3 Eisenberg, D. *et al.* (2000) Protein function in the post-genomics era. *Nature* 405, 823–826
4 Burley, S.K. *et al.* (1999) Structural genomics: beyond the human genome project. *Nat. Genet.* 23, 151–157
5 Hendrickson, W.A. (2000) Synchrotron crystallography. *Trends Biochem. Sci.* 25, 637–643
6 Schweitzer, B. *et al.* (2000) Inaugural article: immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10113–10119
7 Legrain, P. *et al.* (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.* 17, 346–352
8 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934

Yuk Fai Leung
Chi Pui Pang*
Dept of Ophthalmology & Visual Sciences, The Chinese University of Hong Kong, Hong Kong Eye Hospital, 3/F, 147K Argyle Street, Kowloon, Hong Kong.
*e-mail: cppang@cuhk.edu.hk

Techniques & Applications

Small-molecule metabolism: an enzyme mosaic

Sarah A. Teichmann, Stuart C.G. Rison, Janet M. Thornton, Monica Riley, Julian Gough and Cyrus Chothia

Escherichia coli **has been a popular organism for studying metabolic pathways. In an attempt to find out more about how these pathways are constructed, the enzymes were analysed by defining their protein domains. Structural assignments and sequence comparisons were used to show that 213 domain families constitute ~90% of the enzymes in the small-molecule metabolic pathways. Catalytic or cofactor-binding properties between family members are often conserved, while recognition of the main substrate with change in catalytic mechanism is only observed in a few cases of consecutive enzymes in a pathway. Recruitment of domains across pathways is very common, but there is little regularity in the pattern of domains in metabolic pathways. This is analogous to a mosaic in which a stone of a certain colour is selected to fill a position in the picture.**

According to the *Concise Oxford Dictionary*, a mosaic is 'a picture…produced by an arrangement of small variously coloured pieces of glass or stone'. A mosaic is analogous in several ways to small-molecule metabolic pathways. In particular, the enzymes that form the metabolic pathways belong to a limited set of protein families, like the set of different coloured pieces available to the artist to construct the mosaic. Furthermore, the picture of the mosaic as a whole is meaningful, even though there is no discernible repeating pattern in the way the pieces are arranged; instead, each piece has been selected to fill a space with the necessary colour to make the mosaic picture. Likewise, domains in enzymes appear to have been selected from a protein family in an unsystematic way to fill a position in a pathway for the functional features of that family.

Box 1. Determining the domain structure and family membership of enzymes

Structural domains

The domain definitions and evolutionary relationships of the proteins of known structure are described in the Structural Classification of Proteins (SCOP) database[a] (http://scop.mrc-lmb.cam.ac.uk/scop/). In SCOP, domains are structural but also evolutionary units, so a domain has to be observed on its own in a structure or combined with several different domains to be classified as a domain. The phenylalanyl-tRNA synthetase large chain is shown as an example of a multi-domain polypeptide chain (Fig. I).

Domains are classified into superfamilies on the basis of sequence, as well as structural and functional features that are shared by all the domains in a superfamily.

Gough *et al.*[b] used the domains from SCOP version 1.53 as seed sequences to build a type of profile called Hidden Markov Models. (The specific method is described by Karplus *et al.*[c]) The database of Hidden Markov Models is available at http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/.

These models were then scanned against the *Escherichia coli* enzymes to identify domains in the enzymes. The family membership of the *E. coli* domains was inferred from the SCOP superfamily membership of the homologous SCOP domain.

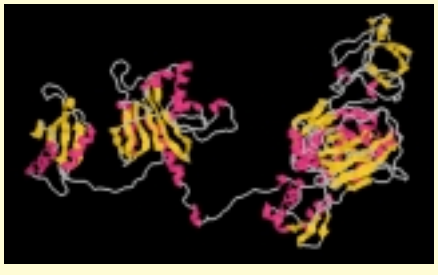

Fig. I An example of a multi-domain polypeptide chain.

Sequence domains

The regions of the *E. coli* enzymes not matched by a structural domain were compared using the multiple sequence comparison procedure PSI-BLAST[d], and then clustered into families as described by Park and Teichmann[e].

References

a Murzin, A. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540
b Gough, J. *et al.* (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313, 903–919
c Karplus, K. *et al.* (1998) Hidden Markov Models for detecting remote protein homologies. *Bioinformatics* 14, 846–856
d Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
e Park, J. and 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

  PII: S0167-7799(01)01813-3

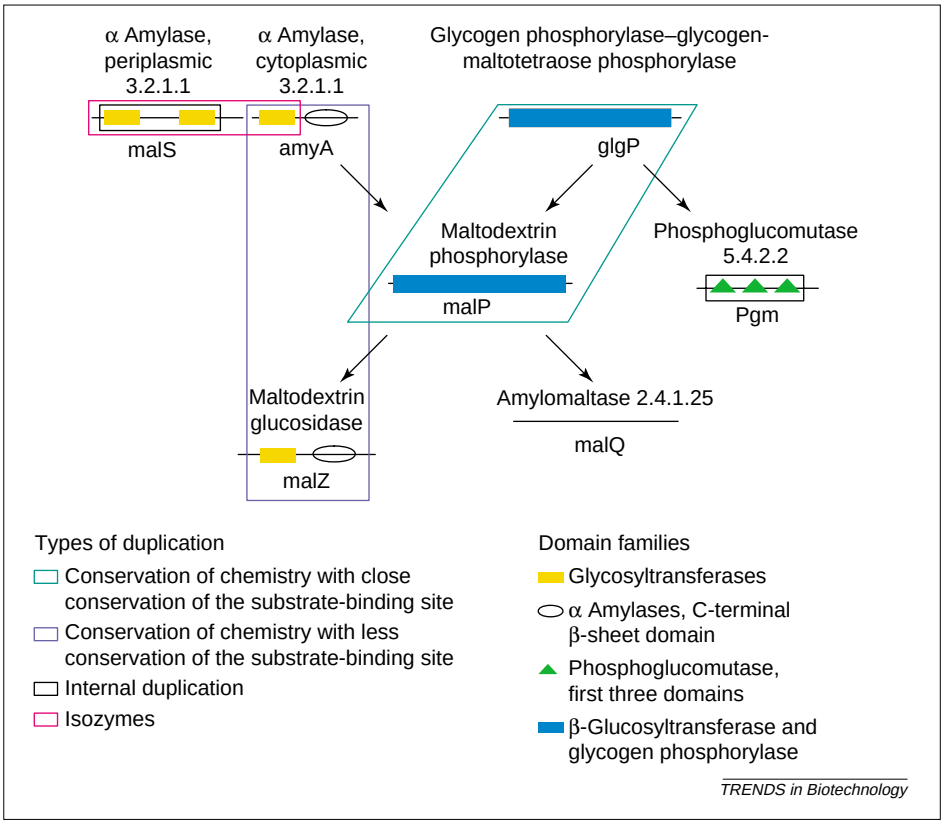

Fig. 1. Glycogen catabolism pathway.The enzymes are represented by black lines and the structural domains by coloured shapes in N-to-C-terminal order on the polypeptide chain.The arrows represent the flux of substrates and products through the pathway.There are two duplications with conservation of catalytic mechanism in this pathway. One is in consecutive enzymes (e.g. glgP and malP), therefore there is also close conservation of substrate-binding site, whereas the other duplication occurs for enzymes one step apart (e.g. amyA and malZ), with less conservation of the substrate-binding site.There are also internal duplications, in which the same type of domain occurs several times in one polypeptide sequence (malS and pgm) and isozymes (malS and amyA).

The 'colours' of the enzymes in the mosaic of *Escherichia coli* small-molecule metabolic pathways were determined by assigning the domains in each enzyme to a protein family. These protein families were derived from a combination of sequence and structural information (Box 1). Like roughly hewn mosaic pieces of one colour, the domains that belong to one family are not identical, but can be very divergent. The result of the domain assignments is a description of the structural anatomy of metabolic pathways and their enzymes, for example those involved in glycogen catabolism (Fig. 1). Such a clarification of the domain structure of enzymes provides a picture of the structural anatomy of the individual enzymes in the metabolic pathways and allows investigation into any patterns in duplicated enzyme domains within and across the metabolic pathways.

Structural anatomy of *E. coli* small-molecule metabolic enzymes

The metabolic pathways in *E. coli* are probably the most thoroughly studied of any organism. Although the details of the enzymes and metabolic pathways will differ from organism to organism, the principles of the structure and evolution of the pathways would be expected to apply across all organisms. The EcoCyc database[1] contains comprehensive information on small-molecule metabolism in *E. coli*, and the 106 pathways and the corresponding 581 enzymes described in this database were used in the present study. The results of the domain assignment procedure (shown in Box 1 and described in detail in Ref. 2) gave a total of 722 domains in 213 families in 510 (88%) of the *E. coli* small-molecule metabolism (SMM) enzymes (summarized in Box 2 and Table 1). There are, on average, 3.4 domains per family, which shows that even this basic set of pathways is the product of extensive duplication of domains within its enzymes. The distribution of family sizes of the 213 families is roughly exponential: 74 families in *E. coli* SMM have only one domain, and the largest family, the Rossmann domains, has 53 domains.

There has been not only extensive duplication of domains but also combinations of domains in these pathways, as exemplified by the fact that 722 domains are assigned to only 510 enzymes. Two-thirds of the 213 families have at least one domain that is adjacent (within 75 residues) to another assigned domain in one of the SMM proteins. Most families have only one or two types of domain partners in a fixed N-to-C-terminal orientation, but the Rossmann domain family has 12 different partner families.

Figure 2 illustrates some of the enzymes that contain Rossmann domains. Half of the SMM enzymes are single-domain proteins, similar to the dihydrobenzoate dehydrogenase (entA) in Figure 2. A quarter of all SMM enzymes contain two domains. For example, the NAD-linked malic enzyme (sfcA) shown in Fig. 2 consists of a Rossmann domain and an amino acid dehydrogenase-like domain. Of the 141 families that are adjacent to another

Box 2. Pathways, proteins, domains and families

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

Table 1. Numbers of domains in enzymes

Number of domains (n)	Numbers of sequences completely matched by n domains	Numbers of sequences partly matched by n domains
1	271	77
2	96	26
3	28	5
4	2	3
5	1	–
6	1	–
Total number of proteins	399	111

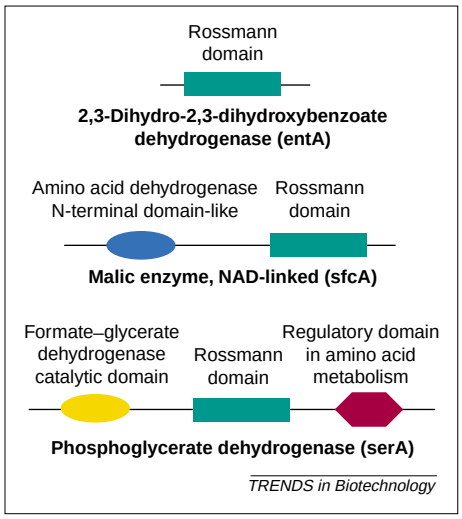

Fig. 2. Rossmann domains in enzymes. The polypeptide chains of enzymes are represented by black lines and the structural domains are represented by shapes from left to right in their N-to-C-terminal order. Examples of single-domain, two-domain and three-domain enzymes containing Rossmann domains are given, showing how domains from this family combine with other domains in different ways.

assigned domain in the SMM enzymes, 73% combine with only one type of domain. The Rossmann domain family, however, is versatile in that it can combine with more than one type of domain. Figure 2 shows two domain neighbours, in addition to those of the amino acid dehydrogenase-like family, in phosphoglycerate dehydrogenase (serA). Like the phosphoglycerate dehydrogenase (serA), a sixth of all *E. coli* SMM enzymes contain three to six domains. Half of the SMM enzymes are multi-domain enzymes, and almost three-quarters of the domain families in these enzymes have at least one domain member that is adjacent to another assigned domain in one of the SMM enzymes.

It is clear that even proteins as fundamental to the functioning of a free-living cell, and also as ancient as the central SMM enzymes, are not all simple single-domain enzymes but are the product of extensive domain combinations. Therefore, either SMM enzymes developed by fusions and recombinations from a more basic set of proteins, which were single-domain proteins, or combinations of two or more domains occurred first, and then domains later split and recombined to crystallize as individual evolutionary units, the domains that are recognized today.

Evolution of *E. coli* small-molecule metabolic pathways

Information about the domain structures of the individual enzymes can be used to investigate aspects of the evolution of metabolic pathways. Of the 213 domain families, 144 have members distributed across different pathways. The 69 families that are active in only one pathway are all small: 67 have one or two members, one has three members and one has four members. This distribution shows that the evolution of metabolic pathways involved widespread recruitment of enzymes to different pathways, which supports Jensen's model of pathway evolution[3].

Types of conservation of domain duplications

It is helpful when discussing pathway evolution to distinguish between different types of duplications of enzymes and their domains. Figure 1 shows multiple copies of the four types of domains in the glycogen catabolism pathway. The glycosyltransferase domain family (yellow) and the phosphoglucomutase domains (green) recur within the individual proteins periplasmic α amylase (malS) and phosphoglucomutase (pgm). This type of duplication is termed internal duplication and can only take place within pathways. Duplication of domains in enzymes that are isozymes can also only occur within pathways. Glycosyltransferase domains are also present in periplasmic (malS) and cytoplasmic α amylase (amyA) and in the maltodextrin glucosidase (malZ). The duplication between α amylase and maltodextrin glucosidase conserves catalytic mechanism because enzymes hydrolyse glucosidic linkages. Similarly, the two phosphorylase domains (shown in blue) conserve reaction chemistry because both glycogen phosphorylase (glgP) and maltodextrin phosphorylase (malP) are phosphorylases acting on different substrates. Recent studies have described this evolutionary mechanism in detail and show how mutations in active site residues produce new catalytic properties for enzymes[4–7]. There are two further types of duplication that do not occur in the glycogen catabolism pathway: duplication of cofactor- or minor substrate-binding domains such as Rossmann domains and duplication with conservation of the substrate-binding site but change in catalytic mechanism.

Duplications within pathways

Of the different types of duplication listed previously, internal duplication and duplication that occurs in isozymes are frequent within pathways. Duplication with conservation of a cofactor- or minor substrate-binding site is also frequent within pathways. Within the entire set of almost 600 enzymes, there are only six examples of duplications in pathways with conservation of the major substrate-binding site and a change in the catalytic mechanism (Table 2). This means that duplications in pathways are driven by similarity in catalytic mechanism much more than by similarity in the substrate-binding pocket. This disagrees with Horowitz' model of retrograde evolution[8], in which it is suggested that enzymes within a pathway are related to each other. In fact, more enzymes that are separated by one catalytic step share a domain (11%) than do consecutive

Table 2. Conservation of the main substrate-binding site with change in reaction catalysed within a pathway.

Superfamily and pathway	Enzymes
Phosphoenolpyruvate and pyruvate $(\alpha/\beta)_8$ barrels in fermentation	pykF/pykA, ppc
Ribulose-phosphate binding $(\alpha/\beta)_8$ barrels in tryptophan biosynthesis	trpA, trpC
[a]P-binding α/β barrels in histidine, purine and pyrimidine biosynthesis	hisA, hisF
Phosphoribosyltransferases (PRTases) in histidine, purine and pyrimidine biosynthesis	prsA,purF and prsA, pyrE
dUTPase domains in deoxypyrimidine nucleotide/nucleoside metabolism	dcd, dut
Inosine monophosphate dehydrogenase $(\alpha/\beta)_8$ barrels in nucleotide metabolism	guaB, guaC

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

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 more rare than change in substrate specificity with conservation of chemistry in metabolic pathways.

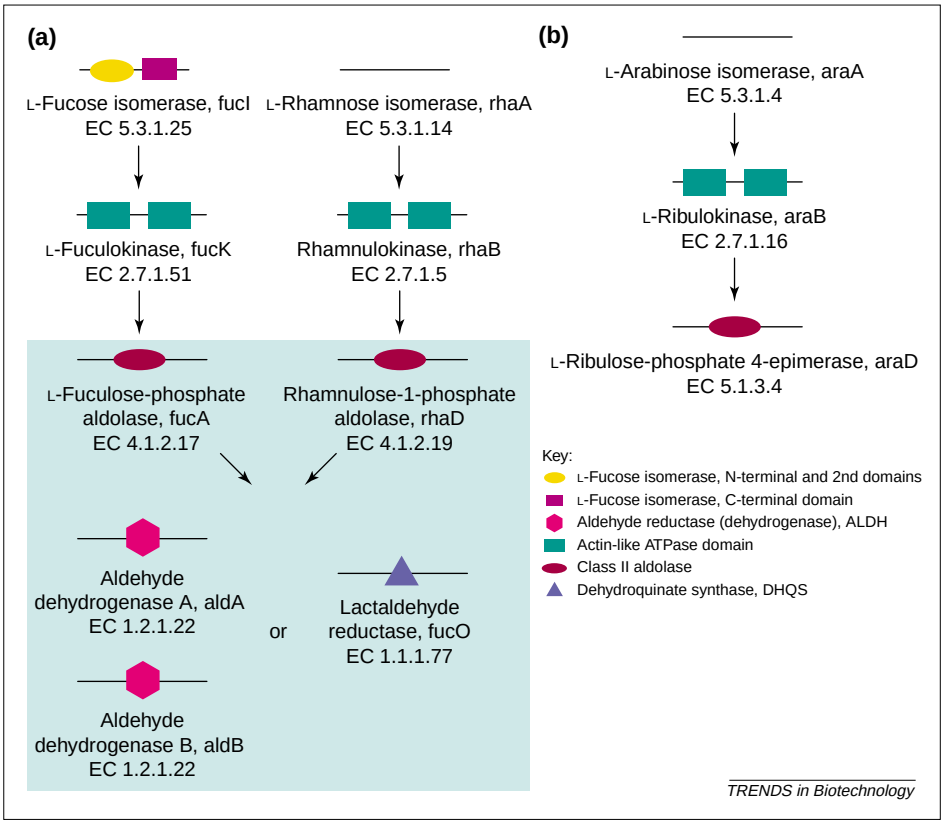

Fig. 3. Fucose, rhamnose and L-arabinose catabolism. (a) Fucose and rhamnose. A superpathway exists in EcoCyc that consists of the fucose and rhamnose catabolism subpathways. An example of serial recruitment and of 'parallel' enzymes is shown (boxed). Serial recruitment has occurred because fucK (L-fuculokinase) 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 or aldB or fucO, and are thus parallel enzymes.The enzyme classification (EC) numbers for each enzyme are given. (b) L-Arabinose. AraB is homologous to fucK and rhaB in (a), and araD is homologous to fucA and rhaD in (a).The three pairs of enzymes are an example of serial recruitment, as supported by their similar positions on the *Escherichia coli* chromosome: the genes in each pair are divided by one gene on the chromosome.

enzymes (6%), indicating that there is no bias for duplication between enzymes that are close to each other in a pathway.

Duplications within pathways occur relatively frequently in situations such as that shown in Fig. 3a, in which L-fuculose-phosphate aldolase (fucA) and rhamnulose-1-phosphate aldolase (rhaD) are homologous. In this type of case, two enzymes are followed by the same enzyme(s) in a pathway and hence have the same or similar products. Alternatively, two enzymes can also be 'parallel' when both have the same precursor enzyme in a pathway and thus have the same or similar substrates. 13% (same or similar substrates) and 17% (same or similar products) of these parallel enzymes in pathways have homologous domains.

Of the eight cases in which two enzymes are followed by the same enzyme, as in fucose and rhamnose catabolism, there are two cases, such as L-fuculose-phosphate aldolase (fucA) and rhamnulose-1-phosphate aldolase (rhaD), in which the two enzymes catalyse similar reactions and have the same product. In all the other cases, the products are merely similar, so that the enzyme that follows in the pathway possesses multiple substrate specificity. In five of the seven cases where two enzymes act on the same substrate, the two enzymes carry out similar reactions, often using a different second substrate in a reaction, such as a transferase or synthase reaction.

Duplications across pathways

As mentioned, all the larger domain families in the metabolic pathways have members in more than one pathway, thus duplications across pathways are extremely common. However, it appears that little of this recruitment takes place in an ordered fashion. Examples of serial recruitment, where two enzymes in one pathway are recruited to another pathway in the same order, such as L-fuculokinase (fucK) and L-fuculose-phosphate aldolase (fucA), rhamnulokinase (rhaB) and rhamnulose-1-phosphate aldolase (rhaD), and L-ribulokinase (araB) and L-ribulose phosphate 4-epimerase (araD) in Fig. 3, are very rare. If duplication of large portions of the bacterial chromosome takes place, and all the genes in a duplicated portion were used to form a new pathway, serial recruitment would be expected. In fact, only 89 out of 26 341 (0.3%) possible pairs of enzymes are homologous in both the first and second enzymes. Only seven of these 89 pairs of doublets of enzymes have the genes for both doublets close to each other on the chromosome, which suggests that the two initial enzymes might have been duplicated as one portion. The three kinase– and aldolase–epimerase pairs of enzymes involved in sugar catabolism are a good example of this rare situation: all three pairs are one gene apart on the *E. coli* chromosome.

Conclusions and discussion

This description of how a relatively small repertoire of 213 domain families constitutes 90% of the enzymes in the *E. coli* small-molecule metabolic pathways is, to some extent, paradoxical. Although the SMM enzymes have arisen by extensive duplication, with an average of 3.4 domain members per SMM family, the distribution of families within and across pathways is complex: there is little repetition of domains in consecutive steps of pathways and little serial homology across pathways. Together with the analysis of the chromosomal locations of genes, it is evident that metabolic pathways have, in general, not arisen by duplication of large portions of the *E. coli* chromosome, either to extend a pathway or to make a new pathway. There are a few well known exceptions to this, such as the enzymes involved in the fucose, rhamnose and arabinose catabolic pathways. Similarly, duplication of enzymes that conserve a substrate-binding site is rare, otherwise the fraction of consecutive homologous enzymes would be larger. The main pressure for selection for enzymes in pathways appears to be either their catalytic mechanism or cofactor-binding properties. This pattern of evolution has resulted in a mosaic of enzyme domains optimized for smooth-functioning

small-molecule metabolism in *E. coli*, with little order in the pattern of domains with respect to position within or between pathways.

Selection based entirely on function, and specifically reaction chemistry, was termed 'patchwork evolution' by Lazcano and Miller and also by Copley in a discussion of the pathway for the degradation of pentachlorophenol by the soil micro-organism *Sphingomonas chlorophenolica*[10]. Pentachlorophenol was introduced into the environment in 1936, and is not produced naturally, so it is probable that the pathway evolved in the past few decades. The pathway involves three enzymes, which were recruited in a 'patchwork' manner from the enzymes that break down naturally occurring chlorinated phenols.

Recently, recruitment of enzymes across metabolic pathways was observed in a study of the distribution of $(\alpha/\beta)_8$ barrels by Copley and Bork[11], and in a review on structural genomics of metabolic pathways by Erlandsen and colleagues[12]. The comprehensive structural assignments to 90% of the enzymes in all *E. coli* small-molecule metabolic pathways described in the present article confirm that pathways are constructed by recruitment on the basis of catalytic mechanism, with few instances of duplication of enzymes within a pathway or serial recruitment across pathways.

Acknowledgements

SAT has a Beit Memorial Fellowship. SCGR acknowledges support from GlaxoSmithkline and MR acknowledges support from the NIH and the NASA Astrobiology Institute. The authors acknowledge computational support from the BBSRC.

References

1 Karp, P. *et al.* (1999) EcoCyc: electronic encyclopedia of *E. coli* genes and metabolism. *Nucleic Acids Res.* 27, 55–58
2 Teichmann, S.A. *et al.* (2001) The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in *Escherichia coli*. *J. Mol. Biol.* 311, 693–708
3 Jensen, R.A. (1976) Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* 30, 409–425
4 Murzin, A.G. (1993) Can homologous proteins evolve different enzymatic activities? *Trends Biochem. Sci.* 18, 403–405
5 Neidhart, D.J. and Petsko, G. (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature* 347, 692–694
6 Babbitt, P.C. and 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
7 Petsko, G.A. *et al.* (1993) On the origin of enzymatic species. *Trends. Biochem. Sci.* 18, 372–376
8 Horowitz, N.H. (1945) On the evolution of biochemical syntheses. *Proc. Natl. Acad. Sci. U. S. A.* 31, 152–157
9 Lazcano, A. and Miller, S.C. (1999) One the origin of metabolic pathways. *J. Mol. Evol.* 49, 424–431
10 Copley, S.D. (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the pathwork approach. *Trends Biochem. Sci.* 25, 261–265
11 Copley, R.R. and Bork, P. (2000) Homology of $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* 303, 627–641
12 Erlandsen, H. *et al.* (2000) Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites. *Curr. Opin. Struct. Biol.* 10, 719–730

Sarah A. Teichmann*
Stuart C. G. Rison
Janet M. Thornton
Dept of Biochemistry and Molecular Biology, University College London, Darwin Building, Gower Street, London, UK WC1E 6BT.
*e-mail: sat@mrc-lmb.cam.ac.uk

Monica Riley
Josephine Bay Paul Centre for Comparative Molecular Biology and Evolution, 7 MBL St, Woods Hole, MA 02543-1015, USA.

Julian Gough
Cyrus Chothia
MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 1TQ.

Center for Molecular Biodiversity and Evolution set up

The biotechnology company BRAIN (Zwingenberg, Germany) and scientists from the Institute of Genetics and Microbiology (Technical University, Darmstadt, Germany) have jointly set up the Center for Molecular Biodiversity and Evolution (ZEB) at theTechnical University of Darmstadt.The Center was set up with the aim of exploring the >99% of microorganisms in a typical soil sample that cannot be cultivated and to search for new enzymes and bioactive molecules. The Center's main goal is to isolate the collective genomes of a microbial community, the 'metagenome', by directly isolating DNA from soil and incorporating it into BioArchives (recombinant DNA libraries containing environmental DNA).The ZEB will be headed by Christa Schelper and represents a promising cooperation between academia and industry.

First results of collaboration between Graffinity and Aventis announced

Graffinity Pharmaceuticals (Heidelberg, Germany) has recently announced the first results in its chemical microarray collaboration with Aventis Pharma (Frankfurt, Germany). Graffinity uses chemical genomics to convert lead targets into small-molecule pharmaceuticals.The agreement between Graffinity and Aventis was first announced in May 2001 – Graffinity was to synthesise exclusive arrays for Aventis to discover novel drug leads.