthesis are homologous to each other, and the two enzymes in carnitine metabolism are adjacent on the *E. coli* chromosome, while the two enzymes in menaquinone biosynthesis are one gene apart on the chromosome.

In the example given in Figure 5(a), one of the enzyme pairs is adjacent, while the other is not. The enzymes in Figure 5(a) are in leucine biosynthesis and the tricarboxylic acid (TCA) cycle. Isopropylmalate dehydrogenase in leucine biosynthesis consists of two chains, leuD and leuC. Each of these has one of the two domains contained in the aconitases A and B in the TCA cycle. 3-Isopropylmalate dehydrogenase, the next enzyme in leucine biosythesis has the same type of domain as isocitrate dehydrogenase (icdA) in the TCA cycle. The genes in leucine biosynthesis are next to each other on the *E. coli* chromosome, while those in the

TCA cycle are scattered around, even the isozymes acnA and acnB. This example of pathway duplication is mentioned by Huynen & Snel,[21] who also point out the pathway duplication of the proteins in the *prp* operon and proteins in the glyoxylate shunt. The proteins in the *prp* operon are not in our dataset, as there is no experimental evidence for their activity in the potential methyl citrate cycle in *E. coli*.

There are potential duplications of consecutive enzymes within nucleotide metabolism and ubiquinone biosynthesis where neither pair of enzymes is close to each other on the *E. coli* chromosome. There is also an example of a multifunctional multi-domain enzyme adhE in fermentation, that contains the same domains as the consecutive enzymes fucO and aldA in glycolate metabolism and rhamnose and fucose catabolism. In this case, there would have been a duplication followed by a gene fusion or fission.

The scarcity of examples like those described above indicates that, in general, recruitment of domains, whether within or across pathways, is not ordered with respect to a chain of consecutive reactions. In general, nature has recruited individual proteins or domains to pathways, not sets of consecutive enzymes. In view of the fact that there are so few pairs of homologous consecutive enzymes as a fraction of the possible pairs, it is not surprising that we could not find any example of a homologous triplet. We are led to conclude that there is little order in the recruitment of domains in the construction of metabolic pathways. It seems that, in general, domains were simply recruited individually for whatever function was needed, without preference for domains close by in the new pathway or in existing pathways.

Proteins in the SMMP pathways that may have been horizontally transferred recently

Lawrence & Ochman[22] identified genes in *E. coli* that are potential candidates for horizontal transfer within the last 10^8 years by testing whether the GC content in the first and third codon positions were atypical when compared to the entire genome. They found 755 candidates in *E. coli* and 15 of these are in our set of enzymes in metabolic pathways, indicating that this set of enzymes has probably not been affected by recent horizontal transfer on a large scale. All but one of these 16 proteins has a structural or sequence assignment, and all of the proteins with an assignment are members of families that contain domains from other SMMP proteins that are not candidates for recent horizontal transfer.

There is one enzyme, glyoxalase II (gloB) that has the only metallo-hydrolase domain in small molecule metabolism. There are two cases where a sequence of multiple genes may have been horizontally transferred recently: galactoside *O*-acetyltransferase (lacA) and beta-galactosidase (lacZ), as well as five genes (glf, rfbC, rfbA, rfbD and rfbB) in O-antigen biosynthesis. (O-antigen biosynthesis is the synthesis of a repeat unit composed of four sugars that are attached to lipids in the outer membrane.)

Homologues of SMMP proteins in other functional categories

The 695 domains in the SMMP proteins that are homologous to proteins of known structure belong to 202 families. The HMMs for these families were matched to all *E. coli* proteins, and 134 of these families were found to have additional members outside small molecule metabolism. In all, the 134 families have 1517 members in *E. coli*. Of these, 577 members are in SMMP proteins and 1039 are in proteins that are outside the SMMP. This means that most of the constituents of SMMP proteins, nearly 85 %, belong to families whose members have been recruited within and between the SMMP, and from and to many other physiological roles or functional classes that have been described for *E. coli* .[23]

The 68 families whose members are found only in the SMMP are all small and contain a total of 118 members.

Conclusions

The mechanisms that generate protein repertoires, the early *ab initio* invention of a set of different domains, and its subsequent elaboration and specialisation through gene duplication, divergence and recombination, have been the subject of analysis and discussion for over 50 years. The idea of homologues forming pathways followed Horowitz' argument for retrograde evolution of pathways, and an example was discussed by Wilmanns *et al*.[18] in the tryptophan biosynthesis pathway. Jensen[5] suggested recruitment across pathways as a mechanism of pathway evolution. Its importance for the formation of pathways that have evolved recently has been described for the mandelate pathway by Petsko *et al*.[9] and for a pathway that degrades a xenobiotic pesticide by Copley.[24] On the basis of the distribution of 38 $(\alpha\beta)_8$ homologous barrel structures in central metabolism, Copley & Bork[25] have also argued that recruitment plays a significant role in the formation of metabolic pathways.

What is new in our work is the quantitative, detailed description of the extent and roles of these different mechanisms in the formation of 510 of the 581 proteins that form the 106 small molecule metabolic pathways of *E. coli*. We have presented here an overview of these results, and the accompanying web site gives the individual results for each of the 510 proteins.

In *E. coli*, close to one-half of the proteins that form these pathways are built from a single domain, whilst the other half have between two and six domains. The evolutionary relationships of the 722 domains that form all or part of the 510 SMMP proteins were determined. The domains belong to one of 213 different families that have

between one and 51 members, and on average, 3.4 members. Domains in almost 70 % of the families undergo recombination with other domains from usually a small number of families and in a fixed N-to-C orientation.

Domains within the same family, and even with the same pairwise domain combination, are widely distributed across different pathways. The presence of homologues within pathways is less common: of the 106 pathways, only 11 have a significant number of homologous domains. Even in these cases, it is common for homologous enzymes to conserve catalytic or cofactor-binding properties and very rare for them to be close in a pathway, conserving substrate recognition and changing their catalytic mechanism. Similarly, recruitment of family members across pathways involves conservation of catalytic mechanism and cofactor-binding domains much more than conservation of substrate recognition with change in chemistry. This suggests that it is more difficult to evolve a new catalytic mechanism than a new substrate-binding site. There was very little order in this process of recruitment of enzymes, as there are very few examples of serial recruitment of consecutive enzymes from one pathway to another.

There are 134 families whose members form nearly 85 % of the SMMP proteins and which also have members outside small molecule metabolism: 68 families, whose members form just over 15 % of the SMMP proteins, occur only in these pathways.

A small proportion of SMMP proteins are not included in this work. When data for these become available to allow their inclusion, the numbers reported here will be somewhat modified. However, it is most unlikely that the general results will be significantly different. The general conclusions we draw about enzyme and metabolic pathway evolution are likely to hold true for all species and metabolic pathways.

The universal presence in cells of the proteins of central metabolism indicates that it was present in the ''last common ancestor'' and was distributed to all descendants. Though during evolution central metabolism has been modified by losses, substitutions and innovations (see Dandekar *et al.*,[26] Huynen *et al.*[27] and Makarova *et al.*[28] for recent work in this area) the enzymes of metabolism are, in general, well conserved across all kingdoms. *E. coli* is a representative of the descendants of the metabolically competent last common ancestor.

Overall, the results reported here show that even in the last common ancestor the functional domains that form the repertoire of proteins in an organism must have had an extensive ''mosaic'' character. Most proteins are formed by families whose members have a function that can be used repeatedly, or can be modified easily for related uses. Only a minority of proteins are formed by small families whose members have a functional role that is required in only one or a few instances, and which cannot be modified easily to perform other roles. This, together with quantitative descriptions that we give for here the gene duplications; recombinations, recruitment across pathways and the use of SMMP domains in other physiological roles, suggests that much of the basic protein repertoire was developed in organisms very much simpler than any known at present.

Methods

A procedure for determination of the domain structure and the evolutionary relationships of *E. coli* SMMP proteins

As described above, we used the information on domain structure and evolutionary relationships contained in the SCOP database[15] to identify even distant relationships between *E. coli* SMMP proteins. We call the sequences corresponding to whole small proteins, or to the SCOP domains in large proteins, PDBD sequences.

Thus, to use structural information to determine the evolutionary relationships of the metabolic proteins of *E. coli* we use the following procedure.

(i) Find which whole *E. coli* sequences, or regions of sequence, match HMMs of PDBD sequences. (The domain structure of the *E. coli* sequences will be given by the number of domains that match the sequence in non-overlapping regions, and the size of any unmatched region.) (ii) Cluster into families the matched *E. coli* sequences on the basis of those known for the homologous PDBD sequences. (For example, all *E. coli* sequences or sequence regions that match PDBD sequences in the Rossmann NAD-binding domain family are members of that family, even though the *E. coli* sequences do not show significant pairwise matches to each other.)

For the SMMP proteins that were not matched by the HMM procedure, and for the unmatched regions in partially matched sequences, we tried to find family relationships using the multiple sequence comparison procedure PSI-BLAST in the manner described by Park & Teichmann.[29]

The details of the iterative HMM procedure SAM-T99, is described by Gough *et al.*,[30] who made HMMs for all PDBD sequences in SCOP version 1.53 that have less than 95 % identity. The SAM-T99 procedure was developed by Karplus *et al.*[31] and the parameters used to construct this library of models were calibrated by Gough *et al.*[30] These models were matched against the 569 known SMMP sequences. The models matched 487 SMMP sequences in 695 non-overlapping regions. Three-quarters of these SMMP sequences were completely, or very largely, covered by the PDBD matches and one-quarter were partially matched, in that they also have a region of 75 or more residues that is unmatched.

As mentioned above, we also tried to find family relationships in the 130 SMMP proteins that were not matched by the HMM procedure, and in the unmatched regions in partially matched sequences, using the multiple sequence comparison procedure PSI-BLAST in the manner described by Park & Teichmann.[29] Using this procedure, matches were made between 27 domains, from (i) the unmatched regions in four proteins partially matched by PDBD sequences and (ii) 23 proteins with no PDBD matches, that cluster into 11 families.

Acknowledgements

S.A.T. has a Beit Memorial Fellowship. S.C.G.R. is funded by GlaxoSmithKline and M.R. acknowledges support from the NIH and the NASA Astrobiology Institute. We acknowledge computational support from the BBSRC.

References

1. Bridges, C. B. (1935). Salivary chromosome maps. *J. Hered.* **26**, 60-64.
2. Lewis, E. B. (1951). Pseudoallelism and gene evolution. *Cold Spring Harbor Symp. Quant. Biol.* **16**, 159-174.
3. Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proc. Natl Acad. Sci. USA,* **31**, 152-157.
4. Horowitz, N. H. (1965). The evolution of biochemical syntheses - retrospect and prospect. In *Evolving Genes and Proteins* (Bryson, V. & Vogel, H. J., eds), pp. 15-23, Academic Press, New York.
5. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409-425.
6. Hartley, B. S. (1970). Homologies in serine proteinases. *Phil. Trans. Roy. Soc. ser. B,* **257**, 77-87.
7. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding proteins. *Nature,* **250**, 194-199.
8. Neidhart, D. J. & Petsko, G. (1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature,* **347**, 692-694.
9. 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.
10. Murzin, A. G. (1993). Can homologous proteins evolve different enzymatic activities? *Trends. Biochem. Sci.* **18**, 403-405.
11. Babbitt, P. C. & Gerlt, J. A. (1997). Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30591-30594.
12. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. & Krummenacker, M. (1999). EcoCyc: electronic encyclopedia of *E. coli* genes and metabolism. *Nucl. Acids Res.* **27**, 55.
13. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparason methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.
14. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
15. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
16. Nomenclature Committee of the International Union of Biochemistry Molecular Biology (NC-IUBMB) (1992). *Enzyme Nomenclature*, Academic Press, New York.
17. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113-11143.
18. Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K. & Jansonius, J. N. (1991). Structural conservation in parallel βα-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry,* **30**, 9161-9169.
19. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science,* **289**, 1546-1550.
20. Todd, A. E., Orengo, C. A. & Thornton, J. M. (1999). Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* **3**, 548-556.
21. Huynen, M. A. & Snel, B. (2000). Gene and context: integrative approaches to genome analysis. *Advan. Protein Chem.* **54**, 345-379.
22. Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA,* **95**, 9413-9417.
23. Riley, M. (1998). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl. Acids Res.* **26**, 54.
24. Copley, S. D. (2000). Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem. Sci.* **25**, 261-265.
25. Copley, R. R. & Bork, P. (2000). Homology of $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627-640.
26. 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.
27. Huynen, M. A., Dandekar, T. & Bork, P. (1999). Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends. Microbiol.* **7**, 281-291.
28. Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (2000). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**, 609-628.
29. 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.
30. Gough, J., Chothia, C., Karplus, K., Barrett, C. & Hughey, R. (2000). Optimal hidden Markov models for all sequences of known structure. In *Currents in Computational Molecular Biology* (Miyano, S., Shamir, R. & Takagi, T., eds), pp. 124-125, Universal Academic Press, Tokyo, Japan.
31. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846-856.
32. Nahum, L. A. & Filey, M. (2001). Divergence of function in sequence-related groups of *E. coli* proteins. *Genome Res.* In the press.

Edited by G. von. Heijne