Of Proteins and Pathways

Investigating Protein Functional Classifications and the Small Molecule Metabolism of Escherichia coli.

Stuart Christopher Gorthorn Rison

Biomolecular Structure and Modelling Unit
Department of Biochemistry and Molecular Biology
University College London

A thesis submitted to the University of London in the Faculty of Science for the degree of Doctor of Philosophy

September 2002
Abstract

In this thesis, proteins are considered in two ways: (i) as entities requiring functional classification, and (ii) as participants in the small molecule metabolism (SMM) of *Escherichia coli*.

The first consideration prompts an investigation of functional classification schemes applicable to gene products. The concepts of depth, breadth and resolution are used as descriptors of the schemes’ scope and architecture, and selected classifications compared on that basis. A “Combination Scheme” (CS) is generated, and a range of representative classifications are mapped against the CS. For comparison, FuncWheels (graphical representations of hierarchical classification schemes) are generated — these illustrate differences in functional space coverage. The survey highlights many issues related to the design and implementation of gene product functional classifications, which are discussed in the light of emerging second-generation schemes such as the “Gene Ontology”.

The thesis then focuses on the structural anatomy of *E. coli* SMM. Domain reuse within and between enzymes, as well as within and between pathways, is considered. Special cases such as the “serial recruitment” of blocks of enzymes, and homology in “parallel enzymes”, are studied. A network view of pathways is then taken, and correlations between four contexts are analysed: the metabolic context (i.e. the spatial organisation of enzymes in the SMM network); the genomic context (i.e. the location of enzyme encoding genes on the *E. coli* chromosome); the evolutionary context (i.e. homologies between SMM enzymes); and the functional context (i.e. the catalytic activity of enzymes). In addition, incidences of “inline reuse” of enzymes (i.e. the use of the same enzyme at different steps of a metabolic pathway) and of isozymes (homologous proteins participating in the same metabolic step) are investigated.

Taken together, these analyses suggest a chemistry-driven patchwork model of pathway evolution, but other localised mechanisms, as well as regulatory constraints, are likely to be involved.
Acknowledgements

First, I would like to thank my supervisor Janet Thornton. She tutted when I was slacking, she cajoled when I was down, and was always encouraging. Her guidance made this thesis possible, and her vision made sure that it was backed by publications. All of this she did with her usual charm, good humour and kindness — even when time was short, and her workload unbelievable. I am greatly indebted to her.

I would like to thank my scientific collaborators, Sarah Teichmann, Monica Riley, Julian Gough and Cyrus Chothia. Sarah deserves special praise; her arrival at UCL truly kick-started my thesis, and I am grateful for her impetus and drive. Many thanks also to Evangelos Simeonidis, David Bogle and Lazaros Papageorgiou from UCL’s Chemical Engineering department who developed the LP pathway distance calculation algorithm used in this thesis. I only wish we had started collaborating much earlier in the course of my thesis, but I look forward to collaborating with you further in the future.

Many people were generous with their time and data, but I would like to mention in particular the creators and curators of the EcoCyc and GenProtEC database. My thanks to Monica Riley, Margrethe Serres and Alida Pellegrini-Toole of GenProtEC for always promptly and efficiently answering my many E. coli queries, and to Peter Karp for the wonderful resource that is EcoCyc, for his assistance, and for allowing me to publish some EcoCyc pathways online. Furthermore, Julio Collado-Vides and Gabriel Moreno-Hagelsieb, who work on the RegulonDB database, provided E. coli operon data, for which I am most grateful. Many thanks also to the “CATH team” for all their help, and to Christine Orengo who leads this great bunch of people.

A number of people have acted in a ancillary supervisory capacity during the course of this Ph.D. I am grateful to Charlie Hodgman of GlaxoSmithKline for accepting at short notice to be my industrial supervisor; to John Ladbury my UCL “mentor”; and to Sean Holden who kindly supervised me during my year spent, under the auspices of a UCL Graduate School cross-disciplinary scholarship, in UCL’s Computer Science department. All of them offered advice and support in abundance. From my time at the Ludwig Institute for Cancer Research, I would like to thank Mike Waterfield and Marketa Zvelebil for their support, and Chris Odell and Rainer Cramer for much merriment.

Two people were kind (or foolish) enough to proofread my thesis: Gail Bartlett (who not only had to make sense of my English, but make sense of my science too) and Erif Rison (who, without the benefit of a formal science education, ploughed through over 200 pages of manuscript — and now leans towards the patchwork theory of pathway evolution). I am extremely grateful for their hard work.

Thanks also to those who had the misfortune to share a desk with me, Neil Stoker, Gail Bartlett, and Ian Sillitoe and Simon Bergqvist with whom I shared the “writing-up office”
until we were evicted to make space for equipment to be stored. Ian, Simon and I shared many pots of over-strong coffee, and many laughs and frustrations. Thanks to both of you for putting up with me.

Ian also deserves much credit for his computer graphics wizardry, and was always generous with his time and knowledge; the JMB cover was thanks to him. I also thank Francisco Barona-Gómez (who generously allowed me to discuss some of his unpublished work in my thesis), Gail Bartlett, James Bray, Daniel Buchan, Alistair Grant, David Lee, Frances Pearl, Adrian Shepherd, Neil Stoker, Annabel Todd and William Valdar for giving me the benefit of their scientific insight.

No bioinformatics Ph.D. could be completed without the help of sysadmins. John Bouquiere, Donovan Binns, Duncan McKenzie, Jesse Oldershaw have been brilliant. Duncan deserves a gold star for coping with my constant nagging, and for helping me way beyond the call of duty.

I don’t how one keeps a sound mind whilst doing a Ph.D., but James Bray, Brian Ferguson and Graham Cahill made sure I kept a somewhat sound body by dragging me to the gym and to circuit training. Thanks for all the aches and pains.

Many thanks also to my friends and colleagues currently, and previously, at UCL, including Jonathan Barker, Andreas Brakoulis, Stephen Campbell, Tina Clarke, Jennifer Dawe, Snezana Djordjevic, Patricia Furtado, Acely Garza, Andrew Harrison, Richard Jackson, Thomas Kabir, Roman Laskowski, Nick Luscombe, Andrew Martin, Christine Mason, Kevin Murray, Nozomi Nagano, Sylvia Nagl, Irene Nooren, Irilenia Nobeli, Shusuke Ono, Craig Porter, Mike Plevin, Oliver Redfern, Gabby Reeves, Hugh Shanahan, Robert Steward, Gillian Urquhart, Cristiana Velloso, David Vines and Gordon Whamond (phew!).

Outside of the lab, Ananyo Bhattacharya, John and Jean Doxey (and Jasper and Pepsi), Jeremy and Ema Gaywood, Morwenna Foxworthy, Liam Hudson, Simon Keane, Claire La Hovary, Barry Mallon, Rebecca Palser, Christopher Pease, Emma Reining, Dennis Rodgers, Dennis Saw and Sarah Tomlin have all been very supportive.

So many people helped in many ways but I would like to personally thank the following. Gail, thanks for your help and good humour, and for picking me up when I am down, I miss not sharing a desk with you. Simon, dudester, you and your Swiss-ball are always welcome back at my flat. James, what can I say, thanks for always being there. Dan, thanks for the beers and cheers, the late night chats and easing my Sunday blues. Gabrielle, thank you for your teasing, your smiling, and always being one of the lads. Ian, you are the Daddy, thanks for being such a great “write-up” buddy... just one more crossword? Erif, thanks for being a wonderful sister, friend and housemate. Dennis, merci pour tout, et plus encore... hasta la victoria siempre!

Finally, a very special thanks to my girl Shelley Doxey (thanks for being there for me) and to my family — you’ve all been great!
Funding

This work was generously supported by a GlaxoSmithKline studentship and a UCL Graduate School one-year scholarship for cross-disciplinary training.
Publications related to this thesis

Some of the “Introduction” (Chapter 1) and part of the “Discussion” (Chapter 6) are derived from a review published in Current Opinions in Structural Biology (Rison & Thornton, 2002).

The chapter analysing functional classification schemes (Chapter 2) is based on an article published in the maiden issue of Functional and Integrative Genomics (Rison et al., 2000).

The “Methods” chapter (Chapter 3) contains a section concerning the Gene3D database which is described in Genome Research (Buchan et al., 2002).

The chapter describing the structural anatomy of *E. coli* metabolic pathways (Chapter 4) is based on an article published in the Journal of Molecular Biology (Teichmann et al., 2001a); a related summary article was also published in Trends in Biotechnology (Teichmann et al., 2001b).

Chapter 5, concerning the relationships between pathway distance, gene intervals and homology in the small molecule metabolism of *E. coli* is based on an article published in the Journal of Molecular Biology (Rison et al., 2002). In addition, the chapter contains details of work completed in collaboration with Evangelos Simeonidis, I. David Bogle and Lazaros Papageorgiou of the Department of Chemical Engineering, UCL. The work, regarding the use of linear programming to calculate pathway distance, has not been published although a manuscript is in preparation and results have been presented at two conferences (Intelligent Systems for Molecular Biology (2001) Copenhagen, Denmark and Metabolic Engineering IV: Applied System Biology (2002) Castelvecchio Pascoli (Tuscany), Italy). The raw data analysed in Section 5.3.4 were generated by Evangelos Simeonidis. Evangelos Simeonidis also wrote most of the description of the Linear Programming algorithm for pathway distance calculation described in the “Methods” chapter.

In all cases, the chapters contain data and analyses not present in the published papers. Reprints for three of the published papers, Rison et al. (2000); Teichmann et al. (2001a); Rison et al. (2002), the basis for chapters 2, 4 and 5 respectively, are bound at the back of this thesis (page 247 onwards).
Contents

Abstract 2

Acknowledgements 3

Funding 5

Publications related to this thesis 6

Contents 7

List of Figures 14

List of Tables 17

Dedication 19

1 Introduction 20

1.1 Pathway evolution ......................................................... 22
  1.1.1 Small molecule metabolism ......................................... 22
  1.1.2 Retrograde evolution ................................................ 22
  1.1.3 Patchwork evolution .................................................. 24
  1.1.4 Further theories ..................................................... 25
  1.1.5 Examples in extant pathways ....................................... 28
    1.1.5.1 Pathway evolution and directed evolution ............... 30
  1.1.6 Revisiting pathway evolution ...................................... 33
Contents

1.2 Escherichia coli ................................................................. 34
  1.2.1 Operons ................................................................. 36
    1.2.1.1 The lac operon ............................................. 36
    1.2.1.2 Operon prediction ....................................... 37
  1.3 Resources ............................................................... 39
    1.3.1 Nucleotide sequence databases ............................. 40
    1.3.2 Protein sequence databases .................................. 40
      1.3.2.1 The Protein Information Resource databases ...... 40
      1.3.2.2 SWISS-PROT and TrEMBL ............................. 41
      1.3.2.3 Other protein sequence databases .................. 42
    1.3.3 Structural databases .......................................... 43
      1.3.3.1 PDB .......................................................... 43
    1.3.4 Structure classification databases .......................... 44
      1.3.4.1 CATH ...................................................... 45
      1.3.4.2 SCOP ...................................................... 47
      1.3.4.3 CATH and SCOP ........................................ 48
    1.3.5 Metabolic databases ............................................ 49
  1.4 Overview of this thesis .............................................. 52

2 Investigating functional annotation schemes for genomes 53
  2.1 Comparison strategy .................................................. 54
  2.2 Methods ................................................................. 56
    2.2.1 Classification scheme uploading ............................ 56
    2.2.2 Design of the “Combination Scheme” and scheme mapping 57
    2.2.3 Generation of FuncWheels ..................................... 58
  2.3 Results ................................................................. 60
    2.3.1 Scheme survey .................................................. 60
    2.3.2 Single and multi-organism schemes .......................... 65
    2.3.3 Scheme depth, resolution and breadth ..................... 65
## Contents

2.3.4 The “Combination Scheme” .................................................. 67  
2.3.5 Functional scheme comparison ........................................... 70  
2.4 Discussion ................................................................................. 74  
  2.4.1 Mapping and scheme comparison limitations ...................... 74  
  2.4.2 Scheme level meaning .......................................................... 75  
  2.4.3 Function, Apples, and Oranges ........................................... 76  
  2.4.4 Multi-dimensionality and multi-functionality ...................... 77  
  2.4.5 Surveyed schemes ............................................................... 78  
  2.4.6 Future schemes ................................................................. 80  
2.5 Recent developments ............................................................... 81  

3 Analysing SMM — Methods ......................................................... 83  
  3.1 The EcoCyc database ............................................................... 84  
    3.1.1 Description ................................................................. 84  
    3.1.2 Sources of EcoCyc data .................................................. 84  
    3.1.3 The EcoCyc ontology of biological function ..................... 85  
  3.2 Generating the SMM pathway set ............................................ 87  
    3.2.1 Downloading EcoCyc data .............................................. 87  
    3.2.2 The “npl” representation of pathways ............................... 88  
      3.2.2.1 Description of the npl format ................................... 89  
    3.2.3 Merging EcoCyc pathways .............................................. 91  
    3.2.4 Ancillary data ............................................................... 93  
      3.2.4.1 Gene identifiers ...................................................... 93  
      3.2.4.2 Chromosomal location and gene intervals ................. 94  
      3.2.4.3 Function and EC Numbers ...................................... 94  
      3.2.4.4 Operons .............................................................. 94  
    3.2.5 Storing data in PostgreSQL ............................................. 95  
    3.2.6 Pathway distance .......................................................... 95  
      3.2.6.1 Calculating pathway distances — DFS ....................... 97
3.2.6.2 Calculating pathway distances — LP ................. 99
3.3 Structural assignments for genomes ................................. 102
  3.3.1 Profile-based search methods ................................. 103
    3.3.1.1 Position Specific Score Matrices and PSI-BLAST ... 104
    3.3.1.2 Hidden Markov Models and SAM-T99 ................. 104
  3.3.2 SCOP assignments ..................................................... 107
    3.3.2.1 The SUPERFAMILY library ................................. 107
    3.3.2.2 Assignment of SCOP superfamilies to genome se-
             quences ................................................................. 108
  3.3.3 CATH assignments .......................................................... 108
    3.3.3.1 CATH-PFDB and DomainFinder ........................... 109
    3.3.3.2 CATH-PFDB and CATH-IMPALA ........................... 112
    3.3.3.3 Gene3D and DRange ........................................... 113
    3.3.3.4 CATH-PFDB and Gene3D ....................................... 119
  3.3.4 Extending the homologous family repertoire with sequence
       families ................................................................. 120
    3.3.4.1 Automated determination of sequence domains ..... 121
    3.3.4.2 Assigning sequence families to structural families .. 122
  3.3.5 Extending superfamilies into hyperfamilies ................. 122
3.4 Final datasets ............................................................... 123
  3.4.1 Small Molecule Metabolism pathway datasets ............... 124
    3.4.1.1 The ANAT_PATHWAYS dataset ......................... 124
    3.4.1.2 The CONTEXT_PATHWAYS dataset ..................... 126
  3.4.2 Structural and sequence assignments ......................... 128

4 Analysing SMM — The Evolution and Structural Anatomy of
Small Molecule Metabolism .................................................. 131
  4.1 Pathways and the domain structure and homology of E. coli SMM
       proteins ................................................................. 132
  4.2 Domain coverage of E. coli SMM proteins ...................... 133
Contents

4.3 Protein families that form the \textit{E. coli} SMM ............................................. 135
  4.3.1 Families of protein domains ............................................................ 135
  4.3.2 Domain combinations in \textit{E. coli} SMM proteins ..................... 135
  4.3.3 Whole-proteins families ............................................................... 137

4.4 The distribution of family members within and across pathways . .. 139
  4.4.1 The distribution of individual domains ........................................ 139
  4.4.2 The distribution of the members of whole-protein families . . . . 141

4.5 Types of conservation within families ................................................. 141

4.6 Homologues within pathways ............................................................... 142
  4.6.1 Pairs of homologues that form consecutive steps in pathways . . 142
  4.6.2 Homologous pairs of enzymes with the same substrate or product 146
  4.6.3 Pathways with a high proportion of homologues ......................... 148
    4.6.3.1 Phosphatidic acid and phospholipid biosynthesis . . 150
    4.6.3.2 Colanic acid biosynthesis .................................................. 150
    4.6.3.3 Nucleotide metabolism ....................................................... 150
    4.6.3.4 3-deoxy-D-manno-octulosonate, peptidoglycan and lipid A-precursor biosynthesis 151
    4.6.3.5 Polyisoprenoid biosynthesis ............................................. 151
    4.6.3.6 Glycogen catabolism ........................................................... 151
    4.6.3.7 Gluconeogenesis .................................................................... 151
    4.6.3.8 Histidine, purine and pyrimidine biosynthesis ............ 153
  4.6.4 Overview of highly-homologous pathways .................................. 153

4.7 Homologues in different pathways ....................................................... 153
  4.7.1 Conservation of function across pathways ................................ 153
  4.7.2 The extent of serial recruitment of proteins across pathways . . 154

4.8 \textit{E. coli} SMM pathways and horizontal gene transfer .................... 158

4.9 Homologues of SMM proteins in other functional categories .......... 159

4.10 Discussion .......................................................................................... 160
5 Analysing SMM — Pathway Distance, Genome Distance, Homology and Function 162

5.1 Small Molecule Metabolism pathways .......................... 163
5.2 Structural annotation, sequence families and evolutionary relatedness 164
5.3 Context based analyses .......................................................... 164
   5.3.1 Homology and pathway distance .......................... 166
   5.3.2 Homology and gene intervals .......................... 168
   5.3.3 Gene intervals and pathway distance ...................... 169
   5.3.4 Pathway distance, chromosomal distance and function .......................... 171
      5.3.4.1 Chromosomal distance and pathway distance ......... 171
      5.3.4.2 Function and pathway distance .......................... 175
5.4 Homology within reaction frames .............................................. 176
5.5 Inline reuse .......................................................... 187
5.6 Discussion .......................................................... 193
   5.6.1 Recruitment of homologous proteins .................. 193
   5.6.2 Nearby pathway enzymes are clustered in the genome .... 194
   5.6.3 Genome distance, pathway distance and homology ............ 195
   5.6.4 Operons, inline-reuse, isozymes and regulation .................. 196

6 Discussion 198

6.1 Towards a network representation of pathways .......................... 198
   6.1.1 Dealing with ubiquitous compounds .................. 199
   6.1.2 Network properties of SMM .......................... 200
6.2 Detecting evolutionary relationships in SMM pathways .............. 202
6.3 Domain composition of SMM enzymes .................................. 205
   6.3.1 Most common domains .......................... 205
   6.3.2 Battery domains ............................................ 206
   6.3.3 Domain combinations ............................................ 206
   6.3.4 Domain fusions ............................................ 207
Contents

6.4 Domain distribution ......................................................... 207

6.5 Context based analyses .................................................... 208
   6.5.1 The genomic and functional contexts .......................... 209
   6.5.2 The genomic and metabolic contexts .......................... 210
   6.5.3 The genomic and evolutionary contexts ....................... 211
   6.5.4 The metabolic, evolutionary and functional contexts ..... 212
   6.5.5 Parallel investigations .............................................. 213
   6.5.6 The evolutionary and functional contexts ................... 214

7 Conclusions ................................................................. 217
   7.1 Using functional annotation for research ...................... 217
   7.2 The evolution of metabolic pathways ........................... 218
   7.3 Future work .......................................................... 219

List of Abbreviations ......................................................... 222

Bibliography ................................................................. 225

Paper reprints ............................................................... 247


List of Figures

1.1 Interweaving of the historical traditions of biochemistry, cell biology, and genetics .................................................. 21
1.2 The retrograde model of pathway evolution .................................................. 23
1.3 The patchwork model of pathway evolution .............................................. 26
1.4 The relationship between prebiotic chemistry and the evolution of metabolic pathways .................................................. 27
1.5 Tryptophan biosynthesis .......................................................................... 29
1.6 Histidine biosynthesis ............................................................................. 31
1.7 Model for the evolution of the \((\beta\alpha)_5\) barrel scaffold by twofold gene duplication ............................................................................................................. 32
1.8 His A and TrpF catalyse similar reactions in the histidine and tryptophan biosynthesis .................................................. 33
1.9 A peritrichously flagellated Escherichia coli ........................................... 35
1.10 Negative and positive transcriptional control of the lac operon ............. 38
1.11 The exponential growth of the GenBank database .................................. 41
2.1 Depth, resolution and breadth of seven representative classification schemes .................................................................................. 68
2.2 The “Combination Scheme” FuncWheel .................................................. 71
2.3 Coverage of the “Combination Scheme” illustrated using FuncWheels 73
3.1 Reaction, Enzymatic-Reaction, Small-Molecule and Protein frames in EcoCyc .................................................................................. 86
3.2 The npl file for the GLYCOLYSIS/E-D pathway of E. coli ..................... 90
3.3 Merging adjacent reaction frames ........................................................... 93
List of Figures

3.4 An E/R diagram of in-house pathway and ancillary data ............... 96
3.5 Pathway distance illustrated in glycolysis and the TCA ................. 98
3.6 Convex and concave sets in Euclidean space ......................... 99
3.7 Bellman’s equations and LP in shortest path computation ............. 101
3.8 The PSI-BLAST procedure .............................................. 105
3.9 A profile hidden Markov model for a sequence alignment ............. 106
3.10 Populating the CATH Protein Family Database (CATH-PFDB) ..... 110
3.11 The DomainFinder algorithm .......................................... 111
3.12 Populating the Gene3D Database ....................................... 114
3.13 DomainFinder’s cautious assignment of consensus regions ........... 115
3.14 DRange: The Collapse module ......................................... 115
3.15 DRange: The MultiParse module ....................................... 116
3.16 DRange: The CleanAssign module ...................................... 118
3.17 The validation of E. coli K-12 MG1655 structural assignments using the DRange protocol ......................................................... 119
3.18 Domain validation for CATH-PFDB and Gene3D ....................... 120
3.19 Assigning sequence families to structural families ..................... 123
4.1 Domain family size and number of pathways in the ANAT datasets 140
4.2 Duplication within and across pathways .................................. 144
4.3 Fucose and rhamnose catabolism ......................................... 149
4.4 Glycogen catabolism ........................................................ 152
4.5 Homologous pairs of consecutive enzymes ............................... 156
5.1 Homology and Pathway distance ........................................ 166
5.2 Gene intervals and Homology ............................................ 169
5.3 Gene intervals and Pathway distance .................................... 170
5.4 Pathway distance and gene intervals ..................................... 172
5.5 Gene intervals, Pathway distance and Homology ....................... 173
5.6 Pathway distance and chromosomal distance ........................... 174
List of Figures

5.7 Pathway distance and enzyme function ............................................ 176
6.1 Attributes of exponential and scale-free network structures .......... 200
6.2 Domain families in EcoCyc pathways ............................................. 204
7.1 The big picture ............................................................................. 220
List of Tables

1.1 CATH and SCOP hierarchy levels .................................................. 46
1.2 The five main classes of the CATH database .............................. 47
1.3 Comparing entries in various CATH and SCOP levels .................. 48
1.4 Major metabolic databases ............................................................ 51

2.1 Some databases with gene product classification schemes .......... 55
2.2 EcoCyc to “Combination Scheme” mapping ................................. 59
2.3 An overview of gene product classification schemes .................... 64
2.4 The “Combination Scheme” ............................................................ 69
2.5 A comparison of the schemes used by Tamames et al. (1997) and Rison et al. (2000) .......................................................... 70
2.6 Description of the different levels in the EC classification .......... 76

3.1 The npf format ............................................................................... 92
3.2 CATH hyperfamilies ...................................................................... 124
3.3 The ANAT_PATHWAYS dataset .................................................... 125
3.4 The CONTEXT_PATHWAYS datasets ........................................... 127
3.5 Homology datasets ........................................................................ 129

4.1 Domains and families in the ANAT_HOMOL dataset ................. 134
4.2 Domain combinations in enzymes of the SMM of E. coli ............. 136
4.3 Domain pairs in E. coli SMM ............................................................ 138
4.4 Distribution of the domains in SMM proteins that function in more than one pathway ...................................................... 139
4.5 Conservation of the main substrate-binding site with change in reaction catalysed within a pathway .................................................... 145
4.6 Homologous pairs of enzymes that act on the same substrate or produce the same product ............................................................... 147
4.7 Pathways with high levels of duplication ........................................... 148
4.8 Homologous pairs of consecutive enzymes and serial recruitment ... 157

5.1 An overview of data analysed in Chapter 5 ........................................ 165
5.2 p-values for the observed percentages of homologous pairs .......... 167
5.3 Domain conservation explanations .................................................... 168
5.4 Isozyme rationales ............................................................................. 179
5.5 Isozymes in E. coli SMM ................................................................. 186
5.6 Inline reuse in E. coli SMM ............................................................... 187
5.7 Instances of inline reuse ................................................................. 192

6.1 The top 5 superfamilies detected in E. coli SMM enzymes in three different surveys ................................................................. 205
6.2 Comparison of the Alves et al. (2002) and the Chapter 5 studies ... 215
A Maman. Tout vient à temps pour qui sait attendre.
   Merci d’avoir attendu.

To Dad. All things come to those who wait.
   Thank you for your patience.

   Love, as always.
Chapter 1

Introduction

Mutability, self-duplication and heterocatalysis comprise a necessary and sufficient definition of living matter (Horowitz, 1959). In basic terms, this means a living organism can do things (catalysis), change and adapt (mutability) and reproduce (duplication). Investigations tying together notions of biochemistry, cell biology and genetics are often assigned to the domain of science known as molecular biology. Key discoveries in the fields of biochemistry, cell biology and genetics, leading to the emergence of molecular biology, are illustrated in Figure 1.1. In this thesis, life is investigated at a micro-level: that of proteins and organic compounds — even at this level, the three life-defining properties of Horowitz remain applicable. First, proteins are considered as entities requiring functional classification, then the proteins and organic compounds of the Small Molecule Metabolism (SMM) of Escherichia coli are investigated. The chapter concerning functional classifications (Chapter 2) is self-contained and little background information regarding it is given in this “Introduction”; rather, this chapter concerns the chapters analysing the SMM of E. coli (Chapters 4 and 5).

In this chapter, theories concerning the evolution of metabolic pathways are described in detail. Furthermore, key resources used in the thesis (such as sequence and structure databases) are described. Programs and algorithms used in this thesis (e.g. sequence comparison software) are not described in this chapter, but in the “Methods” chapter (Chapter 3) and in situ in the analysis chapters.
Figure 1.1: Interweaving of the historical traditions of biochemistry, cell biology, and genetics (from Becker et al., 2000).
1.1 Pathway evolution

1.1.1 Small molecule metabolism

A useful definition of metabolism is given by the "On-line Medical Dictionary" (http://cancerweb.ncl.ac.uk/omd/) as: "The sum of all the physical and chemical processes by which living organised substance is produced and maintained (anabolism) and also the transformation by which energy is made available for the uses of the organism (catabolism)". Metabolism is therefore divided into two sets of pathways: (i) catabolic pathways, and (ii) anabolic pathways. The former are involved in the degradation of metabolites, and the latter involved in their biosynthesis. The catabolic breakdown of complex metabolites is exergonic and the free energy made available during these reactions is harnessed in "high-energy compounds" in reactions such as the synthesis of ATP from ADP and phosphate or the reduction of coenzyme NADP⁺ to NADPH. In turn, molecules such as ATP and NADPH are major sources of free energy for anabolic pathways (Voet & Voet, 1995). Nearly all of these reactions require proteinaceous biological catalysts, in other words, enzymes.

No strict definition of SMM exists, but the term is generally understood to qualify the metabolism of all non-macromolecules. For example, nucleotides and aminoacids are synthesised in SMM, but not DNA and proteins. In *E. coli*, approximately 600 proteins participate in SMM in a network that involves nearly 800 chemical substrates (Ouzounis & Karp, 2000). Often, for convenience and to reflect biological priorities, these enzymes and substrates are subset into distinct pathways (see Chapter 3), but in vivo, SMM is a single complex network (Jeong et al., 2000).

A number of theories have been advanced to explain the evolution of an enzyme catalysed metabolic network from an essentially non-enzymatic "prebiotic soup". As early as 1945, Horowitz proposed the retrograde model of pathway evolution (Horowitz, 1945), and new theories are still being proposed (Lazcano & Miller, 1999). Of the proposed theories, the retrograde model of Horowitz and the patchwork model of Ycas and Jensen have garnered the most support (Horowitz, 1945, 1965; Jensen, 1976; Lazcano & Miller, 1996, 1999).

1.1.2 Retrograde evolution

In the retrograde evolution model, pathways evolve "backwards" from a key metabolite (Figure 1.2). The model presupposes the existence of a chemical environment
Chapter 1. Introduction

Figure 1.2: The retrograde model of pathway evolution (Horowitz, 1945). An organism heterotrophic for key metabolite A uses up all of the environmental supply of the metabolite. The fortuitous recruitment of an enzyme (enz 1) capable of synthesising A from B and C confers a survival advantage to the organism. In turn, environmental concentrations of B and E drop, compensated by the recruitment of enzymes ‘enz 2’ and ‘enz 3’ respectively.

where both key metabolites and potential intermediates are available (Horowitz, 1945). An organism heterotrophic for molecule A will use up environmental reserves of the metabolite to the point where falling availability limits growth; in such an environment, an organism capable of synthesising molecule A from environmental precursors B and C will have a distinct selective advantage. Such a mutant will rapidly spread through the environment. In the continued absence of environmental A, any null-mutation to the enzyme will be lethal, thereby favouring the enzyme’s preservation. In turn, as the environmental concentrations of B or C drop, the process will be repeated with the similar recruitment of further enzymes. Horowitz further suggested that the simultaneous unavailability of two intermediates (say B and C) would favour symbiotic association between two mutants, one capable of synthesising B and the other of synthesising C from other environmental precursors.

In 1965, Horowitz restated his theory to take into account the discovery of operons (Jacob & Monod, 1961; Horowitz, 1965). At the time, the clustering of genes involved in known pathways (e.g. leucine and tryptophan biosynthesis) into operons, along with a consideration of the probable origin of operons, led him to suggest that they would cluster genes with overlapping specificities, suggesting structural homol-
ogy and common ancestry (Horowitz, 1965). Operons are discussed in Section 1.2.1 of this thesis.

The retrograde model of pathway evolution, however, fails to account for the development of pathways which include labile metabolites that could not accumulate in the environment long enough for retrograde recruitment to take place. In addition, the theory can only explain pathway evolution in a metabolic intermediate rich environment; the ultimate destruction of the organic environment would prevent evolution of pathways by retrograde evolution (Horowitz, 1945; Lazcano & Miller, 1999).

1.1.3 Patchwork evolution

In 1969, the outline of a possible alternative to the retrograde model of evolution was beginning to be sketched (Waley, 1969). Waley described two key concepts: (i) the distinction between the evolution of new enzymatic functions based on the conservation of substrate binding with a modification of chemistry and that due to a change in preferred substrate with a change in the type of reaction (with the former suggested as more likely); and (ii) the notion that enzymes performing a specific and specialised catalysis may well have evolved from a broadly catalytic ancestor. Further thought to the issue was given by Hegeman & Rosenberg (1970). Based on the evidence available at the time, Hegeman and Rosenberg suggested that the tandem gene duplication mechanism proposed by Horowitz in 1965 to account for the process of retrograde evolution was questionable. The idea that all enzymes in a pathway — catalysing a number of varying chemistries on sometimes very different substrates — evolved from a repeatedly duplicated gene appeared unlikely. Rather, they suggested that "metabolic pathways grow by making use of pre-existing information in the cell, whatever its origin or genetic location" (Hegeman & Rosenberg, 1970). Genes "borrowed" from other pathways could subsequently be duplicated and specialised and, if advantageous, be brought together into a single operon by translocation.

These concepts all crystallised in the mid 1970s. Considering the possible earlier states of biochemical systems, Yčas (1974) proposed an alternative to the retrograde evolution. Two years later, Jensen proposed his theory of pathway evolution (Jensen, 1976); this theory expands and refines that of Yčas but, in essence, both propose that pathways evolved from a system of broad-specificity enzymes, a concept that has come to be known as the "patchwork evolution" model illustrated in Figure 1.3 (Lazcano & Miller, 1996).
In the patchwork model, enzymes exhibit broad substrate specificities and catalyse classes of reactions (Yeas, 1974). For Yeas, this breadth of catalysed reactions was the probable consequence of an error-prone translation mechanism generating a whole spectrum of enzymes with varying efficacy: the “statistical proteins” described by Woese (1965). For Jensen, the diversity of chemistries derived mostly from “substrate ambiguity”, i.e., the capacity of the enzyme to catalyse a reaction using a range of different substrates; extant enzymes with such properties are known (O’Brien & Herschlag, 1999). In addition to spontaneous non-enzymatic reactions, these enzymes would mean that many paths, some synthesising key metabolites, may have existed, albeit at a very low level, within a large network of possible interactions. Duplication of genes in such key metabolite synthesising paths, followed by their specialisation, would account for extant pathways (Figure 1.3). Such a duplication would be selectively advantageous since increased levels of the enzyme would generate more of the key metabolites. Furthermore, fortuitous evolution of a novel chemistry along with the biological leakiness of such a system could allow for the production of a key metabolite from a novel intermediate, even if it is several enzymatic steps away from the original substrate provided (Jensen, 1976).

1.1.4 Further theories

A number of other pathway evolution theories have been advanced. In the forward evolution model, prebiotic compounds do not play a role; pathways evolve forward because each intermediate in the pathway is of use to the organism, e.g., in the heme and chlorophyll metabolism pathways (Garnick, 1965; Lazcano & Miller, 1999). Thus, any enzyme recruitment capable of catalysing a reaction provides a selective advantage to the organism and the development of multiple genes simultaneously is not required (Garnick, 1965). Recruitment of enzymes to catalyse pre-existing non-enzymatic reactions has also been suggested as a possible explanation for the evolution of metabolic pathways. One theory suggests that extant anabolic pathways may be the result of successive enzyme recruitments to non-enzymatic degradative pathways (Degani & Halmann, 1967); once enzymes developed, the decomposition pathways could be reversed to synthesise the required metabolite (Lazcano & Miller, 1999).

Somewhere between the retrograde and patchwork models lies the “retro-evolution by jumps” model of Roy (1999). This theory borrows from the retrograde model by suggesting that multiple contiguous metabolic steps will have evolved from a single common ancestor, and from the patchwork model by suggesting that the
Figure 1.3: The patchwork model of pathway evolution (Yčas, 1974; Jensen, 1976). In the patchwork model, enzymes may have a favoured substrate and catalytic mechanism (a), but exhibit broad substrate specificities and are capable of catalysing other reactions (b). Therefore, many metabolic chains synthesising key metabolites (e.g. yellow square) may have existed, such as the one catalysed by the olive circle, the green cross and the pink doughnut. (c) Duplication of any gene in such a pathway would be advantageous, as more of the key metabolite would be synthesised. This duplication, followed by enzyme specialisation (d), would account for extant pathways.
Figure 1.4: The relationship between prebiotic chemistry and the evolution of metabolic pathways. There is probably no clear unique origin of metabolic pathways. Rather, a "cloud" obscures the transition from early prebiotic chemistry to a biochemical, enzymatic system (from Lazcano & Miller, 1999).

ancestral enzyme will be capable of performing several chemistries. However, for Roy, the ancestral enzyme would be multifunctional and capable of catalysing several contiguous steps - possibly even the whole sequence of a biosynthetic pathway - at a time, albeit inefficiently. Increased efficiency would arise from the duplication of such enzymes followed by their specialisation, usually at the expense of all but one of their functions. Evidence comes from both the identification of extant multifunctional enzymes and the detection of homologies in enzymes catalysing adjacent steps in a number of pathways. However, much of the evidence presented to support this theory could also be explained by patchwork evolution.

Lazcano & Miller (1999) extend the patchwork theory and suggest a semienzymatic model where the first biosynthetic pathways were partially or wholly nonenzymatic, some intermediary compounds were available due to leakage from existing pathways within cells and enzymes used in the early stages of pathways were formed by gene duplications of "starter-enzymes" giving closely related enzyme activities. The semienzymatic model thus proposes an "earlier" origin of metabolic pathways than the patchwork model since the latter "could operate only after the emergence of protein biosynthesis and the development of enzymes, i.e. after the appearance of the DNA/protein world" (Lazcano & Miller, 1999). The unclear transition between prebiotic chemistry and the emergence of metabolic pathways is depicted in Figure 1.4.
1.1.5 Examples in extant pathways

The previous section presents a conceptual framework in which to consider instances of contiguous homology within extant pathways. Below, two illustrative examples are described.

In tryptophan synthesis, the last three steps are catalysed by phosphoribosylantranilate isomerase (PRAI), indole-glycerol-phosphate synthase (IGPS) and tryptophan synthase (TRPS) as shown in Figure 1.5. The first two of these enzymes form a bienzyme encoded either by two separate genes or, as is the case for *E. coli*, by a single gene (trpC). TRPS is a tetrameric \((\alpha \beta \beta \alpha)\) bienzyme complex composed of the gene products of trpA (\(\alpha\)-domain) and trpB (\(\beta\)-domain). Of the protein chains encoded by these genes, three, PRAI, IGPS and the \(\alpha\)-subunit of TRPS, adopt a parallel \(\beta/\alpha\)-barrel fold — a fold often referred to as a TIM-barrel as the topology was first observed in triosephosphate isomerase or as a \((\beta\alpha)_8\) barrel (Wilmanns et al., 1991). They catalyse three sequential reactions: PRAI catalyses the isomerisation of N-\((5'\text{-phosphoribosyl})\)-anthranilate (PRA) to l-\((\oxygen{\text{-carboxyphenylamino}})l'\text{-deoxyribulose-5'-phosphate}\) (CdRP), the latter is then converted to indole-3-glycerol phosphate in a reaction catalysed by IGPS. Finally, indole-3-glycerol phosphate is cleaved to indole and glyceraldehyde 3-phosphate by the \(\alpha\)-subunit of TRPS. The pyridoxal phosphate dependent \(\beta\)-subunit of TRPS — which has an unrelated fold — catalyses the replacement of the \(\beta\)-hydroxyl group of L-serine by indole to produce L-tryptophan (Wilmanns et al., 1991).

Another example is found in the histidine biosynthesis pathway, here the gene products of hisA and hisF have extensive homology (Alifano et al., 1996). The former gene encodes phosphoribosylformimino-5-amino-1-phosphoribosyl-4-imidazole carboxamide isomerase (ProFARI); hisF encodes a gene which, with the hisH gene product, forms a dimer that constitutes the holoenzyme imidazole glycerol phosphate synthase (identified ImGPS herein to distinguish it from the indole-glycerol-phosphate synthase of the tryptophan pathway). The histidine pathway in *E. coli* is illustrated in Figure 1.6. Not only is the homology between hisA and hisF the likely consequence of a gene duplication event during the evolution of the *his* operon, but both genes exhibit internal homologies (Alifano et al., 1996): "hisA and hisF appear to be composed of four relatively homologous modules, A1, A2, F1 and F2, which are the result of two successive duplication events, the first one involving the hisA1 modules and leading to the extant hisA gene, which in turn duplicated and gave rise to the hisF gene". Indeed, this hypothesis is confirmed by the work of Lang et al. (2000) who solved the structures of the monomeric gene products HisA and HisF.
**Figure 1.5:** Tryptophan biosynthesis. Three domains of the tryptophan biosynthesis enzymes in *E. coli* are homologous and catalyse contiguous reactions: the C-terminal domain of TrpC (PRAI), the N-terminal domain of TrpC (IGPS) and TrpA (TRPS α-subunit). These domains are illustrated next to the reaction they catalyse. The domain illustrations were generated using Molauto and Molscript (Kraulis, 1991). The pathway details were obtained from EcoCyc (Karp et al., 2002b). The table details the structural relationships between the enzymes (Enz.) listing the *E. coli* gene encoding them (Gene), the representative PDB domain (PDBD), the chain residue-range compared (Res.), and the organism for which the PDB structure was solved (Organism). Scores for structural alignments using SSAP (Taylor & Orengo, 1989) are given; scores around 80 and above strongly suggest homology. Sequence identities within overlapping regions (based on the SSAP structural alignments) are also given.
from the hyperthermophile *Thermotoga maritima*. The NH\textsubscript{2} and COOH terminal halves of HisA and HisF were structurally aligned with root-mean-square (rms) deviations of 1.5 to 2.1Å. This overall alignment, as well as the identification of many highly conserved residues, confirmed the twofold duplication model (Figure 1.7). Furthermore, HisA and HisF also have high structural similarity to the TrpC, TrpF and TrpA enzymes (Lang *et al.*, 2000).

Contiguous homologies have been detected *ad hoc* in other pathways, for example in glycolysis (Fothergill-Gilmore & Michels, 1993), in methionine biosynthesis (Belfaiza *et al*., 1986; Roy, 1999) and in the mevalonate-dependent sterol/isoprenoid biosynthesis pathway (Burley & Bonanno, 2002). More recently, analysis of pathway evolution has focused not on single pathways but on larger scale analyses considering, for example, a superfamily (Copley & Bork, 2000), focusing on the metabolic pathways of one organism (Tsoka & Ouzounis, 2001; Teichmann *et al*., 2001a; Saqi & Sternberg, 2001) and attempting to quantitatively assess the relationships between metabolism and enzyme homology (Rison *et al*., 2002; Alves *et al*., 2002).

### 1.1.5.1 Pathway evolution and directed evolution

The identification of contiguous homologous enzymes such as that found in PRAI:IGPS and TRPS, and ProFARI and ImGPS, is suggestive of a classical retrograde model of evolution, but a phylogeny of metabolic \((\beta\alpha)\sb{8}\) barrels does not support the evolution of IGPS from TRPS and of PRAI from IGPS (Copley & Bork, 2000). Two recent directed evolution studies further muddy the waters. The observation that CdrP, the product of PRAI, is a substrate of IGPS and that the \((\beta\alpha)\sb{8}\) barrel tends to segregate substrate-binding residues predominantly to within the barrel itself and the catalytic residues predominantly to within the connecting loop regions and the structural similarity between PRAI and IGPS (see Figure 1.5) led Altamirano *et al.* (2000) to attempt to engineer a \((\beta\alpha)\sb{8}\) barrel with PRAI activity from an IGPS scaffold. They succeeded in doing so by replacing the loop connecting the \(\beta 6\) and \(\alpha 6\) structures of IGPS with a PRAI consensus loop and by generating libraries of mutants with a 4-7 residue replacement for the \(\beta 1\) to \(\alpha 1\) loop. Several rounds of selection by complementation of the PRAI-deficient *E. coli* strain JA300 on tryptophan-free selective media lead to the isolation of iVePRAI (*in vitro* evolved PRAI), a new PRAI with catalytic properties similar to those of the natural enzyme, but with an even higher specificity constant. This directed evolution, with its rounds of selection mimicking natural selection, had pronounced “Horowitzian” retrograde hallmarks: the evolution of an enzyme from the enzyme following it in
Figure 1.6: Histidine biosynthesis. The HisF:HisH complex is a branch point enzyme, one of its products, IniGP, is used as shown in histidine biosynthesis. The other, AICAR, is used in the de novo synthesis of purines. Illustrations of the homologous \((\beta\alpha)_8\) barrels HisA and HisF, generated as described in Figure 1.5, are shown and pathway details were again obtained from EcoCyc. The table details the structural relationships between the enzymes HisA and HisF; the table headings are described in Figure 1.5.
Figure 1.7: Model for the evolution of the \((\beta\alpha)_8\) barrel scaffold by twofold gene duplication. The first gene duplication (possibly of hisA1) generates two identical half-barrels that are the fused and adapted into an ancestral \((\beta\alpha)_8\) barrel with broad functionality. A second gene duplication leads to the diversification of the ancestral \((\beta\alpha)_8\) barrel into two enzymes with distinct catalytic activity (from Lang et al., 2000).

the extant metabolic pathway and a substrate-driven rationale for the recruitment. However, six months after the publication of the ivePRAI paper, Jürgens et al. (2000) presented the results of their directed evolution work. Here, the starting point was the observation that there was strong similarity between the reactions catalysed by HisA (ProFAII) and TrpF (PRAI, in E. coli the C-terminal of the TrpC gene product rather than the product of a separate trpF gene) as illustrated in Figure 1.8.

Using random mutagenesis and selection, several variants of HisA that catalysed the TrpF reaction both \textit{in vivo} and \textit{in vitro} were generated, including one which retained significant HisA activity. It was also observed that a single amino acid exchange could establish TrpF activity on the HisA scaffold. These data suggest a patchwork like recruitment of HisA and TrpF from an ancestral enzyme of broader substrate specificity, but conservation of chemistry (Jürgens et al., 2000). In support of this theory comes the observation that, in \textit{Streptomyces coelicolor}, no gene for TrpF has been identified. However, a gene, provisionally called priA, has been cloned. The priA gene product can complement both a \textit{trpF}~\textsuperscript{–} and a \textit{hisA}~\textsuperscript{–} \textit{E. coli} strain. Likewise, the epigenic presence of trpF and hisA in a \textit{priA}~\textsuperscript{–} \textit{S. coelicolor} strain restores tryptophan and histidine auxotrophy (Barona-Gómez & Hodgson, 2003, and Francisco Barona-Gómez, University of Warwick, personal communication). PriA has convincing homology with many trpF and hisA genes and may be the
closest relative of the (βα)₈ barrel enzymes to their last common ancestor (LCA) (Barona-Gómez & Hodgson, 2003).

Sadly, the results presented by Altamirano et al. (2000) were recently retracted (Altamirano et al., 2002). This further undermines support for the retrograde evolution of the homologous histidine biosynthesis metabolic pathways. Nevertheless, the possibility that a PRAI-like enzyme can be evolved from an IGPS scaffold must not be excluded.

1.1.6 Revisiting pathway evolution

Horowitz's theory for retrograde evolution is generally supposed to lead to three observable consequences: (i) clustering of evolutionarily related proteins in metabolic pathways, (ii) within such clusters, identification of the enzyme catalysing the last step in a metabolic chain as the deepest branching (i.e. ancestral) when a phylogeny is constructed (Copley & Bork, 2000) and (iii) a tendency for substrate-driven recruitment. In 1965, Horowitz restated his theory to take into account the discovery of operons (Horowitz, 1965). At the time, clustering of genes involved in known pathways into operons (e.g. leucine and tryptophan) along with a consideration of the probable origin of operons led Horowitz to suggest that they would cluster genes with overlapping specificities favouring structural homology and common ancestry; a clustering not however thought to be essential (Horowitz, 1965). Indeed, in its purest form, "the stepwise backwards route does not demand that the enzymes are evolutionarily related" (Waley, 1969).
If evolutionary relatedness of recruited enzymes is not a *sine qua non* condition for retrograde evolution, the theory should perhaps be thought of as an extension of the patchwork model — both based on the *ad hoc* recruitment of enzymes, but driven by different selective advantages. There is the difference that the retrograde model is thought to be substrate-driven (i.e. enzymes recruited because nearby metabolic enzymes are likely to act on similar chemical moieties) and the patchwork model is thought to be chemistry-driven (i.e. enzymes recruited for their catalytic properties). Again though, substrate-driven recruitment is not a necessary condition for the retrograde model to be valid, just an interpreted speculation. Therefore, it may be unwise to see these two theories as competing ones. Indeed, the two theories may be complementary (Fani, 1995). The retrograde theory offers a Darwinian rationale for the evolution of pathways: the selective pressure is the disappearance of a key metabolite and the adaptive response is the recruitment of an enzyme capable of supplying the metabolite to the organism. In a sense, whether this enzyme originates from that catalysing the previous step or from another broad specificity enzyme is immaterial, as long as the selective pressure can be overcome. Hence, for Waley (1969), “the ‘stepwise backwards’ route seems [...] likely for the evolution of the pathway, but less likely in general for the evolution of the enzymes”. Similarly, Hegeman & Rosenberg (1970) offer the retrograde model as the evolutionary framework for pathway evolution but suggest that “new catalytic steps are gained by recruiting enzyme activities from pre-existing metabolic pathways.”

Finally, theories of pathway evolution are not mutually exclusive. Some of the earliest pathways may have evolved in a retrograde fashion, some from the semienzymatic proposal and others, perhaps later pathways, from patchwork recruitment or forward following Garnick’s hypothesis (Lazcano & Miller, 1999; Rison & Thornton, 2002). The theories of pathway evolution are experimentally investigated in this thesis in Chapters 4, 5 and further discussed in Chapter 6.

### 1.2 *Escherichia coli*

*Escherichia coli* was first isolated in 1885 by the German bacteriologist Theodor Escherich, as a normal inhabitant of the intestinal tract (Madigan *et al.*, 1997). *E. coli* is a gram-negative flagellated heterotrophic rod-shaped bacterium (phylum proteobacteria (purple bacteria), subkingdom eubacteria) approximately 1µm long (Margulis & Schawrtz, 1998, and Figure 1.9).

As the chief object of study by molecular biologists, *Escherichia coli* is better known
than any other single organism on earth. The organism has become one of the prime model organisms in molecular biology because:

- it is very easy to manipulate in the laboratory; it grows rapidly and has simple nutritional requirements;
- it exhibits much interesting biochemistry and physiology;
- it is capable of undergoing conjugation (bacterial "sex") which makes it suitable for performing genetic experiments such as genetic crosses;
- it is able to support the growth of a whole range of bacterial viruses (phages)

Unsurprisingly, the genome of non-pathogenic reference \textit{E. coli} strain, \textit{E. coli} K-12, was one of the first to be fully sequenced (Blattner \textit{et al.}, 1997). In this thesis, unless specifically mentioned, the term \textit{E. coli} refers to \textit{E. coli} K-12. The genome of \textit{E. coli} K-12 is constituted by a single double-stranded circular DNA chromosome; certain other \textit{E. coli} strains –particularly pathogenic ones– also carry one or more plasmids, pieces of extrachromosomal cytoplasmic DNA which can be transferred from one organism to another. The \textit{E. coli} K-12 chromosome is approximately 4.6Mbp long and encodes 4405 genes of which 115 are non-protein encoding RNA species (Blattner \textit{et al.}, 1997, and http://www.genome.wisc.edu/). The complete genomes of other \textit{E. coli} strains are now available, including that of \textit{E. coli} O157:H7, the causative agent of haemorrhagic colitis (Perna \textit{et al.}, 2001).

The \textit{E. coli} bacterium is a free-living organism and, therefore, has a set of the small molecular metabolic pathways sufficient for independent life. Similar, if not identical, sets of pathways are believed to exist in all free-living bacteria and eukaryotes.
Chapter 1. Introduction

(Teichmann et al., 2001a). The very extensive experimental work that has been carried out on *E. coli*, including the determination of its genome sequence, means that knowledge of these pathways is probably close to complete. Many databases are dedicated to *E. coli* and some, described below, were exploited extensively during the course of this thesis.

1.2.1 Operons

Not all genes are expressed continuously; such a strategy would be wasteful and unsustainable. It is crucial for living organisms to control their levels of gene expression. Furthermore, gene expression must be modulated, for example, to take into account external conditions or, in the case of multicellular organism, to switch-on tissue-specific genes. Therefore, gene activity is regulated at the level of transcription (i.e. by controlling the production of mRNA from its DNA template). Indeed, in prokaryotes, control of transcription is the *de facto* strategy for the on-off control of genes.

The analysis of the regulatory control of gene expression in prokaryotes is intimately tied to the seminal work of François Jacob and Jacques Monod and their discovery of the operon (Jacob et al., 1960). Jacob and Monod were awarded, along with André Lwoff, the 1965 Nobel Prize in Medicine and Physiology “for their discoveries concerning genetic control of enzyme and virus synthesis” (http://www.nobel.se/).

An operon is a collection of adjacent genes regulated by an operator and a repressor (Hartl, 1991). The gene or genes (usually the latter) controlled by an operon are downstream of a promoter (a region of DNA to which the RNA polymerase binds). The operator is a region of DNA either overlapping or adjacent to the promoter. Binding of a repressor protein to the operator obstructs access to the promoter and RNA polymerase is unable to bind, thereby preventing the synthesis of mRNA.

1.2.1.1 The lac operon

Much of Jacob and Monod's work on operons concerned the regulation of genes involved in the metabolism of lactose in *E. coli* and the regulatory mechanism of this system was first explained by their *lac* operon model (Figure 1.10).

In the *lac* operon, the repressor protein is the product of the *lacI* gene which happens to be upstream of the promoter regions (although this co-localisation is not
compulsory and many repressors are encoded by genes some distance from the op­
erator region to which they bind). The promoter region is identified as the lacP region, the operator region is identified as the lacO region. Three genes are found downstream of the promoter: lacZ, lacY and lacA which encode β-galactosidase, lactose permease and thiogalactoside acetyltransferase respectively. These three genes are polycistronic, i.e. transcribed on a single mRNA molecule from which more than one polypeptide is translated. Thus, when switched on, the lac operon guarantees that all the genes required for lactose degradation will be available. The lac operon is turned-off until lactose is available to the cell. In the presence of lactose, a small sugar molecule called allolactose is formed in the cell (Alberts et al., 1989). Allolactose -known as an inducer- binds the LacI (repressor) protein and the inducer-repressor complex can no longer bind the operator, leaving the promoter region accessible.

Operons are also subject to positive regulation, as opposed to the negative regulation described above. In positive regulation, more translation occurs in the presence of a DNA-bound gene regulatory protein than in its absence. Indeed, the lac operon is subject to both positive and negative regulation. In the absence of environmental lactose, the operon is kept switched off by the LacI repressor, but even in the presence of lactose, the operator is not always active. This is because the lac promoter is a very weak promoter. The lac operon is only switched-on in the presence of environmental lactose and in the absence of glucose; lactose is catabolised to glucose in the cell, and this process is therefore frivolous when environmental glucose is available. However, in the absence of glucose, E. coli elevates the internal concentration of cyclic adenosine monophosphate (cAMP). cAMP binds the gene activator protein catabolic-activator-protein (CAP) of E. coli (Alberts et al., 1989). The cAMP-CAP complex binds upstream of the promoter regions and aids the binding of the RNA polymerase to the promoter. Thus, the lac operon genes are effectively switched-on only in the presence of lactose and the absence of glucose, as illustrated in Figure 1.10.

1.2.1.2 Operon prediction

A number of operons have been investigated experimentally. The RegulonDB, a database of transcriptional regulation and operon organisation in E. coli K-12, lists 694 experimentally verified operons of which 262 contain more than one gene (Salgado et al., 2001, and Gabriel Moreno-Hagelsieb, personal communication). Nevertheless, it is likely that many more operons exist in E. coli, and knowledge of
Figure 1.10: Negative and positive transcriptional control of the lac operon. (a) In the absence of lactose, no lac mRNA is formed because the repressor bound to the lac operator prevents transcription. (b) In the presence of glucose and lactose, the lac repressor binds lactose and undergoes a conformational change, so that it does not bind to the lac operator. However, cAMP is low, because glucose is present, and thus cAMP-CAP does not bind to the CAP site in the operator. As a result, RNA polymerase does not bind efficiently to the lac promoter and only a little lac mRNA is synthesized. (c) In the presence of lactose and the absence of glucose, maximal transcription of the lac operon occurs. In this situation, the lac repressor does not bind to the lac operator, the concentration of cAMP increases, and the cAMP-CAP complex that forms binds at the CAP site, stimulating binding and initiation by RNA polymerase (from Lodish et al., 2000).
Operons in other prokaryotes is even more incomplete. A number of techniques have therefore been developed for computational prediction of operons. RegulonDB relies on distances between adjacent genes and functional gene annotation to predict a further 2165 operons of which 583 contain more than one gene (Salgado et al., 2000, 2001, and Gabriel Moreno-Hagelsieb, *ibid*). Earlier methods based on finding signals that occur at the boundaries of operons, transcription promoters on the 5' end, and terminators on the 3' end, also exist (Yada et al., 1999). Another method uses a combination of gene expression data, functional annotation and other experimental data; such a method is primarily applicable to well studied genomes such as *E. coli* (Craven et al., 2000). Finally, some methods predict operons by identifying gene clusters where gene order and orientation is conserved in two or more genomes (Ermolaeva et al., 2001; Lathe et al., 2000).

Operons and polycistronic genes are explored in Chapter 5 of this thesis; as the most complete resource available for *E. coli*, the known and predicted operons of the RegulonDB database, kindly supplied in May 2001 by Gabriel Moreno-Hagelsieb of the Universidad Nacional Autónoma de México, were used.

### 1.3 Resources

Much of the data exploited in this thesis originate from on-line databases. These resources are the principal "material" for any bioinformatics research; they are electronic repositories of biological information. Without doubt, the two most investigated biological macromolecules are DNA and proteins. It therefore comes as no surprise that the most utilised databases relate to these entities. Nevertheless, a plethora of other on-line biological information is available to researchers, the scope of which can be witnessed by consulting the annual January "Database issue" of the Nucleic Acids Research journal (http://nar.oupjournals.org/). In the 2002 Database issue, no less than 336 entries are listed in the Molecular Biology Database Collection (Baxevanis, 2002). Biological databases are generally divided into primary and secondary, or derived, databases. Primary databases are repositories for raw data. Significant processing of these data is often performed. This processing may add extra information to the primary data in the form of annotations or in the identification of relationships in the data (e.g. homologies between proteins or similarities in structure). A database that includes relationships or other data not found in the primary database is referred to as a derived, or secondary database. In addition, certain databases are known as "composite databases" (Luscombe et al., 2001); in
these, data found in primary databases are compiled and filtered to produce combined non-redundant sets that are more complete than the individual databases. In this section, some primary, secondary and composite databases are surveyed with particular emphasis on resources exploited in this thesis. Also surveyed are metabolic databases. These can be thought of as secondary databases, but in view of the scope and complexity of some of them, they deserve to be considered in a category of database of their own.

### 1.3.1 Nucleotide sequence databases

DNA sequences are stored at three major sequence databases: the United States of America based GenBank (Benson et al., 2002), the European Molecular Biology Laboratory (EMBL) Nucleotide sequence database (Stoesser et al., 2002) and the DNA Data Bank of Japan (DDBJ) (Tateno et al., 2002). Data are exchanged between these three repositories to ensure comprehensive and consistent worldwide coverage (Benson et al., 2002). These data may be accessed through integrated retrieval systems such as the National Center for Biotechnology Information (NCBI) Entrez system (Wheeler et al., 2002), which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical literature via PubMed. No introduction regarding sequence databases is complete without a mention of their exponential growth and a plot illustrating this trend (Figure 1.11). In 1982, GenBank contained 680,338 base pairs in 606 sequences; by 2001, these numbers were 15,849,921,438 base pairs in 14,976,310 sequences. Of late, this astounding growth has been fuelled in no small part by the the availability of complete genome sequences for different organisms.

### 1.3.2 Protein sequence databases

#### 1.3.2.1 The Protein Information Resource databases

The Protein Information Resource Protein Sequence Database (PIR-PSD) is the present day incarnation of the “Atlas of Protein Sequence and Structure” begun in 1965 by Margaret Dayhoff. The PIR-PSD, produced in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID) is a public domain database of annotated protein sequences (Wu et al., 2002). In June 2002, the PIR-PSD contained over
Chapter 1. Introduction

Figure 1.11: The exponential growth of the GenBank database. The number of sequences and the number of nucleotides in the GenBank database for the 1982-2001 period are plotted (data from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

283,000 entries. The PIR-PSD is divided into four sections, PIR1 to PIR4, that differ in terms of the quality of data and annotation provided. In addition, the PIR provides the iProClass secondary database, which includes family classification at the superfamily, domain and motif levels, structural and functional features of proteins, as well as cross-references to over 40 biological databases (Wu et al., 2002). The composite database PIR-NREF is also available. This non-redundant protein database consists of proteins collected from PIR-PSD, SWISS-PROT, TrEMBL, GenPept, RefSeq and PDB, with composite protein names and literature data (Wu et al., 2002). The SWISS-PROT, TrEMBL and PDB databases are described below, GenPept is produced by parsing the corresponding GenBank release for translated coding regions as defined in GenBank entries and RefSeq provides reference sequence standards for genomes, transcripts and proteins (Pruitt & Maglott, 2001).

1.3.2.2 SWISS-PROT and TrEMBL

SWISS-PROT is a manually curated protein sequence database. The main asset of the database is its extensive annotation of entries describing, among other things, the function, domain structure, post-translational modifications and domains and site of the protein (Bairoch & Apweiler, 2000). The database has high-integration with other sequence related resources (nucleic acid and protein sequence and protein tertiary structure) by means of cross-references to other databases (e.g. EMBL
DNA database, PDB, PROSITE). In addition, SWISS-PROT strives for minimal redundancy. Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports (Bairoch & Apweiler, 2000). In SWISS-PROT, these separate entries are merged; if there are discrepancies between the merged entries, these are flagged in the feature table of the corresponding SWISS-PROT entry. SWISS-PROT can be thought of as a primary database, providing minimally-redundant protein sequences, and as a secondary database, providing extensive annotation and cross-linking to other databases. SWISS-PROT entries are manually curated and annotations regularly revised. This time-consuming procedure means that the rate of entry of new sequences in SWISS-PROT is slow compared to the rate of discovery of new protein sequences in general. SWISS-PROT is therefore extended by the TrEMBL. TrEMBL is a computer annotated supplement to SWISS-PROT which consists of entries derived from the translation of all coding sequences (CDSs) in the EMBL database, except for CDSs already included in SWISS-PROT. TrEMBL is split into two sections, SP-TrEMBL and REM-TrEMBL. The former contains entries awaiting incorporation into SWISS-PROT; the latter (REMaining-TrEMBL) collects entries that are not to be included in SWISS-PROT, but are nevertheless identified as CDSs.

1.3.2.3 Other protein sequence databases

Many other protein sequence databases exist (for an extensive list see Baxevanis, 2002) but many of these tend to specialise on a particular family of proteins (e.g. ABCdb: a database of ABC transporter (Quentin & Fichant, 2000)) or are part of a larger database such as an organism specific database (e.g. SubtiList for the *Bacillus subtilis* genome (Moszer et al., 2002)). Since many protein sequence databases are available, it makes sense to centralise sequences into a preferably non-redundant database. One such composite database is the PIR-NREF database mentioned above. Another is the NRDB (Non-Redundant DataBase) provided by the NCBI. The NRDB is derived from GenPept, SWISS-PROT and PIR-PSD (Benson et al., 2002); no two sequences in the NRDB are identical. However, it has been argued that the database still has redundancy since multiple copies of the same proteins are retained in the database as a result of polymorphisms and/or minor sequencing errors (Attwood & Parry-Smith, 1999). Many protein sequence database derived secondary databases also exist, such as the protein family databases PROSITE (Falquet et al., 2002) and Pfam (Bateman et al., 2002). The PROSITE database contains both regular-expression patterns and profiles to describe protein families identified.
in SWISS-PROT. Recently, the use of regular-expression patterns has been all but superseded by the use of hidden Markov model (HMM) like profiles (Falquet et al., 2002). HMM profiles are also used in the Pfam database: accurate manually curated seed alignments of related SWISS-PROT proteins are used to generate a HMM which can then pull more related proteins into the family (Bateman et al., 2002). In addition to these manually curated families (Pfam-A families), automatically generated families (Pfam-B) are also available to provide completeness in terms of sequence coverage. HMMs are discussed further in the "Methods" chapter of this thesis (Chapter 3).

1.3.3 Structural databases

1.3.3.1 PDB

The protein related resources listed above deal with the primary structure of proteins (i.e. their amino-acid sequence). A number of databases store information related to the tertiary (3D) structure of proteins, but these tend to be subsets of the comprehensive Protein Data Bank (PDB). Bar the "Atlas of Protein Sequence and Structure", the PDB is probably the oldest molecular biology collection. The PDB was established in 1971 at the Brookhaven National Laboratory (Bernstein et al., 1977; Berman et al., 2002). The PDB remained at the BNL until October 1998, when the task of maintaining it was taken up by the Research Collaboratory for Structural Bioinformatics (RCSB), a non-profit consortium composed of: the Department of Chemistry and Chemical Biology and the Center for Molecular Biophysics and Biophysical Chemistry at Rutgers, the State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; the Biotechnology Division of the National Institute of Standards and Technology (NIST); and the Department of Biochemistry at the University of Wisconsin-Madison (see http://www.rcsb.org/).

In 1972, two structures were deposited into the PDB; 3298 structures were deposited in 2001. Much like the sequence databases mentioned above, the growth in PDB entries has been exponential, fuelled by improvements in technology (e.g. the use of NMR for structure resolution) and structural genomics initiatives (Burley & Bonanno, 2002). However, there is a twist in this tale; while the number of entries in the PDB has grown exponentially, the number of new folds (i.e. previously unobserved structural topologies) has not shown such an increase. There are, in all likelihood, a limited repertoire of folds in nature (Chothia, 1992; Orengo et al., 1994); the iden-
tification of new folds may therefore be subject to a "law of diminishing returns". The discrepancy between the number of novel sequences and novel structures has been described as the "protein sequence/structure deficit" (Attwood & Parry-Smith, 1999).

1.3.4 Structure classification databases

The availability of a large number of protein 3D structures naturally generated the need for useful classification of structures (Attwood & Parry-Smith, 1999; Hadley & Jones, 1999). In fact, because a protein may be composed of several discrete globular structural domains (generally defined as compact, local, semi-independent folding units built from secondary structure elements), it is usual to separate proteins into separate domains prior to classification (Hadley & Jones, 1999). Two structure classification databases are used extensively in this thesis: SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homologous superfamily). In SCOP, domain identification is mostly performed manually. In CATH, the procedure is semi-automated: three domain identification algorithms are applied; if the domains they detect match, the domain assignments are accepted; if they clash, the domains are resolved after manual inspection (see Jones et al., 1998, for details of the algorithms and of the consensus approach). Following identification of domains, both CATH and SCOP classify domains within a tree-like hierarchy. Both databases strive to identify homologous domains, even when the homology is difficult to detect. It has been shown that proteins with significant sequence similarity (>30% sequence identity) adopt similar folds (see for example Sander & Schneider, 1991) thus most structural classifications use sequence comparisons with conservative thresholds to assign "close" homologues (e.g. historically CATH has used 35% sequence identity over 80% of the larger sequence). The difficulty comes in identifying homology between domains with sequence identity in the "twilight zone" (usually defined to be 15-25% identity) and below. Here, knowledge of the 3D structure of the domain is invaluable since structural similarities can exist even between proteins with very low sequence similarity (Murzin, 1998). The skill, even when in possession of the 3D structure of domains, is to distinguish between analogous domains (which share a common fold but lack further evidence of evolutionary propinquity and may be the product of convergent evolution) from homologous domains. Thus, both in CATH and SCOP, structurally similar domains with low sequence-identity are not considered homologous unless there is supplementary functional evidence (e.g. the presence of key conserved residues involved in structure stabilisation, substrate or
co-factor binding or catalysis) to confirm this status.

1.3.4.1 CATH

The CATH database is a hierarchic classification of protein domain structures (Orengo et al., 1997). The name CATH is derived from the first four levels of the classification, namely Class, Architecture, Topology and Homologous superfamily. Details of these levels are given in Table 1.1. There are further levels in the CATH classification; CATH domains are clustered on the basis of sequence similarity. Thus, the S-level (Sequence family level) clusters domains having sequence identities >35% and at least 60% overlap. The N-level (Near-identical level) clusters sequences with >95% identity and 85% overlap and the I-level (Identical level) clusters sequences with 100% sequence identity and 100% overlap. The S, N and I-levels are often referred to as S-35, S-95 and S-100 families respectively. Within each N-level cluster, the domain with the best resolution is identified and its protein sequence obtained; this is the N-rep (for N-representative); the combination of all such representatives forms the N-reps dataset as used in Chapter 3.

PDB entries are classified into the CATH database using an eight step procedure (Orengo et al., 1997). The salient features of the procedure are briefly described below. First, PDB chains are clustered using HOMOL, a Needleman & Wunsch (1970) based algorithm, into families with >95% identity (S-95 PDB chain families). For each of these families, a representative is selected and analysed, using a consensus approach (Orengo et al., 1997; Jones et al., 1998), to determine the number of domains in it and the corresponding domain boundaries. These domain boundaries are inherited by each member of the S-95 family and are used to extract the related domain sequences. Domain sequences—as opposed to whole chains—are clustered into I, N and S families using the HOMOL algorithm. The structural class of the protein (C-level) is assigned next, using an automatic procedure applied to each S-level representative. This stage requires manual intervention in approximately 10% of cases. In addition to the 4 structural classes (classes 1–4), the CATH database has 5 further classes. Classes 6 to 9 are internal and are a “side-product” of the CATH database population protocol. Class 5 contains clusters of fully classified multi-domain protein chains (i.e. the clusters are based on the whole protein chain rather than its constituent domains, but the constituent domains are known and have been classified individually into one of the first four classes). Class 5 clusters are used by the MultiParse assignment clean-up module described in the “Methods” chapter (see Section 3.3.3.3). CATH classes 1 to 5 are described in Table 1.2.
CATH level | SCOP level | Description | Definition
---|---|---|---
Class | Class | Class describes the gross secondary structure content. | In CATH, four structural classes are defined: (i) mainly α, (ii) mainly β, (iii) mixed α/β and (iv) few secondary structures. In SCOP, domains composed of a mixture of α-helices and β-strands are either classified as α/β proteins (composed of beta-alpha-beta units) or α+β (where alpha and beta regions are segregated). The SCOP database also has classes for membrane and cell surface proteins and peptides, small proteins and multi-domain proteins (i.e. proteins composed of more than one domain but not segregated into individual domain classes, as these domains are never seen separated in the PDB).

Architecture | no equivalent | Architecture describes the gross spatial arrangement of secondary structures ignoring their connectivities. | Currently, architecture is assigned manually where possible using descriptions commonly cited in literature. There are only 5 all α architectures but 13 mixed αβ and 19 β ones. The SCOP database does not have an equivalent level.

Topology | Fold | At this level, the connectivity of secondary structures is considered. Proteins with the same major secondary structures in the same arrangement and with the same topology are clustered at this level; this does not necessarily imply homology, which is defined at the next level. | In CATH, the SSAP algorithm is used to compare protein structures (Taylor & Orengo, 1989). SSAP returns a normalised score between 1 and 100. Two domains with a SSAP score greater than 70 are classified in the same topology (subject to the condition that the smaller domain must overlap at least 60% of the larger domain) although such assignments are subject to manual corrections. In SCOP, classification is performed manually, although it is strongly guided by the use of sequence comparison and alignment methods.

Homologous superfamily | Superfamily | Proteins clustered at this level are thought to be homologous (i.e. have a common ancestor). Since homology can be declared even in the absence of reasonable sequence identity, this level is often said to cluster “distant homologues”. | In CATH, the criteria for inclusion of two domains into the same H-level are a SSAP score greater than 80 (again subject to the condition that the smaller domain must overlap at least 60% of the larger domain) as well as proof that the domains are functionally similar to the other members of the H-level. In SCOP, superfamily membership is determined manually; even proteins with low sequence identity can be found in the same superfamily of the basis of structural and functional similarity.

Sequence family | Family | Proteins clustered in the same family have high levels of similarity, indicating highly similar structures and usually very similar functions. Members of the same family may be thought of as “close homologues”. | The S-level in CATH clusters sequences with 35% or more sequence identity. In SCOP, the level of similarity required for clustering in the same family is generally 30% or more, although proteins with low identities but whose functions and structures are very similar may also be clustered (e.g. globins with sequence identities of ≈10%).

Table 1.1: CATH and SCOP hierarchy levels.
Chapter 1. Introduction

Table 1.2: The five main classes of the CATH database (adapted from Buchan et al., 2003).

<table>
<thead>
<tr>
<th>#</th>
<th>CATH Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All $\alpha$ domains</td>
<td>Domains With $&gt;15%$ $\alpha$ helix and $&lt;15%$ $\beta$ sheet.</td>
</tr>
<tr>
<td>2</td>
<td>All $\beta$ domains</td>
<td>Domains With $&lt;15%$ $\alpha$ helix and $&gt;15%$ $\beta$ sheet.</td>
</tr>
<tr>
<td>3</td>
<td>Mixed $\alpha$ and $\beta$ domains</td>
<td>Domains With $&gt;15%$ $\alpha$ helix and $&gt;15%$ $\beta$ sheet.</td>
</tr>
<tr>
<td>4</td>
<td>Domains with little secondary structure</td>
<td>Domains With $&lt;15%$ $\alpha$ helix and $&lt;15%$ $\beta$ sheet.</td>
</tr>
<tr>
<td>5</td>
<td>Fully classified multi-domain protein chains</td>
<td>Multi-domain proteins that have been divided into domains which have been classified in Classes 1-4.</td>
</tr>
</tbody>
</table>

Next, the H and T-level assignments are derived. These assignments rely on the Sequential Structure Alignment Program (SSAP) (Taylor & Orengo, 1989) which returns a normalised score between 1 and 100. SSAP scores for protein pairs are stored in a two-dimensional matrix and structure pairs that have a sufficiently high SSAP score (and a significant proportion, at least 60%, of the larger fold equivalent to the smaller) are merged into structure-based families using single-linkage clustering. Two cutoffs on the SSAP score are applied, 70 to generate T-levels and 80 to generate the H-level of the CATH database. However, it is not sufficient for proteins to have a SSAP score greater than 80 for them to be assigned to the same homologous superfamily (H-level). The proteins must exhibit a similar function to those other proteins within the superfamily. Many sources of information are used to validate the homology, such as the SWISS-PROT database (Bairoch & Apweiler, 2000), the PDB file itself or the literature. Finally, the architecture (A-level) is manually determined. Further details on the classification of PDB entries within CATH are given in Table 1.1.

1.3.4.2 SCOP

The SCOP database is conceptually similar to the CATH database and organises proteins in a hierarchy from class down to fold, superfamily and family (Murzin et al., 1995, and Table 1.1). Curation of the SCOP database is essentially performed manually, although some automation is used for the most routine tasks such as clustering of protein chains on the basis of similarity. Superfamilies are further divided into sequence families on the basis of sequence similarity. Although proteins are usually separated into domains, this is not always the case. For example, if two domains are only ever seen together in PDB structures, proteins containing this domains pair are classified in the same SCOP superfamily and the individual
Table 1.3: Comparing entries in various CATH and SCOP levels. For the three main classes in CATH and SCOP (\(^\dagger\) \(\alpha/\beta\) and \(\alpha+\beta\) combined for SCOP), the number of entries at conceptually identical levels is given. For CATH: \(#T = \text{Number of topologies}, \#HS = \text{Number of homologous superfamilies}, \#SeqF = \text{Number of sequence families};\) for SCOP: \(#F = \text{Number of folds}, \#SupF = \text{Number of superfamilies}, \#F = \text{Number of families}.\) Details of analogous CATH and SCOP levels are given in Table 1.1. \(^*\)CATH release 2.4, 14/01/2002; \(^\dagger\) SCOP release 1.59, 01/03/2002.

domains not classified alone — until one of them is found separated from the other in Nature. The class, fold, superfamily and family levels of SCOP are described in Table 1.1.

1.3.4.3 CATH and SCOP

Unlike CATH, the SCOP classification does not describe architecture but, just like CATH, the superfamily level collects together domains with convincing evidence for homology, even in the absence of telling sequence similarity. The SCOP database also distinguishes between alpha and beta (\(\alpha/\beta\)) and alpha plus beta (\(\alpha+\beta\)) classes. In the former, proteins are composed of beta-alpha-beta units; in the latter, alpha and beta regions are segregated. Such a distinction is not made in CATH. Nevertheless, the CATH and SCOP databases have very similar finalities. In both databases, membership in the same family or superfamily implies homology — a property which is exploited extensively in Chapters 4 and 5. The levels in the CATH and SCOP databases are compared in Table 1.1. A summary of the number of entries at each of these levels for the major classes is given in Table 1.3. The CATH and SCOP databases, as well as the completely automatically generated Families of Structurally Similar Proteins (FSSP) database (Holm et al., 1992), are extensively compared by Hadley & Jones (1999).
1.3.5 Metabolic databases

The study of biochemical pathways is age-old and yet the advent of metabolic databases is relatively recent (Karp, 1998). The major metabolic databases are summarised in Table 1.4 (see also reviews by Karp, 1998; Gerrard et al., 2001). Such databases range from simple online reproductions of textbook pathways to complex interactive databases listing pathways, reactions, enzymes, reactants, cofactors, etc. Metabolic databases are the logical consequence of the accumulation of large amounts of biochemical and genomic data: putative enzymes are deduced from genomes, and the patterns of interaction between the enzymes and their substrates are derived from biochemical experiments.

Some metabolic databases, such as the EcoCyc database, are specific to one organism (Karp et al., 2002b). Others databases pertain to many organisms, for example the PATHWAY database of the “Kyoto Encyclopedia of Genes and Genomes” (KEGG) (Kanehisa et al., 2002) and the “What is there?” (WIT) database (Overbeek et al., 2000), or focus on a particular aspect of metabolism, for example the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) (Ellis et al., 2001).

PATHWAY and WIT employ different strategies to deal with multiple organisms. In the former, pathways are consensus views not specific to a particular organism. For each consensus pathway view, enzymes thought to exist in a particular organism can be highlighted. In WIT, consensus views exist, but pathway collections are organised by species. However, the EcoCyc metabolic pathway set has the advantage of being thought to be complete and experimentally verified (Karp, 1998). Recently, the repertoire of species integrated within the EcoCyc architecture has been extended to include 11 further species but pathway derivations for these were computationally derived using the PathLogic program (Karp et al., 2002b).

As well as the PATHWAYS database, KEGG also contains GENES, a collection of gene catalogues of all the completely sequenced genomes and some partial genomes, and LIGAND, for information about chemical compounds, enzyme molecules and enzymatic reactions (Kanehisa et al., 2002). Similarly, WIT incorporates part of the “Enzymes and Metabolic Pathways” (EMP) database (Selkov et al., 1996) and the “Metabolic Pathways Database” (MPW) database (Selkov et al., 1998), both of which contain encodings of pathways derived from an extensive literature corpus. The EMP project began in the early 1960s when Evgeni Selkov, the EMP project leader, was collating enzyme activities, and linking them into pathways where papers described them. EMP, and the associated MPW database, were moved from paper to electronic format in the late 1980s. The EMP/MPW may, therefore, be thought
Chapter 1. Introduction

of as the first metabolic database, which led to the development of the WIT genome annotation environment (Selkov et al., 1996, and T. Charles Hodgman, personal communication).

The EcoCyc database, used extensively in this thesis, is described in detail in the Methods chapter (Chapter 3) and further information on the WIT and KEGG databases can be found in Chapter 2.
<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcoCyc/MetaCyc</td>
<td><a href="http://biocyc.org/">http://biocyc.org/</a></td>
<td>Literature-derived (and therefore experimentally confirmed) SMM pathways of <em>E. coli</em>. Computationally derived pathways for other microbial species are also available. The MetaCyc database is not organism specific but describes literature derived pathways, reactions, and enzymes from a variety of organisms, admittedly with a microbial focus.</td>
</tr>
<tr>
<td>KEGG</td>
<td><a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a></td>
<td>The PATHWAY database of the KEGG system is a collection of metabolic and signalling pathways. The pathways, derived from 30 organisms, present a global composite view, i.e. not all organisms will perform all of the steps described in a pathway. Steps specific to a particular organism may be highlighted in a pathway diagram. The PATHWAY database is closely linked with a number of other related KEGG databases, for example the LIGAND database of compounds, enzymes and reactions.</td>
</tr>
<tr>
<td>WIT</td>
<td><a href="http://wit.mcs.anl.gov/WIT2/">http://wit.mcs.anl.gov/WIT2/</a></td>
<td>The “What Is There?” (WIT) environment is a system designed for the curation of function assignments made to genes and the development of metabolic models. WIT partly incorporates the Enzyme and Metabolic Pathways database (EMP), a repository of 30,000 literature derived metabolic and enzymatic records. WIT therefore covers more organisms than any other publically available metabolic database.</td>
</tr>
<tr>
<td>UM-BBD</td>
<td><a href="http://umbbd.ahc.umn.edu/">http://umbbd.ahc.umn.edu/</a></td>
<td>The University of Minnesota Biocatalysis/Biodegradation Database contains information on microbial biocatalytic reactions and biodegradation pathways for primarily xenobiotic chemical compounds.</td>
</tr>
<tr>
<td>Boehringer Mannheim</td>
<td><a href="http://www.expasy.org/cgi-bin/search-biochem-index/">http://www.expasy.org/cgi-bin/search-biochem-index/</a></td>
<td>The Boehringer Mannheim Biochemical Pathways Wallchart adorning many a laboratory wall is available online as a series of scanned images. Enzymes on these images are linked with ExPASy’s ENZYME database. ENZYME is a repository of information related to the nomenclature of enzymes and is closely linked with the SWISS-PROT and PROSITE databases.</td>
</tr>
</tbody>
</table>

Table 1.4: Major metabolic databases.
1.4 Overview of this thesis

This thesis investigates proteins: their functions, and their participation in metabolism.

Chapter 2 investigates available functional classification schemes for genomes. A method is developed to permit their direct comparison, and some of the shortfalls of the schemes are discussed. Solutions to these problems are discussed in the light of “second generation” classification schemes such as the “Gene Ontology”. The following three chapters are concerned with the SMM of *E. coli*.

Chapter 3 describes the datasets exploited in this thesis, as well as the methodologies used. Three key resources are described in detail, the EcoCyc metabolic database, and the SUPERFAMILY and Gene3D structural assignment databases. Two methods for the calculation of “pathway distance”, a metric exploited in Chapter 5, are also presented.

In Chapter 4, the structural anatomy of *E. coli* SMM is investigated; domain family combinations are analysed and the distribution of domain family members within and between pathways is considered. Special cases of domain recruitment are studied, such as the “block recruitment” of sequential enzymes and “parallel enzymes” (i.e. enzyme pairs with identical substrates or products). Instances of pathways with unusually high levels of domain duplication are described.

Four contexts are considered in Chapter 5: the metabolic context, the evolutionary context, the genome context and the functional context. These represent respectively: the spatial arrangement of enzymes within the metabolic network; homologies between SMM enzymes; the spatial arrangement of SMM enzyme-encoding genes on the chromosome; and the catalytic (biochemical) activities performed by the enzymes. Correlations between a number of these contexts are explored. The reuse of enzymes at distinct locations of the SMM and instances of isozymic enzymes are also studied. The chapter therefore presents a “network” view of SMM in which, to a certain extent, relationships between contexts can be quantified.

The themes explored in Chapters 2 to 5 are drawn together in Chapter 6. Notably, the findings regarding the anatomy and evolution of SMM are reconsidered in the light of recent literature. Finally, overall conclusions are drawn in Chapter 7, and possible future work is discussed.
Chapter 2

Investigating functional annotation schemes for genomes

The analysis of genes and gene products is usually performed in order to discover, confirm or clarify their function. The function of a gene product is its raison d'être; understanding this function is key to understanding how a limited number of interacting gene products can generate life, from simple unicellular organisms to incredibly complex multi-cellular organisms such as *Homo sapiens*.

The association of functional data with a gene product (functional annotation) first appeared in databases of gene products such as SWISS-PROT or PIR, where protein entries are accompanied by careful human-generated annotations of their empirically determined or predicted role (Bairoch & Apweiler, 1999; Barker et al., 1999). However, these annotations, whilst including keywords chosen from a controlled vocabulary, are currently not formally organised in a functional annotation scheme (although there have been many efforts to classify such databases on the basis of their annotation (Eisenhaber & Bork, 1999; Tamames et al., 1998; Licciulli et al., 1999)).

The first extensive gene product functional classification scheme was devised in 1993 by Monica Riley to catalogue the 1,171 *E. coli* genes known at the time (Riley, 1993). This was some four years before the complete genome for *E. coli*, currently estimated to have approximately 4,300 genes, was sequenced (Blattner et al., 1997). An updated version of the classification scheme was published in 1996 (Riley & Labeledan, 1996) and regular updates can still be found in GenProtEC (Riley, 1998a) and EcoCyc (Karp et al., 1999b). More recently, genome sequencing projects have
been the driving force behind the development of alternative functional annotation schemes.

Once a genome is sequenced, the first step is to identify genes and attempt to annotate the functions of their products. However, in order to understand the mechanisms operating, the genes need to be organised according to the biological processes they perform. Such an organisation needs a standardised classification scheme. Functional classification schemes are usually simple hierarchies which begin by defining function in very general terms and become increasingly specific as one progresses down the hierarchy. When dealing with genomes, such schemes allow the gene complement of an organism to be sub-divided into sets of functionally related gene products and also help to provide an overview of the biology of an organism. There are currently many different schemes used to annotate genomes. Even the interpretation of the *Mycoplasma genitalium* genome, which with 470 genes is the smallest so far completed (Fraser *et al.*, 1995), greatly profits from organisation into such a scheme.

In this chapter, a number of gene-product functional classification schemes are surveyed. All these schemes are applicable to genome annotation initiatives. The concepts of depth, breadth and resolution as descriptors of the schemes are introduced and selected classifications are compared according to these criteria. A “Combination Scheme” (CS) is generated, against which six representative classification are mapped. The mapping allows the generation of “FuncWheels”, which are graphical representations of hierarchical classification schemes. Many issues related to the design and implementation of gene product functional classifications are highlighted by this survey. These are discussed in the light of emerging “second generation” schemes.

The work presented in this chapter is, to a certain extent, independent from that presented in Chapters 3, 4 and 5. The “Methods” section of this chapter has therefore been left *in situ* rather than being incorporated into the thesis' methods chapter (Chapter 3).

### 2.1 Comparison strategy

A number of WWW sites with functional classification schemes were surveyed; literature references and URLs for these can be found in Table 2.1.
<table>
<thead>
<tr>
<th>Functional Classification</th>
<th>URL</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenProtEC</td>
<td><a href="http://genprotec.mbl.edu/">http://genprotec.mbl.edu/</a></td>
<td>Riley &amp; Labedan (1997); Riley &amp; Serres (2000)</td>
</tr>
<tr>
<td>EcoCyc</td>
<td><a href="http://ecocyc.org/">http://ecocyc.org/</a></td>
<td>Karp et al. (2002b)</td>
</tr>
<tr>
<td>Sanger Centre - <em>M. tuberculosis</em></td>
<td><a href="http://www.sanger.ac.uk/Projects/M_tuberculosis/">http://www.sanger.ac.uk/Projects/M_tuberculosis/</a></td>
<td>Cole (1999)</td>
</tr>
<tr>
<td>MIPS: <em>Arabidopsis thaliana</em> Database</td>
<td><a href="http://mips.gsf.de/proj/thal/db/">http://mips.gsf.de/proj/thal/db/</a></td>
<td>Mewes et al. (2002)</td>
</tr>
<tr>
<td>MIPS: PEDANT</td>
<td><a href="http://pedant.gsf.de/">http://pedant.gsf.de/</a></td>
<td>Frishman et al. (2001)</td>
</tr>
<tr>
<td>MGI: Mouse Genome Database (MGD)</td>
<td><a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a></td>
<td>Blake et al. (2002)</td>
</tr>
<tr>
<td>TIGR: Microbial Databases</td>
<td><a href="http://www.tigr.org/tdb/mdb/">http://www.tigr.org/tdb/mdb/</a></td>
<td>Peterson et al. (2001); see also TIGR genome papers (e.g. Fleischmann et al., 1995; Fraser et al., 1995)</td>
</tr>
</tbody>
</table>

Table 2.1: Some databases with gene product classification schemes with relevant URLs and references. *no longer freely available since purchase by Incyte Genomics (see https://www.incyte.com/proteome/).
Six functional classification schemes, representative of the range available at the time of this survey (August 1999) were considered in detail: EcoCyc (essentially identical to GenProtEC), TIGR, SubtiList, MIPS/PEDANT, KEGG and WIT. EcoCyc and GenProtEC are updated versions of Riley's original scheme (Riley, 1998a; Karp et al., 1999b), while TIGR (Fleischmann et al., 1995) and SubtiList (Moszer et al., 1996) are adaptations of it. The MIPS/PEDANT scheme was developed by researchers at the Munich Information Centre for Protein Sequences (MIPS) (Frischman & Mewes, 1997; Mewes et al., 1997). Finally, KEGG (Kyoto Encyclopedia of Genes and Genomes) (Ogata et al., 1999) and WIT (What Is There?) (Selkov et al., 1998) mainly address regulation and metabolic pathways. Mapping of these schemes onto a "Combination Scheme" made it possible to compare them. This analysis included the generation of FuncWheels, a novel way of graphically depicting gene product functions. Certain schemes, although independently implemented, were not included on the basis of their similarity with schemes already present in the selection. For example, the Mycobacterium tuberculosis genome classification scheme employed at the Sanger Centre is essentially the same as the Riley scheme (Cole et al., 1998) and the COGs scheme is a "generalisation" of the Riley scheme into broad functional categories (Tatusov et al., 1997). Therefore, in both of these cases, a similar coverage of functional space is to be expected. Furthermore, the "Gene Ontology" scheme (Ashburner et al., 2000; The Gene Ontology Consortium, 2001, and http://www.geneontology.org/) was not included for mapping, even though it represents a separate type of functional classification scheme. There were two reasons for this: firstly, its scope is much larger, and its structure more complex, than the chosen schemes; secondly, the "Gene Ontology" represents a radical rethink of gene product function classification. Therefore, its direct comparison with the chosen schemes would have been difficult and ineffective.

2.2 Methods

2.2.1 Classification scheme uploading

All analysed functional classification schemes were available on-line during the course of August 1999 when the data for this chapter were collected. Where relevant and possible, they were uploaded locally and converted to a format applicable for storage into the publicly available PostgreSQL relational database management system (see http://www.postgresql.org).
The schemes were conceptualised as trees — a set of connected nodes organised in a hierarchical arrangement. The nodes are functions or functional categories (e.g. "DNA synthesis" or "Transport"). Increasingly specific functions are encountered when progressing from the top (level-1) nodes down to the terminal nodes. The functions can be identified by means of a key (for example, function 5.3.1) in which the first number (5) refers to level-1, the second (3) to level-2, etc.

All of the uploaded classification schemes were easily stored in such a format with the exception of the "Gene Ontology". The latter is implemented as a directed acyclic graph (DAG), which has a more complex data structure than a tree. A DAG allows a node to have more than one parent and for the edges to distinguish between different types of relationships between nodes. It is not possible to convert a DAG data structure onto a tree structure without some concessions; in particular, the capacity to distinguish between relationship types was lost, and nodes with more than one parent had to be duplicated and inserted as separate instances within the tree structure. Nevertheless, in order to estimate the depth, breadth and resolution of the "Gene Ontology" (see below), such a conversion was performed.

2.2.2 Design of the “Combination Scheme” and scheme mapping

In order to compare the six chosen functional classification schemes, a “Combination Scheme” (CS) of gene product functions was designed. The CS is not intended as a replacement scheme, but was designed solely to facilitate a comparison of the current schemes to appreciate their similarities and differences.

The generation of the CS was iterative. It involved the collation of all functional nodes described in the selected schemes and their organisation into a tentative combined scheme. As the aim was to generate as simple as possible a CS, the first attempted scheme only had two levels. However, it was soon evident that such a scheme was not viable. Therefore, a three level scheme was designed and modified during two rounds of mapping. This generated a CS with a broad coverage of all functions described in the selected schemes without excessive bias towards any one of them. The details of the design of the CS are given below.

All nodes in all the schemes investigated were collected and obviously duplicated functions or functional categories were eliminated from the list. The resulting list was re-organised into a three level tree with six nodes at the top-level and seventy-three level-3 nodes. For each of the investigated schemes, all level-1, 2, 3 and 4
nodes (a total of 1,315 nodes) were compared with nodes in the CS and mapped to the lowest (most specific) CS node possible. To simplify the mapping process, only a one to one relationship between a node in the mapped schemes and the CS was allowed. In certain cases, such a rule made the mapping impossible. For example, the node "Cell growth, Cell division and DNA synthesis" in the MIPS/PEDANT scheme could be mapped onto three different CS nodes. In some instances, where such multi-functional categories overwhelmingly pointed towards one CS node, such nodes were mapped to the appropriate CS node, but usually these functions were skipped rather than incorrectly assigned.

In order to keep the CS as universal as possible, species specific nodes were avoided. For example, "Sporulation", a property specific to certain organisms including Bacillus subtilis, was present as a function in the SubtiList scheme and could have justifiably been included as an additional "Organism process" in the CS. However, because the function is specific to a very limited number of organisms, it was subsumed into the more generalised "Adaptation" category, on the basis that "Sporulation" is usually initiated in response to nutrient starvation.

To identify and eliminate scheme-specific nodes from the CS, the results of the first-round of mapping were analysed, and all CS nodes associated with only one or two distinct schemes were identified. Each of these nodes was reviewed and either: subsumed into another node, combined with other scheme-specific nodes, deleted, reclassified or, in rare instances where the function was considered critical, left unchanged. The mapping process was then repeated to determine coverage of the CS by each of the six selected schemes. For illustrative purposes, the mapping of the first 50 EcoCyc nodes onto their "Combination Scheme" equivalent is shown in Table 2.2. A similar iterative process has previously been used to classify SWISS-PROT function annotations (Tamames et al., 1996) in which SWISS-PROT entry keywords were mapped onto a one level scheme, based on the segregation of the Riley scheme into three nodes: "Energy", "Information" and "Communication". All the mapping procedures were performed using database backed Perl scripts. Further details on the the mapping results and the mapped schemes can be found at http://www.biochem.ucl.ac.uk/~rison/FuncSchemes.

2.2.3 Generation of FuncWheels

A FuncWheel is a graphical representation of all the nodes in a three level classification scheme. The wheel is separated into differently coloured segments each
Chapter 2. Functional annotation schemes

<table>
<thead>
<tr>
<th>EC key</th>
<th>EcoCyc function</th>
<th>CS key</th>
<th>“Combination Scheme” function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MACROMOLECULE METABOLISM</td>
<td>1.2</td>
<td>Macromolecules</td>
</tr>
<tr>
<td>1.1</td>
<td>aa-tRNAs</td>
<td>5.3.2</td>
<td>aminoacyl-tRNA synthetases/aminocarboxyl-tRNA transfers/aminoacyl-tRNA</td>
</tr>
<tr>
<td>1.1.1</td>
<td>tRNA</td>
<td>5.3.2</td>
<td>aminoacyl-tRNA synthetases/aminocarboxyl-tRNA transfers/aminoacyl-tRNA</td>
</tr>
<tr>
<td>1.1.2</td>
<td>amino acyl tRNA synth; tRNA modification</td>
<td>5.3.2</td>
<td>aminoacyl-tRNA synthetases/aminocarboxyl-tRNA transfers/aminoacyl-tRNA</td>
</tr>
<tr>
<td>1.2</td>
<td>Basic proteins</td>
<td>5.3</td>
<td>Protein related</td>
</tr>
<tr>
<td>1.2.1</td>
<td>basic proteins - synthesis, modification</td>
<td>5.3.1</td>
<td>Protein synthesis</td>
</tr>
<tr>
<td>1.3.1</td>
<td>phospholipids</td>
<td>1.2.3</td>
<td>phospholipids, glycolipids, lipoproteins</td>
</tr>
<tr>
<td>1.3.2</td>
<td>DNA - replication, repair, restr./modif.</td>
<td>5.1.1</td>
<td>DNA synthesis and replication</td>
</tr>
<tr>
<td>1.3.2</td>
<td>DNA - replication, repair, restr./modif.</td>
<td>5.1.2</td>
<td>DNA restriction/modification and repair</td>
</tr>
<tr>
<td>1.3.3</td>
<td>proteins &amp; peptides - translation and modification</td>
<td>5.3.4</td>
<td>Protein modification/phosphorylation</td>
</tr>
<tr>
<td>1.3.3</td>
<td>proteins &amp; peptides - translation and modification</td>
<td>5.3.3</td>
<td>Translation related</td>
</tr>
<tr>
<td>1.3.4</td>
<td>RNA synthesis, modification, DNA transcrip.</td>
<td>5.2.1</td>
<td>RNA Synthesis</td>
</tr>
<tr>
<td>1.3.4</td>
<td>RNA synthesis, modification, DNA transcrip.</td>
<td>5.2.3</td>
<td>RNA modification</td>
</tr>
<tr>
<td>1.3.4</td>
<td>RNA synthesis, modification, DNA transcrip.</td>
<td>5.2.2</td>
<td>Transcription related</td>
</tr>
<tr>
<td>1.3.5</td>
<td>poly saccharides - (cytoplasmic)</td>
<td>1.2.1</td>
<td>polysaccharides, lipopolysaccharides</td>
</tr>
<tr>
<td>1.3.6</td>
<td>lipoproteins</td>
<td>1.2.3</td>
<td>phospholipids, glycolipids, lipoproteins</td>
</tr>
<tr>
<td>1.3.7</td>
<td>glycoproteins</td>
<td>1.2.2</td>
<td>proteoglycans, glycoproteins</td>
</tr>
<tr>
<td>1.3.8</td>
<td>lipopolysaccharides</td>
<td>1.2.1</td>
<td>polysaccharides, lipopolysaccharides</td>
</tr>
<tr>
<td>1.4.1</td>
<td>degradation of DNA</td>
<td>5.1.4</td>
<td>DNA degradation</td>
</tr>
<tr>
<td>1.4.2</td>
<td>degradation of proteins, peptides, glycopept.</td>
<td>5.3.6</td>
<td>Protein degradation</td>
</tr>
<tr>
<td>1.4.3</td>
<td>degradation of polysaccharides</td>
<td>1.2.1</td>
<td>polysaccharides, lipopolysaccharides</td>
</tr>
<tr>
<td>1.4.4</td>
<td>degradation of RNA</td>
<td>5.2.4</td>
<td>RNA degradation</td>
</tr>
<tr>
<td>2</td>
<td>ELEMENTS OF EXTERNAL ORIGIN</td>
<td>6.1</td>
<td>Elements of external origin</td>
</tr>
<tr>
<td>2.1</td>
<td>colicin-related functions Non-bacterial functions</td>
<td>6.1.3</td>
<td>Plasmid/colicin related</td>
</tr>
<tr>
<td>2.2</td>
<td>phage-related functions and prophages</td>
<td>6.1.1</td>
<td>Phage/virus related</td>
</tr>
<tr>
<td>2.3</td>
<td>transposon-related functions</td>
<td>6.1.2</td>
<td>Transposon and IS related</td>
</tr>
<tr>
<td>2.4</td>
<td>plasmid-related functions Non-bacterial functions</td>
<td>6.1.3</td>
<td>Plasmid/colicin related</td>
</tr>
<tr>
<td>3.3</td>
<td>not classified</td>
<td>6.2.1</td>
<td>Unclassified</td>
</tr>
<tr>
<td>4</td>
<td>SMALL MOLECULE METABOLISM</td>
<td>1.3</td>
<td>Small molecules</td>
</tr>
<tr>
<td>4.1</td>
<td>Amino acid biosynthesis</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.1</td>
<td>isoleucine/valine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.2</td>
<td>leucine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.3</td>
<td>methionine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.4</td>
<td>glycine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.5</td>
<td>serine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.6</td>
<td>cysteine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.7</td>
<td>phenylalanine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.8</td>
<td>tyrosine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.9</td>
<td>tryptophan</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.10</td>
<td>chorismate</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.11</td>
<td>histidine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.12</td>
<td>alanine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.13</td>
<td>glutamate</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.14</td>
<td>glutamine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.15</td>
<td>arginine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.16</td>
<td>proline</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.17</td>
<td>aspartate</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.18</td>
<td>asparagine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>4.1.19</td>
<td>lysine</td>
<td>1.3.1</td>
<td>Amino acid metabolism</td>
</tr>
</tbody>
</table>

Table 2.2: EcoCyc (EC) to “Combination Scheme” (CS) mapping. The mapping of the first 50 EC nodes onto their CS equivalent is given as an example. The complete list of mappings can be found at http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/view_mappings.html.
Chapter 2. Functional annotation schemes

representing a top-level node and proportional in size to the number of level-3 nodes in them. The wheel is also divided into an inner disc and an outer ring. The inner disc of the wheel is divided into segments representing level-2 nodes, again of a size proportional to the number of level-3 nodes in them, whilst the outer ring is divided into equally sized segments each representing a level-3 node (see Figure 2.2 for an example of a FuncWheel).

Coverage of the CS by mapped schemes was illustrated by FuncWheels with non-matching level-3 nodes blanked out. Additionally, if more than a third of level-3 nodes belonging to a level-2 node parent were unmatched, the level-2 node was considered unmatched and blanked out. Where level-2 nodes spanned only two level-3 nodes, they were considered unmatched only if both level-3 nodes were also unmatched (see Figure 2.3 for examples of such "coverage" FuncWheels).

All FuncWheels were generated using a modified version of the software used to generate "CATH wheels" (Martin et al., 1998). Data for the generation of these wheels were extracted from the scheme database using Structured Query Language (SQL) queries (Bowman et al., 1996) and Perl scripts (Wall et al., 1996).

2.3 Results

2.3.1 Scheme survey

The functional classification scheme data gathered during this survey are summarised in Table 2.3, which also includes some data on related classification schemes (e.g. classification of gene products by subcellular localisation or by Enzyme Commission code). Schemes discussed in this chapter are in bold in the table; other schemes encountered are included for completeness.
<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resol.</th>
<th>1:M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single organism</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenProtEC</td>
<td><strong>Physiological role</strong></td>
<td>Cell process; Metabolism of small mols; Structural elements</td>
<td>3</td>
<td>10</td>
<td>118</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Gene type</td>
<td>Enzyme; RNA; Membrane</td>
<td>1</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EC scheme</td>
<td>Oxidoreductases; Transferases; Ligases</td>
<td>4</td>
<td>6</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EcoCyc</td>
<td><strong>Taxonomy of genes</strong></td>
<td>Cell process; Lipid biosynthesis; Broad regulatory functions</td>
<td>3</td>
<td>6</td>
<td>150</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Taxonomy of pathways</td>
<td>Signal-transduction pathways; Intermediary metabolism; Biosynthesis</td>
<td>3</td>
<td>6</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxonomy of reactions</td>
<td>EC-Reactions; Binding reactions; Transport reactions</td>
<td>4</td>
<td>6</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxonomy of compounds</td>
<td>Hormones; Lipids; All carbohydrates</td>
<td>5</td>
<td>16</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Institut Pasteur:</td>
<td><strong>Functional categories</strong></td>
<td>Intermediary metabolism; Information pathways; Cell envelope and cell processes</td>
<td>3</td>
<td>6</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>SubtiList</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institut Pasteur:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TubercuList</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPS: MYGD (S. cerevisiae)</td>
<td><strong>Functional catalogue (FunCat)</strong></td>
<td>Transport facilitation; Energy; Cellular organisation</td>
<td>4</td>
<td>14</td>
<td>75</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Protein classes (ClassCat)</td>
<td>Cellular organisation; ATPases; Transcription factors; Molecular chaperones</td>
<td>4</td>
<td>22</td>
<td>187</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Subcellular localisations (SubcellCat)</td>
<td>Plasma membrane; Nucleus; Mitochondria</td>
<td>2</td>
<td>15</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: continued

<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resol.</th>
<th>1:М</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC scheme</td>
<td>As above</td>
<td></td>
<td>4</td>
<td>6</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Protein complexes (CompCat)</td>
<td>Replication complexes; Transcription complexes; Cell-cycle checkpoint complexes</td>
<td></td>
<td>4</td>
<td>70</td>
<td>317</td>
<td></td>
</tr>
<tr>
<td>Phenotypes (PhenCat)</td>
<td>Stress response defect; Sensitivity to antibiotics; Auxotrophy defects</td>
<td></td>
<td>4</td>
<td>11</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td>Amino acid metabolism; Energy; Signal transduction</td>
<td></td>
<td>2</td>
<td>10</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>MIPS: MATD (Arabidopsis thaliana)</td>
<td>Functional catalogue (FunCat)</td>
<td>Transport facilitation; Energy; Cellular organisation</td>
<td>4</td>
<td>17</td>
<td>255</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>tRNA; 5'UTR; Gene/protein</td>
<td></td>
<td>1</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>YPD (S. cerevisiae)/WormPD (C. elegans)</td>
<td>Functional categories</td>
<td>Binding protein; Ligase; Protein kinase</td>
<td>1</td>
<td>1</td>
<td>55/55</td>
<td>Y</td>
</tr>
<tr>
<td>Cellular role</td>
<td>Amino-acid metabolism; Pol I transcription; Signal transduction</td>
<td></td>
<td>1</td>
<td>1</td>
<td>42/45</td>
<td>Y</td>
</tr>
<tr>
<td>Subcellular localisations</td>
<td>Plasma membrane; Golgi; Nuclear nucleolus</td>
<td></td>
<td>1</td>
<td>1</td>
<td>24/32</td>
<td>Y</td>
</tr>
<tr>
<td>Molecular environment</td>
<td>Actin-cytoskeleton associated; Protein synthesis factor; Soluble</td>
<td></td>
<td>1</td>
<td>1</td>
<td>10/10</td>
<td></td>
</tr>
<tr>
<td>Genetic properties</td>
<td>XX animals are male; 1 intron; Null lethal</td>
<td></td>
<td>1</td>
<td>1</td>
<td>23/119</td>
<td></td>
</tr>
<tr>
<td>Post-translational modifications</td>
<td>Phosphorylation; Ubiquitination; N-linked glycosylation</td>
<td></td>
<td>1</td>
<td>1</td>
<td>30/25</td>
<td></td>
</tr>
</tbody>
</table>

continued on next page
<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resol.</th>
<th>1:M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger Centre (M. tuberculosis)</td>
<td>Gene list</td>
<td>Cell process; Lipid biosynthesis; Broad regulatory functions</td>
<td>4</td>
<td>6</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>MGI/MGD</td>
<td>Phenotypic classification</td>
<td>Biochemical; Anatomical; Physiological</td>
<td>2</td>
<td>8</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple-organism databases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIGR</td>
<td>Gene identification list</td>
<td>Amino Acid biosynthesis; Regulatory functions; Cell envelope</td>
<td>2</td>
<td>16</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EGAD Cellular roles</td>
<td>Metabolism; Cell signalling/communication; Cell structure/motility</td>
<td>3</td>
<td>6</td>
<td>49</td>
<td>Y</td>
</tr>
<tr>
<td>MIPS: PEDANT</td>
<td>FunCat</td>
<td>Transport Facilitation; Energy; Cellular organisation</td>
<td>4</td>
<td>16</td>
<td>240</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>EC scheme</td>
<td>As above; All alpha; All beta; Alpha beta</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Structural classes</td>
<td>All alpha proteins; Alpha plus beta protein; Membrane and cell surface</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCOP scheme</td>
<td>proteins</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pathway related</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG: GENES</td>
<td>Gene catalogue (functional and metabolic)</td>
<td>Energy metabolism; Membrane transport; Signal transduction</td>
<td>2</td>
<td>14</td>
<td>105</td>
<td>Y</td>
</tr>
<tr>
<td>KEGG: PATHWAYS</td>
<td>Pathway classification (Metabolic)</td>
<td>Energy metabolism; Metabolism of complex lipids; Metabolism of macromols.</td>
<td>2</td>
<td>10</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

*continued on next page*
<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resol.</th>
<th>1:M</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG: LIGAND</td>
<td>Pathway classification (Regulatory)</td>
<td>Signal Transduction; Ligand-Receptor Interaction; Molecular Assembly</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>KEGG: ENZYME</td>
<td>Ligand (Compound)</td>
<td>Carbohydrate; Lipid; Nucleic acid</td>
<td>3</td>
<td>5</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ligand (Enzyme EC)</td>
<td>Oxidoreductases; Transferases; Ligases</td>
<td>4</td>
<td>6</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>WIT</td>
<td>General overview</td>
<td>Intermediate metabolism and bioenergetics; Information pathway; Structure and function of the cells</td>
<td>9</td>
<td>6</td>
<td>3,002</td>
<td></td>
</tr>
<tr>
<td><strong>Ontologies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>Functional primitive</td>
<td>Protein; Ribozyme; Nucleic acid</td>
<td>9</td>
<td>3</td>
<td>1,740</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cellular component</td>
<td>Extracellular; Intracellular; Unlocalised</td>
<td>9</td>
<td>3</td>
<td>385</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Process</td>
<td>Cell growth and maintenance; Cell communication; Physiological processes</td>
<td>11</td>
<td>6</td>
<td>1,667</td>
<td></td>
</tr>
<tr>
<td><strong>Miscellaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COGs</td>
<td>Functional annotation</td>
<td>Information storage and processing; Cellular processes; Metabolism</td>
<td>2</td>
<td>3</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: An overview of gene product classification schemes identified during a survey of 16 genome related WWW sites. All encountered gene product classification schemes are listed; those discussed in the paper are in bold type, and where applicable three examples of nodes in the top level of the classification are given. The table also indicates the breadth, depth and resolution of the schemes. Schemes empirically determined to support multiple functional annotations of single gene products are flagged in the "1:M annotation" column.
The surveyed classification schemes were for the most part related to genome sequencing initiatives or analysis of genomes. Of the remaining sites, 2 were databases related to pathways and metabolism and the third was the “Gene Ontology” project. It is worth pointing out that the ‘Gene Ontology’ scheme—originally developed by a consortium of researchers affiliated with the Saccharomyces Genome Database (SGD), the Drosophila melanogaster database FlyBase and the Mouse Genome Informatics/Mouse Genome Database (MGI/MGD) group—is actually composed of three functional classification schemes concerning cellular localisation, biological processes and biological function (Ashburner et al., 2000; The Gene Ontology Consortium, 2001). This distinction of biological process and function is extremely pertinent to the design and implementation of functional classification schemes and is discussed later in this chapter.

2.3.2 Single and multi-organism schemes

Some of the genome-related WWW sites were single organism databases, and the others dealt with multiple organisms. Whilst the MIPS databases included two single organism databases (MYGD and MATD for Saccharomyces cerevisiae and Arabidopsis thaliana respectively) and a multiple organism database (PEDANT), they all shared one functional classification scheme, FunCat, originally based on yeast gene products, but adapted to be applicable to a number of other organisms (Mewes et al., 1997, 1999). Whilst the FunCat is used in PEDANT to classify many gene complements including the partially completed human one, it nevertheless remains yeast oriented, although efforts are being made to extend the scope of the classification (Dmitrij Frishman, personal communication). The Gene Ontology is being developed with the aim of being applicable to many organisms (Riley, 1998b)—it is considerably more complex than previous schemes by an order of magnitude.

2.3.3 Scheme depth, resolution and breadth

To gain an understanding of the scope and structure of the surveyed schemes, data were collected on the number of levels, the number of nodes at the top level, and the total number of nodes in these schemes (see Table 2.3). These three elements can be used to represent respectively the depth, breadth and resolution of the classifications and, for a selection of schemes, these are plotted in Figure 2.1.

Depth can be thought of as the sub-setting potential of the scheme: the greater the depth, the further the scheme allows sub-setting into functional groups. For
example, the MIPS/PEDANT scheme has a depth of 4 and, when applied to the *S. cerevisiae* genes, yields sets of 742 ORFs involved with transcription (level-1), 539 ORFs involved with mRNA transcription (level-2), 411 ORFs involved with mRNA synthesis (level-3) and 30 ORFs involved with chromatid modification (level-4). The depth of a scheme represents the amount of magnification that can be applied to functions; much like a microscope, the higher the magnification, the more specifically a particular sub-set of functions can be resolved. The depths of the mapped schemes along with that of the “Gene Ontology” function and process classifications are plotted in Figure 2.1a. The depth indicated in the bar chart is the maximum depth encountered and not all branches of the functional tree necessarily extend that far. Depths ranged from 2 (TIGR) to 11 (“Gene Ontology” process scheme). Only the Gene Ontology schemes and the WIT scheme have depths greater than 4 levels. The WIT database, constructed to aid in the reconstruction of metabolic pathways, contains a painstakingly detailed classification of metabolism and information pathway functions (404 terms related to these functions are found at a depth greater than 6). The “Gene Ontology” function and process schemes has a maximum depth of 9 and 11 respectively which reflects the intricacy of these schemes.

The next parameter is the resolution of a scheme (Figure 2.1b). The intuitive hypothesis is that schemes with a large number of function nodes are likely to have more specific functional descriptions. To use an analogy from the computer world, if all gene functions are represented as a screen –where the fundamental unit is function rather than pixels– the greater the number function nodes, the higher the resolution. Resolutions ranged from 52 for SubtiList to 3,002 for WIT. Again the size of the WIT scheme is apparent, as is that of the “Gene Ontology” schemes, which have a combined resolution of over 3,500 nodes, illustrating the minutiae that have gone into designing these schemes. Depth and resolution are closely linked: the greater the depth and resolution of a scheme, the finer its granularity.

The breadth of the schemes, represented by the number of nodes at the top-level and plotted in Figure 2.1c, helps to illustrate the coverage of the scheme. The broadest schemes, TIGR and MIPS/PEDANT, has 16 nodes at the top-level. The narrowest is the section of the “Gene Ontology” dedicated to gene product function with a breadth of 3. TIGR and MIPS/PEDANT do offer good coverage of function, but judging a scheme by its breadth can be misleading. Whilst the “Gene Ontology” function ontology has a depth of 3, this is because, at the top level, the ontology distinguishes between proteins, ribozymes and nucleic acids. The protein node itself has 16 level-2 nodes (e.g. “signal transduction” and “structural protein”), many of
which tend to be top-level nodes in the other schemes. Therefore, a scheme with a limited breadth does not necessarily have a narrow coverage of function.

2.3.4 The “Combination Scheme”

The “Combination Scheme” (CS), designed to allow comparison between schemes, was generated by compiling all level-1, level-2 and level-3 nodes in the six selected schemes and joining, splitting, deleting or renaming them during two rounds of mapping. The first pass mapping was performed to identify CS nodes biased towards one particular node. 17 of the first-pass level-3 CS nodes were found to be associated only with one scheme and 12 nodes to be associated only with two. Since one of the aims when generating the CS was to avoid such bias, the CS was modified accordingly. Of the 29 level-3 nodes identified as potentially scheme-specific, 15 nodes were variously grouped into combined nodes, 4 nodes were deleted, 3 nodes subsumed into other nodes and 7 nodes kept unchanged. In the first-pass mapping, 149 of the 1,315 nodes were skipped.

The resulting version of the CS still had three levels and six level-1 nodes, but now only had 55 level-3 nodes and, with minor modifications, became the working version as shown in Table 2.4 and illustrated as a FuncWheel in Figure 2.2. Second-pass mapping of the selected schemes to this CS confirmed that the incidence of over-specific nodes had been minimised. Only 139 nodes were skipped during this second round mapping. The mapping also generated the data used in the generation of coverage FuncWheels for the six selected schemes. Nevertheless, mapping of selected schemes onto the CS was difficult to complete. All schemes use umbrella terms (especially at the higher levels) and some of these did not resolve well onto the CS; by extension, it was not always trivial to reclassify unambiguously the children nodes of such umbrella terms within the CS. The CS is amongst the “smallest” of the schemes, with only 77 nodes, and yet it could accommodate all of the other schemes combined, even those with markedly more nodes such as the 254 belonging to the MIPS/PEDANT scheme. This is a good indicator of the level of subsuming involved in the generating and mapping to the CS, and explains in part why it is not recommended as a substitute scheme.

The mapping was also subject to a number of arbitrary assignments, for example when distinguishing function relating to energy metabolism from that concerned with small molecule metabolism. As far as possible, mapping was kept consistent. For example, the tricarboxylic acid (TCA) cycle, a functional node found in many
Figure 2.1: Depth, resolution and breadth of seven representative classification schemes. The resolutions of schemes WIT, GO function and GO process all exceed 500 nodes; their actual resolution is shown in bold above the columns.
Chapter 2. Functional annotation schemes

1. METABOLISM

1.1 Energy
1.1.1 autotrophic (energy) metabolism
1.1.2 energy metabolism (carbon)
1.1.3 energy transfer/atp-proton motive force

1.2 Macromolecules
1.2.1 polysaccharides, lipopolysaccharides
1.2.2 proteoglycans, glycoproteins
1.2.3 phospholipids, glycolipids, lipoproteins

1.3 Small molecules
1.3.1 amino acid metabolism
1.3.2 nitrogen/sulphur metabolism
1.3.3 nucleotide/nucleoside metabolism
1.3.4 phosphorus metabolism
1.3.5 carbohydrate metabolism
1.3.6 lipid, fatty acid and sterol metabolism
1.3.7 biosynthesis of vitamins, co-factors and prothetic groups
1.3.8 secondary metabolism

2. PROCESSES

2.1 Cell processes
2.1.1 cell division
2.1.2 signal transduction
2.1.3 protein targeting/protein destination
2.1.4 cell regulation

2.2 Organism processes
2.2.1 adaptation
2.2.2 protection responses/detoxification
2.2.3 responses to stimuli

3. TRANSPORT

3.1 Large molecules
3.1.1 protein, peptide transport
3.1.2 transport of nucleic acids

3.2 Small molecules
3.2.1 ion channels/porins/ion transporters
3.2.2 sugar and carbohydrate transporters
3.2.3 amino-acid/amine transporters
3.2.4 nucleotide, nucleoside, purine and pyrimidine transporters
3.2.5 ABC transporters/transport ATPases

4. STRUCTURE AND ORGANISATION OF STRUCTURE

4.1 Cell envelope/membrane
4.1.1 cell membrane
4.1.2 cell wall

4.2 Cell exterior
4.2.1 surface structures
4.2.2 surface polysaccharides/antigens

4.3 Ribosome related
4.3.1 ribosomal RNAs
4.3.2 ribosomal proteins

4.4 Other structural elements
4.4.1 chromosome related
4.4.2 organelle related

5. INFORMATION PATHWAYS

5.1 DNA related
5.1.1 DNA synthesis and replication
5.1.2 DNA restriction/modification and repair
5.1.3 DNA recombination
5.1.4 DNA degradation

5.2 RNA related
5.2.1 RNA synthesis
5.2.2 transcription related
5.2.3 RNA modification
5.2.4 RNA degradation

5.3 Protein related
5.3.1 protein synthesis
5.3.2 aminoacyl-tRNA synthetases/transferases
5.3.3 translation related
5.3.4 protein modification/phosphorylation
5.3.5 protein folding/chaperoning
5.3.6 protein degradation

6. MISCELLANEOUS

6.1 Elements of external origin
6.1.1 phage/virus related
6.1.2 transposon and is related
6.1.3 plasmid/colicin related

6.2 unclassified/unknown
6.2.1 unclassified
6.2.2 unknown function

Table 2.4: The “Combination Scheme” (CS). The hierarchical CS was used as a common reference to compare various classification schemes. The CS has 6 level-1 nodes, 16 level-2 nodes and 55 level-3 nodes. Numbers represent the key of the functions (e.g. CS 4.1.1 is “cell membrane”).
Chapter 2. Functional annotation schemes

Table 2.5: A comparison of the schemes used by Tamames et al. (1997) and Rison et al. (2000). The left hand side scheme was used by Tamames et al. (1997) to analyse clusters of functionally related genes in *E. coli* and *H. influenzae*. The right hand scheme is used in this chapter and in Rison et al. (2000). Most of the top nodes in each scheme map quite easily. The “Protein processing” node of the Tamames scheme (for which three sub-nodes are shown in italics) maps onto two of the Rison nodes and the “Information pathways” node of the Rison scheme encompasses both the “DNA & RNA” and “Translation” nodes of the Tamames scheme.

<table>
<thead>
<tr>
<th>Tamames et al. (1997) scheme</th>
<th>Rison et al. (2000) scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>Metabolism</td>
</tr>
<tr>
<td>DNA &amp; RNA</td>
<td>Information pathways (DNA, RNA &amp; Protein)</td>
</tr>
<tr>
<td>Translation</td>
<td>Transport</td>
</tr>
<tr>
<td>Signal</td>
<td>Processes (Cell)</td>
</tr>
<tr>
<td>Environment</td>
<td>Processes (Organism)</td>
</tr>
<tr>
<td>Structural</td>
<td>Structure and organisation of structure</td>
</tr>
<tr>
<td>Protein processing</td>
<td>Information pathways (Protein)</td>
</tr>
<tr>
<td><em>Chaperones</em></td>
<td></td>
</tr>
<tr>
<td><em>Protein and peptide secretion and transformation</em></td>
<td></td>
</tr>
<tr>
<td><em>Detoxification</em></td>
<td>Processes (Organism)</td>
</tr>
</tbody>
</table>

of the schemes, was always mapped under “Energy Metabolism” (CS 1.1) regardless of whether it was under a different parent node (e.g. “Carbohydrate metabolism”) in the mapped scheme.

It is interesting to note that the top level of the final CS is similar to the 8 node top level of the scheme employed by Tamames et al. (1997) in their analysis of functionally related genes in *Haemophilus influenzae* and *E. coli*, although it was designed independently. The Tamames scheme was adapted from the TIGR scheme (Fleischmann et al., 1995), itself derived from Riley (1993), and found to be a good compromise between functional specificity and ease of use for the analysis of genomes. The top-levels of the two schemes are compared in Table 2.5.

2.3.5 Functional scheme comparison

A full list of mapping assignments, along with further details regarding the mapping process, can be found at http://www.biochem.ucl.ac.uk/~rison/FuncSchemes. The
Figure 2.2: The “Combination Scheme” FuncWheel. Level-3 nodes are labelled in the outer ring, level-2 nodes in the inner disc. All identically coloured segments belong to the same level-1 node; these nodes are labelled on the edge of the FuncWheel.
mappings made it possible to compare schemes by generating a set of coverage FuncWheels as shown in Figure 2.3. In these, CS nodes not represented in each of the six selected schemes are blanked out (see the Methods section of this chapter for full details on the generation coverage of FuncWheel). Blanked nodes can be determined by reference to the CS FuncWheel shown in Figure 2.2.

The most extensive coverage of the CS is provided by the MIPS/PEDANT scheme, with only five level-3 nodes unoccupied (Figure 2.3d). The MIPS/PEDANT scheme is also the only scheme to have all its level-2 nodes occupied. Conversely, the KEGG scheme, with 11 out of 15 level-2 nodes blanked, has the lowest overall coverage: level-1 segments for "Processes", "Transport", and "Information pathways" are almost entirely blanked out, although the scheme has good coverage of metabolism (Fig. 2.3e).

The WIT scheme has good overall coverage except for the "Processes" level-1 segment (and the relatively trivial "Miscellaneous" level-1 segment). WIT is also unsurpassed in its coverage of metabolism and is the only scheme with no blanks in that segment (Figure 2.3f).

It comes as no surprise that EcoCyc (Figure 2.3a) and TIGR (Figure 2.3b) exhibit very similar coverage as they are both based on the Riley scheme. The SubtiList scheme is partially based on the Riley scheme, but a number of functions have been combined and adapted for *B. subtilis* with consequent partial loss of CS coverage (Figure 2.3c).

FuncWheels could be used to depict graphically the functional coverage of fully sequenced genomes, providing insight into the functional distinctions that characterise genomes. A similar comparison of genomes on the basis of their gene product function distribution was performed by Tamames *et al.* (1996) on 44 organisms using a very simplified 3 node scheme and highlighted differences between viruses, bacteria, eukaryotic unicellular organisms, plants and animals.
Figure 2.3: Coverage of the “Combination Scheme” illustrated using FuncWheels: (a) EcoCyc, (b) TIGR, (c) Subtilist, (d) MIPS, (e) KEGG, (f) WIT. In these FuncWheels, nodes in the CS not represented in the illustrated scheme are blanked out. CS functions present and absent can be identified by reference to Figure 2.2.
Chapter 2. Functional annotation schemes

2.4 Discussion

2.4.1 Mapping and scheme comparison limitations

Clearly, the comparison of schemes depends on their mapping to the CS. This mapping is not straightforward and is constrained by the requirement for one-to-one correspondence between a node in the mapped schemes and a node in the CS. Therefore, the absence of mapping to a CS node can mean one of three things:

1. The CS node is not represented in the mapped scheme. For example, the KEGG scheme, at the time of data gathering, did not explicitly describe functions pertaining to "(Cellular) structure", therefore this segment in the KEGG FuncWheels (Figure 2.3f) is completely blanked out.

2. The mapping process has assigned nodes that could have mapped to the "missing" CS node elsewhere. For example, Figure 2.3b shows that the CS nodes "Cell membrane" (CS 4.1.1) has not been "mapped-to" by the TIGR scheme, yet this scheme has a "Cell Envelope" node. Two of the nodes under the TIGR "Cell envelope" node could have been mapped to CS 4.1.1: "biosynthesis of surface polysaccharides and lipopolysaccharides" and "lipoproteins". The former was more accurately mapped to "Surface polysaccharides/antigens" (CS 4.2.2), the latter was mapped to "metabolism of phospholipids, glycolipids and lipoproteins" (CS 1.2.3) and therefore, the CS 4.1.1 node appears unoccupied.

3. The mapping process could not resolve ambiguity of broad-coverage nodes. In the SubtiList scheme FuncWheel, the majority of transport related functions (third segment) are blanked out. However, the SubtiList scheme does include the function node "Transport/binding proteins and lipoproteins"; most of the CS level-3 transport related nodes could be subsumed by this broad function, but because this node cannot specifically be mapped to any CS node, they appear unoccupied.

In view of these limitations, it is worth reiterating that the comparison of the mapped schemes to the CS is a means of getting an approximate overview of the schemes' coverage. A different person repeating the mapping would doubtless have emerged with somewhat different coverage FuncWheels but, in all likelihood, would draw very similar conclusions from them.

In addition, although the breadth, depth and resolution descriptors offer a good handle to compare functional classification schemes, they do not reflect the quality
of schemes. It would be unwise to assume that a wide, deep scheme with high resolution is necessarily better than a scheme of small breadth, depth and of low resolution. Broad schemes tend to be used to offer end-users rapid access to large functional categories, but this means that super-sets of these must be constructed manually. For example, to generate the equivalent of the “Information Pathways” node in the CS, the “Transcription”, “Translation” and “DNA metabolism” nodes of the TIGR scheme have to be combined. Deeper schemes allow end-users to identify gene products associated with quite specific functions without having to resort to alternative functional information databases, but complicate access to gene product data. High-resolution schemes may indicate focus on a particular area of functional classification (e.g. the WIT scheme and metabolism) or simply reflect the extensiveness of the scheme. A high-resolution scheme may be crucial for the expert user, but may prove dauntingly complex to others. Different depths, breadths and resolutions reflect different functional classification strategies and goals on the part of their implementers and cater to different needs on the part of their users. This is well illustrated by the COGs scheme (Tatusov et al., 1997), where the combination of some nodes in the “Riley scheme” generates nodes with broader functional coverage (i.e. coverage remains the same, but depth, breadth and resolution are reduced). This is needed to classify the COGs, which group together related proteins, with similar but sometimes non-identical functions. Conversely, the WIT scheme requires very detailed description of function (i.e. a deep, broad scheme with high-resolution) to allow the development of metabolic models (Selkov et al., 1998).

2.4.2 Scheme level meaning

In some classification schemes, levels have a semantic value. For example, in the Enzyme Commission (EC) scheme, a four level hierarchical scheme of enzyme catalysed reactions, the first level represents the major class of enzyme activity (e.g. “transferases” or “hydrolases") and the second, the group, substrate or bond acted upon (e.g. “transferring phosphorus containing groups” or “acting on peptide bonds") (Enzyme Nomenclature Committee, 1992). Further details about the semantic meanings of the EC classification scheme levels are given in Table 2.6. Such semantic “level-meaning” is absent in the surveyed schemes. Levels are often used to sub-set functions, but the rationale for this sub-setting is dependent on the parent node (e.g. if the parent node is “amino-acid metabolism”, the children nodes usually relate to the metabolism of a specific amino-acid) rather than an intrinsic property of the level. Resolution and depth in the schemes is therefore not consistent for all
Chapter 2. Functional annotation schemes

| Table 2.6: Description of the different levels in the EC classification. An enzyme reaction is assigned a four-digit EC number, where the first digit denotes the class of reaction. The meaning of subsequent levels depends upon the primary number, e.g. the substrate acted upon by the enzyme is described at the second level for oxidoreductases, whereas it is described at the third level for hydrolases. Different enzymes clustered together at the third level are given a unique fourth number, and these enzymes may, for example, differ in substrate/product specificity or cofactor-dependency. Peptidases (EC 3.4.-.-) have a different classification scheme. The EC scheme is a classification of overall enzyme reactions, and not enzymes, and takes no account of the details of the reaction chemistry involved (from Todd et al., 2001).

<table>
<thead>
<tr>
<th>First figure</th>
<th>Second figure</th>
<th>Third figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>OXIDOREDUCTASES</strong></td>
<td>Describes substrate acted on by enzyme</td>
<td>Type of acceptor</td>
</tr>
<tr>
<td>Substrate is oxidised-regarded as the hydrogen or electron donor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. <strong>TRANSFERASES</strong></td>
<td>Describes group transferred</td>
<td>Further information on the group transferred</td>
</tr>
<tr>
<td>Transfer of a group from one substrate to another</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. <strong>HYDROLASES</strong></td>
<td>Describes type of bond</td>
<td>Nature of substrate</td>
</tr>
<tr>
<td>Hydrolytic cleavage of bond</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. <strong>LYASES</strong></td>
<td>Type of bond</td>
<td>Further information on the group eliminated</td>
</tr>
<tr>
<td>Cleavage of bonds by elimination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. <strong>ISOMERASES</strong></td>
<td>Type of reorganisation</td>
<td>Type of substrate</td>
</tr>
<tr>
<td>Convert substrate into one of its isomers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. <strong>LIGASES</strong></td>
<td>Describes type of bond formed</td>
<td>Describes type of compound formed</td>
</tr>
<tr>
<td>Enzyme catalysing the joining of two molecules in concert with hydrolysis of ATP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.4.3 Function, Apples, and Oranges

One of the main issues bearing on functional classification schemes derives from a more philosophical question: “What is function?” Function is an umbrella term and, for example, a gene product can be described in terms of its biochemistry, molecular activity, cellular function and physiological role (Rastan & Beeley, 1997).
These functions are distinct and different. Consider the human serine protease trypsin: biochemically, it catalyses the hydrolysis of peptide bonds following lysine or arginine residues in peptides, its molecular activity is as a proteolytic enzyme, its cellular function is protein degradation, and its physiological role is to aid digestion. Such distinctions are rare in functional classification schemes. In her review of systems for cataloguing the functions of gene products, Riley (1998b) points out that many schemes juxtapose the “apples and oranges” of function and combine different aspects of gene product function, such as biochemical and physiological function, into a one-dimensional list. This problem is inherent in the surveyed schemes, which all mix “apples and oranges”. Similarly the CS includes, for example, the nodes “cell regulation” (CS 2.1.4), a physiological function, and “ion channels” (CS 3.2.1), a molecular function. The current schemes cannot be merely re-arranged to tackle this; separating the apples from the oranges requires a fundamental rethink of classification. This remains one of the most pressing and complex issue which must be resolved for effective gene product functional classification.

The “Gene Ontology” illustrates a possible solution to this problem by distinguishing function in terms of three organising principles: gene product function, process, and cellular localisation (Ashburner et al., 2000; The Gene Ontology Consortium, 2001). Gene product function is defined as “a capability that a physical gene product (or gene product group) carries as a potential”. To avoid confusion with the more general use of the term function, this organising principle is also known as a “functional primitive”. Examples of functional primitives include broad terms (e.g. “enzyme” and “transporter”) and narrower ones (e.g. “adenylate cyclase”). Process is defined as “a biological objective accomplished via one or more ordered assemblies of functions” for example “cell growth and maintenance” or more specifically “pyrimidine metabolism”. The division between organising principles is, however, not always definitive. The term “signal transduction”, for example, exists within both the function and process categories.

### 2.4.4 Multi-dimensionality and multi-functionality

The evident solution to dealing with the umbrella term “function” would be to distinguish carefully all these different aspects of function and to describe a gene product’s function in terms of each of them. This solution is encapsulated in the concept of multi-dimensionality of classification schemes as proposed by Riley (Riley, 1998b). The three organising principles of the “Gene Ontology” represent three functional dimensions (biochemical for “functional primitive”, cellular and physiological for “pro-
cess" and spatial for the "cellular localisation"). Such a classification is invaluable in understanding the role of a gene product. This is illustrated by the comprehensively annotated Yeast Protein Database (YPD) (Hodges et al., 1999). Each gene product in YPD is annotated in up to six different dimensions: genetic properties, functional category, post-translational modification, cellular role and subcellular localisation (see Table 2.3 for examples of nodes in these categories). Although each of these dimensions is only a list (i.e. a scheme with only one level and resolution equal to breadth), the combined information described by these six parameters permits the gene product to be positioned very accurately within the functional space.

Another aspect of multi-dimensionality concerns the hierarchical classification of functions within schemes; certain functions can be involved in a number of more generalised functional classes. In the "Gene Ontology", the functional node "ATP-binding and phosphorylation-dependent chloride channel" is an instance of an "intracellular ligand-gated ion channel", a "chloride channel" and a "transmembrane conductance regulator". This is handled in the "Gene Ontology" by conceptualisation of the scheme as a DAG; a simple tree-like hierarchy could not contend with such complexity.

Finally, many proteins are multi-functional: capable of performing a variety of biological roles, sometimes, particularly with multi-domain proteins, simultaneously. The biological role of a protein may also be dependent on its environment or localisation. In Table 2.3, the schemes that were found empirically to include multiple functional assignments for gene products are flagged.

### 2.4.5 Surveyed schemes

In this chapter, the focus is on six schemes mapped to the CS (EcoCyc, TIGR, SubtiList, MIPS, WIT and KEGG) and the "Gene Ontology". Two broad families of schemes emerged from this survey: i) genome related schemes and ii) gene product interaction pathway related schemes.

The genome related schemes are EcoCyc, TIGR, SubtiList, MIPS and the "Gene Ontology". Two of them (EcoCyc, TIGR) are current implementations or derivations of the original Riley classification (Riley, 1993). As a consequence, they can essentially be thought to represent the same scheme (implemented with trees of different breadth, depth and resolution). The SubtiList scheme was derived from an adapted combination of parts of the WIT related Metabolic Pathways Database (MPW), and of the Riley scheme (Moszer et al., 1996; Selkov et al., 1998). In ad-
tion, the scheme includes a number of *B. subtilis* specific functions. In terms of their coverage of the CS, no doubt because of their relation to the Riley scheme, the three schemes are quite similar, even though SubtiList appears to have a noticeably smaller coverage of the CS than the other two schemes. This is partly due to mapping limitations and partly because the SubtiList scheme was designed with the specific needs of the *B. subtilis* research community in mind and therefore focuses on aspects of functions of major relevance to them. The original Riley scheme was designed for the unicellular prokaryotic eubacterium *E. coli* and this bias will exist in all derivative schemes. With schemes such as TIGR that are applied to diverse gene complements, such a bias could be problematic.

The MIPS scheme shares a lot in common with the Riley scheme, but extends it to encompass a number of further functions. Some of these functions (e.g. "signal transduction") exist in all organisms but are not explicitly listed in the Riley based schemes; others are present to allow better coverage of eukaryotic functions by the scheme (e.g. organelle-related functions). The MIPS scheme can, in essence, be thought of as a superset of Riley schemes which begins to address the issue of generating functional schemes applicable to multiple and diverse organisms.

As previously mentioned, genome sequencing initiatives are the main driving force in the development of functional classification schemes. Nevertheless, the majority of genome sequencing initiatives have been focused on unicellular micro-organisms (for the latest status on completed genomes see http://igweb.integratedgenomics.com/GOLD/, Bernal *et al.*, 2001). Both Riley's scheme and the MIPS scheme were designed for classification of the genomes of unicellular organisms. Therefore, in many schemes, there is a great paucity of functional nodes concerning the interaction between cells.

WIT and KEGG are databases of gene product interactions. They deal with functions performed by the concerted actions of gene products in pathways and complexes. Both the WIT and the KEGG functional classification schemes generally classify gene product function by association with a pathway or complex. This helps explain why both these classification schemes have good coverage of metabolism. At the time of data gathering, the KEGG scheme had only minimal coverage of non-metabolism related functions. The WIT scheme had good coverage of the CS dealing with transport, structure and information pathway related pathways in addition to metabolism. This association of gene products with pathways and complexes is very relevant to their function: all but the simplest of biological roles in cells are performed by gene product interactions.
Chapter 2. Functional annotation schemes

The "Gene Ontology" is representative of the "next generation" of functional classification schemes. Rather than updating existing schemes, the "Gene Ontology" has been designed from scratch and addresses many of the problems and issues discussed in this chapter. The "Gene Ontology" is multi-dimensional and segregates the concepts of "functional primitive", "process" and "localisation". Its more complex architecture allows it to accommodate functional descriptions that are instances of more than one parent node. The scheme is being developed for classification of the gene complements of both unicellular and multi-cellular organisms. Mapping of the "Gene Ontology" to the CS was not attempted, but its scope is such that without doubt, it would have occupied all of the CS nodes had it been.

The most extensive scheme in use, and applied to a significant number of genomes, at the time of the survey was the MIPS scheme. However, perhaps one of the most notable conclusions is that all genome related schemes (other than the "Gene Ontology") cover broadly the same set of functions and there is little to make one scheme overwhelmingly superior to another. The schemes are tantalisingly similar, but unfortunately different enough to make direct comparisons between them difficult. With respect to the pathway and metabolism schemes, WIT has the most extensive functional classification, but KEGG has built more generalised pathways that may be more accessible to many researchers. Certainly, considering gene product interaction in pathways and complexes will pay an important part in any future functional classification scheme.

2.4.6 Future schemes

Functional classification schemes will become an increasingly critical element of genome databases. Ideally, they should be applicable to all species but still be capable of accommodating very specific functions and allow cross-species functional comparisons where possible. They will also have to be able to contend with environment and location dependent changes in gene product function. This will only be possible if multiple functional assignments for gene products are permitted. Furthermore, the most effective functional classification schemes will be multi-dimensional, which will allow for accurate positioning of gene product in the function space. To deal with these multiple parameters, the schemes will undoubtedly need to explore more complex structures than simple trees.

The increasing availability of multi-cellular genomes demands the development of more complete schemes that will have to classify not only the functions related to
intercellular communication but also functions related to the more complex organisation of multi-cellular organisms (e.g. tissues and organs).

The “Gene Ontology” represents a promising development. Nevertheless, its very complexity and scope may be an obstacle to its widespread implementation. There is, perhaps, a need for a less extensive scheme, spanning the gap between simple, tree-like classification schemes and the “Gene Ontology”, or perhaps it may be worth maintaining two schemes, one highly detailed, the other simplified, catering to different sets of end-users (Gelbart, 1998).

2.5 Recent developments

The work presented in this chapter was begun in August 1999 and completed in December of the same year. Naturally, many things have changed since then (for example, no publications regarding the “Gene Ontology” were available prior to 2000, only limited online information at http://www.geneontology.org/). This section attempts to present a brief overview of some of these recent developments.

The “Gene Ontology” has now become the de facto large-scale annotation scheme for a number of model organisms databases, for example, SGD (Dwight et al., 2002), FlyBase (Adams et al., 2000) and the MGI databases (Hill et al., 2001). In addition, the “Gene Ontology” is being applied to databases other than model organisms databases. For example, efforts are being made to annotate a non-redundant set of proteins described in the SWISS-PROT, TrEMBL and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms (http://www.ebi.ac.uk/GOA/). For a number of the annotation schemes, conversion files are now available to permit the transfer of annotations to a “Gene Ontology” based annotation. Notably, conversion files are available for InterPro, SWISS-PROT keywords and the EC classification (see http://www.geneontology.org/#indices). Not only are genomes being annotated with the “Gene Ontology”, but these annotations are themselves being exploited for further research (Hvidsten et al., 2001). The http://www.geneontology.org/ website is the main repository of “Gene Ontology” information; the latest version of the ontology can be downloaded from the site or visualised with a number of browser applications. “Gene Ontology” assignments for gene products are also available, as are details of relevant literature (The Gene Ontology Consortium, 2001). The site also keeps up-to-date statistics of the level of annotation in a number of databases implementing the “Gene Ontology”.

Shortly after the completion of the work presented in this chapter, Serres & Riley (2000) described MultiFun, an update of the original Riley scheme. In this scheme, cellular functions are divided into ten major categories: Metabolism, Information Transfer, Regulation, Transport, Cell Processes, Cell Structure, Location, Extra-chromosomal Origin, DNA Site, and Cryptic Gene. The scheme is currently implemented in the GenProtEC and EcoCyc databases. Not only was the scope of the scheme expanded, but more than one functional category could be assigned to a single *E. coli* gene products, thus reflecting the multifunctionality of most gene products.

A number of the other schemes seem to have extended their scope, in particular the FunCat scheme used in the MIPS/PEDANT databases and, to a lesser extent, the KEGG scheme. Finally, access to some of the databases has been stopped or modified. The elegant proteome.com databases (which included YPD and WormPD) have been bought-up by IncyteGenomics and access is no longer free. The WIT2 database has become the subscription-requiring ERGO database belonging to Integrated Genomics (http://igweb.integratedgenomics.com/IGwit/). The WIT2 database is still accessible, but only a limited number of metabolic reconstructions are available and much of the development seems limited to the ERGO resource.
Chapter 3

Analysing SMM — Methods

In the previous chapter, functional classification schemes were investigated. These are crucial tools when trying to define the function of gene products. However, these schemes are not directly exploited in the rest of this thesis. Rather, in the following three chapters, analyses are performed on a restricted set of proteins whose shared function is their participation in the small molecule metabolism (SMM) of *E. coli*.

Since *E. coli* is a model organism, it has been extensively investigated. In particular, the metabolism of *E. coli* is extremely well characterised (Neidhardt *et al.*, 1996). Furthermore, *E. coli* was one of the first completely sequenced organisms (Blattner *et al.*, 1997). Thus, knowledge of the SMM of *E. coli*, at least in terms of the proteins and metabolites involved, is essentially complete. The SMM data used in this thesis were principally derived from the EcoCyc database (Karp *et al.*, 2002b). EcoCyc is therefore described in detail below, as is the related pathway distance metric used extensively in Chapter 5.

Another key element of this thesis was the identification of homologous SMM proteins. Homologous proteins were principally identified using resources derived from the CATH (Pearl *et al.*, 2001) and SCOP (Lo Conte *et al.*, 2002) structural classification databases. The identification of these homologies is described. The chapter ends with a summary of the datasets used in Chapters 4 and 5.
3.1 The EcoCyc database

3.1.1 Description

EcoCyc, originally described as "an encyclopedia of *E. coli* genes and metabolism" (Karp *et al.*, 1999b), has developed into "an organism-specific pathway/genome database (PGDB) that describes the metabolic and signal-transduction pathways of *Escherichia coli*, its enzymes, its transport proteins and its mechanisms of transcriptional control of gene expression" (Karp *et al.*, 2002b).

The EcoCyc system is implemented in Common Lisp with a graphical-interface toolkit called the Common Lisp Interface Manager (CLIM) (Karp *et al.*, 1997). The system also uses the Grasper-CL component to display metabolic pathways and the online server capabilities (see http://www.ecocyc.org/) are based on the CWEST software tool (Karp *et al.*, 1997). Currently, EcoCyc is available in four forms (Karp *et al.*, 2002b):

1. online through the World Wide Web at http://ecocyc.org/ (this version supports a subset of the Graphical User Interface (GUI) functionality of the X-windows and PC versions);
2. as an X-windows application for Sun workstations (bundling together Pathway/Genome Navigator software with the EcoCyc database);
3. as a PC version of the above package;
4. as a flatfile version suitable for global analyses.

The current version of EcoCyc is version 6.0 (released February 2002). At the time the SMM dataset was gathered for this thesis (Spring 2000), the EcoCyc version was 5.4 and distributions were only available in forms 1 and 2 were available. Therefore it was necessary to develop tools to acquire and exploit the EcoCyc data in-house, a process which required the design of a pathway storage format known as an ‘npl’ file.

3.1.2 Sources of EcoCyc data

EcoCyc can be viewed as an electronic review article because it is a carefully collated collection of information drawn largely from primary literature and supplemented with data from a variety of sources. Sources of data for key elements of EcoCyc are:
Chapter 3. Methods

Genes Until 1997, most of the information on \textit{E. coli} genes was obtained from the EcoGene DB (Rudd, 2000). Data provided by EcoGene included synonyms for gene names, physical positions for all sequenced genes, and the direction of transcription for each gene (Karp \textit{et al.}, 1997). Subsequent releases of EcoCyc have incorporated the complete \textit{E. coli} K12 genome as sequenced in 1997 by Blattner \textit{et al.} (1997) (Karp \textit{et al.}, 1999b). In addition, the dataset is supplemented by additional information derived from the literature, external databases such as SWISS-PROT (Bairoch & Apweiler, 2000) and by corrections and additions provided by internal and external curators. Genes in EcoCyc are classified on the basis of their physiological role using a scheme derived from Riley (Riley, 1993).

Reactions The initial set of reactions described in EcoCyc were derived from the ENZYME DB (Bairoch, 2000). The ENZYME database is a complete representation of the International Union of Biochemistry and Molecular Biology (IUBMB) nomenclature and classification of enzymes (Enzyme Nomenclature Committee, 1992). This set of reactions has been extended with reactions not currently classified in the EC scheme, but found to occur in the metabolism of \textit{E. coli} and other organisms (Karp \textit{et al.}, 1997).

Pathways The EcoCyc pathways were derived from literature, textbooks and additionally curated by human experts such as Monica Riley at the Marine Biological Laboratory, Woods Hole, USA. Many were originally derived from the celebrated Gerhard Michal "Biochemical Pathways" wallcharts which adorn many laboratory walls (Michal, 1998).

Transcription Units The addition of information on regulation of transcription initiation and organization of genes into transcription units (TUs) is relatively recent (Karp \textit{et al.}, 2002b). The original TU data were derived from the RegulonDB database of transcriptional regulation and operon organization in \textit{Escherichia coli} K-12 (Salgado \textit{et al.}, 2001). Although the RegulonDB database collects both experimentally confirmed and computationally predicted TUs, EcoCyc includes only the former.

3.1.3 The EcoCyc ontology of biological function

In EcoCyc, data are stored in frames managed within a frame knowledge representation system (FRS) known as OCELOT (Karp \textit{et al.}, 2002b). In essence, frames
Chapter 3. Methods

Figure 3.1: Reaction, Enzymatic-Reaction, Small-Molecule and Protein frames in EcoCyc. EcoCyc frames are boxed and identified in blue type. Example attributes are shown in red. The Enzymatic-Reaction frame acts as a “bridge” between the Reaction and Protein frames.

are computational objects, and the FRS is a variant of the object-oriented (OO) approach of object-oriented database management systems (OO-DBMS) (Karp, 2000). The OCELOT FRS is discussed extensively by Karp (2000), but the main features of the system are given below.

There are two sorts of frames: classes and instances. Classes are generic (e.g. all genes or all proteins) whilst instances are specific entities (e.g. a particular gene or polypeptide). Frames have slots, and these may contain attributes or describe relationships of the class or instance with other classes or instances. Thus, Pathway frames have slots for Reaction frames. Reaction frames have slots for Small-Molecules (reactants and products) and, using an associative Enzymatic-Reaction frame, link to a Protein frame. In turn, this frame details the polypeptide(s) that catalyse the reaction. The relationships between the Reaction, Enzymatic-Reaction, Small-Molecule and Protein frames are schematised in Figure 3.1.

These frames describe in particular the pathway aspect of EcoCyc, but EcoCyc is a pathway/genome database (PGDB), and many other frames exists describing the chromosome, genes, transcriptional units, etc. (Karp, 2001). The use of a hierarchy of classes means that frames can be functionally classified (Karp, 2000). For example, under the class Gene is the gene-classification system developed by Monica Riley (Riley, 1993) and under the class Reactions is the entire EC classification system for enzyme catalysed reactions (Enzyme Nomenclature Committee, 1992).

The decoupling of biological entities (e.g. enzymes, substrates) from the representation of function (e.g. reaction performed) gives the system flexibility and al-
allows representation of features as diverse as metabolic transition in metabolism, the "charging" of tRNA moieties with amino-acids, signaling pathways, genetic-regulation functions and transport events (Karp, 2000).

The EcoCyc architecture is also used by the MetaCyc database (Karp et al., 2002a). MetaCyc is a metabolic-pathway database compiling literature derived pathways regardless of the organism in which they are identified. In addition, the BioCyc online resource (http://biocyc.org/) contains pathway/genome databases for 12 additional organisms —with pathways computationally predicted using the PathLogic program (Karp et al., 1999a)— including Mycobacterium tuberculosis, Haemophilus influenzae and Saccharomyces cerevisiae.

3.2 Generating the SMM pathway set

3.2.1 Downloading EcoCyc data

Whilst both the online and stand-alone version of EcoCyc provide a flexible and information rich environment, they are not well suited for global analyses. EcoCyc data were therefore converted to a format suitable for the analyses presented in the next two chapters. Data came from three main sources:

- EcoCyc pathway predecessor lists;
- large scale downloads of HTML source from the EcoCyc online site;
- personal communications and ad hoc access to other related databases.

Predecessor lists are connectivity lists and describe relationships of Reaction frames within a pathway (Karp & Paley, 1994). The predecessor list gives a minimised representation of biochemical pathways. In EcoCyc, predecessor lists are used to infer unambiguous pathway-graph representations using LISP (for LISt Processing) production rules. In turn, these pathway-graphs serve as inputs for the EcoCyc pathway-drawing algorithm. In addition, the conversion of predecessor lists into pathway-graphs requires heuristics to determine which of the compounds in the pathway are main compounds (i.e. lie along the backbone of the pathway) and which compounds are side compounds. Predecessor lists for EcoCyc pathways are generated manually, principally by Monica Riley and colleagues. These lists formed the foundations of the SMM pathway data explored herein. Rather than use the
Chapter 3. Methods

EcoCyc LISP-based system (Karp, 2000), a Perl-based system (Wall et al., 1996) was used in this thesis. The predecessor lists therefore had to be converted into a format suitable for the Perl system. In addition, because the predecessor lists are minimal representation and the pathway-graph generation algorithm was not available, Perl compatible pathway-graphs were generated manually by combined analysis of predecessor lists, parsing of the downloaded EcoCyc HTML pages (an unsatisfactory procedure which has been described as “screen scraping” (Stein, 2002)) and personal communications from members of both Peter Karp and Monica Riley’s research groups.

In keeping with the EcoCyc architecture, data describing the Pathway, Reaction and Enzymatic-Reaction frames were downloaded. Some of the data were stored locally using a relational database management system (postgresql). The Perl system used for analyses of SMM pathways in this thesis therefore relied on two main datasets: (i) an “in-house” form of predecessor lists and (ii) a relational database of Reaction and Enzyme-Reaction derived data as well as other ancillary data. These two datasets are described below.

The use of an in-house system was prompted by expediency — knowledge of Perl (and the absence of knowledge of LISP), as well as the unavailability of the flatfile version of EcoCyc at the time data were collected. In hindsight, the use of the EcoCyc LISP-system, and of the associated Generic Frame Protocol (GFP) (Karp et al. 1997) may have been more efficient. Finally, it is worth noting that a Perl-based interface to EcoCyc and MetaCyc, PerlCyc, has recently been developed (Peter Karp, personal communication).

3.2.2 The “npl” representation of pathways

The Perl language is capable of supporting object-oriented (OO) programming. These facilities, although not as developed as that of more formal OO languages such as C++ or Java, were exploited by the Perl SMM analysis system written for this work. The system, more a collection of related analysis programs, relied heavily on a two Perl classes: CPathwayStruct and CEcocycReaction. The former encapsulates pathways and the metabolic connections between the Reaction frames that compose them, the latter encapsulates the Reaction frames. The details of the object are not given in this thesis; however, in order to provide persistence, these objects were stored as easily parseable text-files. These text-files are known as npl files (for new predecessor list) and constitute the in-house equivalent of EcoCyc predecessor lists.
3.2.2.1 Description of the npl format

Npl files are similar to predecessor lists in that connected reaction frames are listed. However, npl files collect unbranched linear chains of reactions into entities known as "threads". Thus, two levels of connectivity are described, first that of reaction frames within threads, and secondly that of threads between themselves. An example npl file for the glycolysis and Entner-Doudoroff pathways is shown in Figure 3.2, along with a graphical representation of the pathway. The pathway illustration was obtained from the X-windows version of EcoCyc (release 5.4). The pathway contains four threads and these have been colour coded. The npl files use the same frame identifiers as EcoCyc. Thus, the Pathway frame identifier is "GLYCOLYSIS/E-D" and the reaction catalysed by phosphoglucose isomerase (top right) is described in reaction frame "PGLUCISOM-RXN". The first and last frames of each thread in Figure 3.2 are labelled with these identifiers. So, for example, thread_3 begins with reaction frame "GAPOXNPHOSPHN-RXN" hence the line "{thread_3 GAPOXNPHOSPHN-RXN}". The next reaction in the thread is that catalysed by phosphoglycerate kinase. Adjoining reaction frames are listed in the order "current-frame, preceding-frame", hence the line "(PHOSGLYPHOS-RXN GAPOXNPHOSPHN-RXN)". Thread_3 is preceded by threads 2, 4 and 5, hence the line "thread_2 thread_4 thread_5" and thread_4 is followed by threads 3 and 5, hence "$thread_3 thread_5$". Further details about the line format of npl files are given in Table 3.1.

The npl format has a number of advantages: it is easy to read, understand and manipulate manually and it is easily parseable by Perl scripts. A number of programmes were used to manipulate npl files, including programmes to semi-automate the process of modifying, detecting overlap between and merging npl files as well as programmes used to analyse npl files and associated data.
>GLYCOLYSIS/E-D
\glycolysis+Entner-Doudoroff

{thread_2 PGLUCONDEHYDRAT-RXN}
\( \text{KDPGALDOL-RXN} \)

{thread_3 GAPOXNPHOSPHN-RXN}
\( \text{PHOSGLYPHOS-RXN} \)
\( \text{GAPOXNPHOSPHN-RXN} \)
\( \text{3PGAREARR-RXN} \)
\( \text{2PGADEHYDRAT-RXN} \)
\( \text{PEPDEPHOS-RXN} \)

{thread_4 PGLUCISOM-RXN}
\( \text{6PFRUCTPHOS-RXN} \)
\( \text{F16ALDOLASE-RXN} \)

{thread_5 TRIOSEPISOMERIZATION-RXN}

Figure 3.2: The npl file for the GLYCOLYSIS/E-D pathway of *E. coli*. A graphical view of EcoCyc's GLYCOLYSIS/E-D is given on the left, the npl file for the pathway is shown on the right. Threads are colour-coded identically in both representations. The first and last reaction frames in each thread of the graphical view are labelled with their EcoCyc identifier.
3.2.3 Merging EcoCyc pathways

In the EcoCyc database, certain pathways are represented both in isolation and as a subpathway of larger pathways. For example, glycolysis is represented on its own, as well as in combination with the tricarboxylic acid (TCA) cycle: glycolysis is considered a subpathway of the latter “combined” pathway and conversely, the latter is a superpathway of glycolysis. Other pathways are “atomic” in that they have neither subpathways, nor are they themselves a part of a superpathway. To minimise redundancy, it is reasonable to begin by considering only atomic pathways and “terminal” superpathways (i.e. superpathways not themselves a subpathway of an even larger superpathway).

However, even if only atomic and superpathways are considered, the pathways still exhibit some overlap as a consequence of the arbitrary definition of pathways (Ger-rard et al., 2001). To deal with this problem, pathways were further merged until no two pathways in the dataset overlapped by more than two contiguous EcoCyc reaction frames. This procedure was semi-automated with the use of a Perl program capable of implementing nine basic operations to npl files: (i) splitting one thread into two threads, (ii) merging two threads into one, (iii) adding a new thread, (iv) deleting a thread, (v) adding a reaction to a thread, (vi) deleting a reaction from a thread, (vii) connecting threads, (viii) disconnecting threads and (ix) renumbering threads. The merging procedure was iterative (see Section 3.4.1.2).

Finally, any set of two adjacent reaction frames catalysed by identical enzymes were fused. In EcoCyc, these usually represent enzymes or complexes which generate an identifiable intermediate compound. For example, as illustrated in Figure 3.3, the fcl gene encodes a protein that carries out both the epimerase and reductase reactions for GDP-L-fucose synthesis in the colanic acid building blocks biosynthesis pathway of EcoCyc (pathway ID: COLANSYN-PWY). The epimerase and reductase reaction are separated in EcoCyc, but merged in this thesis. This merging of reaction frames is only relevant for the analyses presented in Chapter 5 where the pathway distance metric is used (see Section 3.2.6) — without the merging, the pathway distance between two enzymes might not accurately reflect the number of intervening enzyme catalysed steps.
<table>
<thead>
<tr>
<th>Character at beginning of line</th>
<th>Line description</th>
<th>Character at end of line</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;</td>
<td>Start of file; followed by pathway (EcoCyc) ID.</td>
<td>N/A</td>
</tr>
<tr>
<td>\</td>
<td>Pathway description.</td>
<td>N/A</td>
</tr>
<tr>
<td>{</td>
<td>Start of thread; contains thread number and reaction ID for the first reaction frame in the thread.</td>
<td>}</td>
</tr>
<tr>
<td>^</td>
<td>Preceding threads: lists the identifiers for the threads connected “upstream” to the current one. [This line appears after the {thread} identifying line. If no threads precede the current thread, the line appears as ^^.]</td>
<td>^</td>
</tr>
<tr>
<td>(</td>
<td>Reaction-frame pair: Line contains “current” reaction frame identifier followed by the previous frame’s identifier.</td>
<td>)</td>
</tr>
<tr>
<td>$</td>
<td>Subsequent threads: lists the identifiers for the threads connected “downstream” to the current one. [This line appears at the end of all threads. If no threads follow the current thread, the line appears as $$.]</td>
<td>$</td>
</tr>
<tr>
<td>#</td>
<td>Comment line.</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Table 3.1:** The npl format. In the npl format, the significance of a line is indicated by its first character. Certain lines are compulsory, others optional. For example, a single-reaction thread will have a “Start of thread” line, but no “Reaction-frame pair” lines. Optional lines are described in italics.
Figure 3.3: Merging adjacent reaction frames. The fcl gene encodes GDP-L-fucose synthase which catalyses an epimerisation (FCLEPIM-RXN reaction frame) and a reduction (FCLREDUCT-RXN). Adjacent frames catalysed by the same enzyme(s) are merged into a single frame, here FCLEPIM_FCLREDUCT-RXN.

3.2.4 Ancillary data

The pathway datasets described the reaction frames and their relationships. By using the Reaction, Enzymatic-Reaction and Protein frames it was therefore possible to assign one or more genes to each metabolic step in the pathway. Most of these assignments were obtained directly from EcoCyc with some additional manual correction, but some ancillary data, required to perform all the analyses presented in this thesis, were derived from other sources.

3.2.4.1 Gene identifiers

In EcoCyc, genes are commonly described by their gene symbol (e.g. gapA or pgk) or their Blattner number (e.g. b3919 or b2926), but the Gene3D structural assignment procedure (see below) required GenBank protein identifiers (PIPs) (Benson et al., 2000). Gene symbols and Blattner numbers were converted to GenBank identifiers using a conversion list obtained from GenProtEC (Riley, 1998a) which was manually edited following update reports (Margrethe Serres, personal communication).
3.2.4.2 Chromosomal location and gene intervals

Genes were assigned a chromosomal location by consulting the Gene Table for *E. coli* (see http://www.genome.wisc.edu/pub/analysis/m52orfs.txt and Blattner *et al.* (1997)) using the GenBank identifiers described above. Genes were ordered, irrespective of their strand, on the basis of their boundaries (i.e. starting at position 1 on the circular chromosome and numbering genes by scanning clockwise for boundaries regardless of whether the boundary was a start codon, as would be the case for genes on the (+) strand, or as stop codon, as would be the case for genes on the (-) strand); this ordering was near-identical to the Blattner numbering (Blattner *et al.*, 1997). The gene interval is a measure of the number of genes separating two genes as derived from the aforementioned ordering (e.g. a gene interval of 0 for genes side by side, of 1 for two genes separated by a third gene, of 2 for genes separated by two other genes etc.)

3.2.4.3 Function and EC Numbers

In Chapter 5, the relationship between pathway distance and enzyme function is investigated. To perform this analysis, Enzyme Commission (EC) numbers were used. The EC scheme is a four level hierarchical scheme classifying enzyme catalysed reactions (see Table 2.6 and Enzyme Nomenclature Committee, 1992). Because EC numbers classify reactions, one gene product could have more than one EC number if it catalyses more than one reaction (e.g. the fcl gene product discussed in the previous section). In addition, since several distinct gene products can participate in a single metabolic step (e.g. in a protein complex), each may have a separate EC number defining its particular catalysis whilst the overall reaction performed by the concerted action of these gene products might be given a different EC number. The EC numbers used in Chapter 5 were the EC numbers assigned in EcoCyc to each reaction frame and not to individual gene products.

3.2.4.4 Operons

The current release of EcoCyc incorporates data defining *E. coli* Transcriptional Units (TUs) obtained from the RegulonDB database (Salgado *et al.*, 2001; Karp *et al.*, 2002b). At the time of data collection for this thesis, these data were not available from EcoCyc. Furthermore, EcoCyc only includes experimentally verified TUs, whereas the RegulonDB database also contains predicted *E. coli* operons (Salgado *et al.*, 2000). For the analyses in Chapter 5, a list of all *E. coli* known and
predicted operons was obtained in May 2001 from Gabriel Moreno-Hagelsieb and Julio Collado-Vides of the Universidad Nacional Autónoma de México.

### 3.2.5 Storing data in PostgreSQL

Whilst pathway objects were used and manipulated within the Perl analysis scripts and text files were employed to store persistent representations of these, much of the data were stored in a relational database. This was made relatively easy because most of these data were essentially tabular. Storing of the ancillary data in a relational database meant that it was possible to:

- easily update and correct the information when provided data by curators;
- use the database to verify the integrity of the datasets;
- easily generate reports, lists, statistics, etc.

The only data stored in flat-file format were the npl files describing the pathways. All other data, such as data associated with Reaction frames (e.g. genes, EC number, etc.), data associated with Protein frames (e.g. length, amino-acid sequence) and identifier conversion lists were stored in a relational database using the PostgreSQL relational database management system (RDBMS) (see http://www.postgresql.org/). Structural and sequence family assignments obtained for SMM proteins (see below) were also stored in the same database. The entities stored in a database, as well as the relationships between them, can be described in an Entity/Relationship (E/R) diagram (Date, 1995). A simplified E/R diagram showing some of the information included in the database exploited herein is shown in Figure 3.4.

### 3.2.6 Pathway distance

In this thesis, use is made of a measure called pathway distance: the number of distinct metabolic steps separating two enzymes (see Figure 3.5). The term “metabolic step” refers to the enzyme catalysed modification of one or more substrates into chemically distinct compounds. The concept of a metabolic step is similar to that of reaction frames found in the EcoCyc database (Karp, 2000). EcoCyc reaction frames are computational objects encapsulating an enzyme catalysed substrate modification. The frame contains the reactants and products of the modification and is
Figure 3.4: An E/R diagram of in-house pathway and ancillary data. The E/R shows entities (database tables, boxed with their name in bold) along with some of their key attributes (normal type). The relationships between tables are given by the connecting lines. Connectors terminated by a "crow's foot" describe "one-to-many" relationships. Primary Keys (PKs) are unique identifiers for each row in a table. Some primary keys are formed by more than one attribute, these are Composite Primary Keys (CPKs). Attributes used as PKs or as part of CPK are flagged. Not all the tables and relationships are shown.
associated with one or more pathways. The reaction frame is associated with one or more enzymes using a linking object. Conceptually, metabolic steps are the product of this linking, i.e. the enzyme(s) catalysing the transition and reactant(s) and product(s). Indeed, in most cases, the number of metabolic steps and the number of reaction frames in an EcoCyc pathway are identical and only differ when two EcoCyc reaction frames are merged into one (see above). Pathway distance has been used independently, for example by Kolesov et al. (2001) (where it is called "metabolic distance") and more recently by Alves et al. (2002).

Using this measure of pathway distance, adjacent enzymes have a pathway distance of 1. By extension, enzymes catalysing the same metabolic step in a pathway (for example, in Figure 3.5, pfkA and pfkB) can be thought of as having a pathway distance of 0.

Two methods for calculating pathway distances were used. The first, based on a depth-first search (DFS) of pathway trees, was developed by the author. The second method, based on linear programming (LP), was developed by Evangelos Simeonidis, Lazaros Papageorgiou and David Bogle of the Department of Chemical Engineering, UCL. Both methods are described below.

3.2.6.1 Calculating pathway distances — DFS

The depth-first search algorithm can be used to visit all nodes in a tree (Orwant et al., 1999). A tree is a simple, undirected, connected, acyclic graph and a graph is composed of nodes and connecting edges. The pathway illustrated in Figure 3.5 can be thought of as a graph with genes as nodes (enzymes, in pink) and connecting arrows (in blue, labelled with key metabolites in red) as edges. However, it is not a tree for two reasons. First, the pathway is not acyclic (the TCA cycle is by definition cyclic) and secondly, it is not undirected since the edges are arrows. To use the DFS, the pathway representation was "converted" into a tree. First, all edges were considered undirected, and secondly cycles were arbitrarily "snipped" (i.e. certain edges were deleted to make the graph acyclic — in this case, the edge connecting tpiA and epd, gapA and the edge connecting mdh and gltA). Once the pathway is converted into a tree, the root and leaves can be identified (here pgi and tipA and mdh respectively).

The DFS algorithm was used to identify all possible Reaction-frame traversals between the root and the leaves. From the traversals, pairs of reaction frames at user defined distances (i.e. specified number of steps) were extracted. Duplicate pairs
Figure 3.5: Pathway distance illustrated in glycolysis and the tricarboxylic acid (TCA) cycle. A protein-centric (Gerrard et al., 2001) representation of glycolysis and the TCA cycle is shown (adapted from EcoCyc). Genes encoding enzymes are shown in pink. Substrates are shown in red with only key metabolites shown. The arrows can be read as “produces a substrate for”. Thus enolase (gene eno) produces phosphoenolpyruvate used by the pyruvate kinases encoded by pykA and pykF, whilst isocitrate dehydrogenase (icdA) synthesises 2-oxoglutarate used by the oxoglutarate dehydrogenase complex (sucA, sucB and lpdA). However, the oxoglutarate dehydrogenase never produces 2-oxoglutarate, so only a single arrow connects the two. Minimal pathway distances (shown in green) are given from enolase (eno). The first number indicates the minimal pathway distance if the arrow directions are considered, the second number indicates minimal pathway distance if directionality is not considered (i.e. all edges are assumed to be bi-directional).
Figure 3.6: Convex and concave sets in Euclidean space. In a convex set, all line segments connecting any pair of its points is fully contained. This is not the case with a concave set.

(i.e. same reaction frames at the same pathway distance, but reached via alternative routes) were eliminated. By identifying the gene(s) involved in each reaction frame, the pathway distance between \textit{E. coli} enzymes can thus be calculated.

The DFS method, although relatively fast and easy to implement, suffered from two main disadvantages: (i) the arbitrary snipping of cycles meant some enzyme pairs were ignored; and (ii) directionality could not be taken into account. Nevertheless, this method was used for all the analyses presented in Chapter 5, with the exception of the enzyme function and chromosomal distance analyses (see Sections 5.3.4.1 and 5.3.4.2).

3.2.6.2 Calculating pathway distances — LP

Linear programming is the most commonly used form of constrained optimisation (i.e. the identification of the best possible solution within the limits of a set of conditions). The quality of a solution may be assessed by use of an objective function and the best solution(s) will maximise the objective function. As the name implies, linear programming requires both the objective function and the constraints to be linear. One of the consequences of this is that the feasible region (the region containing all valid, although not necessarily optimal, solutions) defined from the linear constraints is convex (see Figure 3.6), therefore a local optimum is actually a global optimum. Furthermore, an optimal solution to the problem corresponds to a vertex of the feasible region, thus always resting on a constraint or the intersection of two or more constraints, and not the interior of the convex region.

Solutions to LP problems can be obtained very fast and efficiently using LP-solver packages which range from simple spreadsheet add-ons to very powerful complete
packages that use their own algebraic modeling languages. The most prominent method for solving LP problems is the simplex method (Dantzig, 1963). An LP model (Lawler, 1976) applied to metabolic networks was developed, capable of finding in a single pass the minimal pathway distances (shortest path lengths) of all enzymes reachable from a source enzyme \((i^*)\), a step then repeated for each metabolic pathway and its constituent enzymes. Formulating and solving network problems (such as the shortest route or maximum flow problems) via linear programming is called network flow programming (Chinneck, 2001). A classic solution to the shortest path problem is that described by Bellman (1958), but the problem can be recast as an LP problem as described in Figure 3.7.

Figure 3.7 gives the background for use of LP in the calculation of shortest paths, but the details of the LP implementation used to calculate minimal pathway distances in metabolic networks are described below.

Considering the following notations:

Indices

\[ p = \text{pathways} \]
\[ i, j = \text{enzymes} \]

Sets

\[ E_p = \text{set of enzymes in pathway } p \]

Parameters

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

Positive continuous variables

\[ D_i = \text{distance from the } i^* \text{ source enzyme to enzyme } i \]

Overall, the algorithm can be outlined as follows:

For each metabolic pathway \((p)\) and for each source enzyme \((i^*)\) in that pathway, find the minimal pathway distances to all other enzymes by solving the following LP optimisation model:

\[
\text{maximise } \sum_{i \in E_p} D_i \tag{3.1}
\]

subject to

\[
D_j \leq D_i + 1 \ \forall (i, j) : L_{ij} = 1 \tag{3.2}
\]
\[
D_{i^*} = 0 \tag{3.3}
\]
\[
D_i \geq 0 \tag{3.4}
\]
The recognition of the shortest possible directed path from a specified source node to some other node of a weighted, directed graph is known as a shortest path problem. Instead of finding the shortest path from one specified source node to one specified destination node separately, it is more convenient to compute all shortest paths from a single source node to all other nodes in the network.

In this box, the appropriate expressions to derive the length of shortest paths \(D_j\) from the source node to nodes \(i\) are discussed. First, a directed graph with \(n\) nodes which is characterised by the following parameters is assumed.

\[a_{ij} = \text{the (finite) weight of edge } (i, j) \text{ (if there is such an edge; } + \infty \text{ otherwise)}\]

The source node is numbered 1; the aim here is to calculate the shortest path distances from node 1 to all other nodes in the network. If there are no negative-weight cycles (i.e. no cycles for which the sum of edge weights is less than zero) reachable from the source node, then \(D_1\) is equal to zero. Then, for each node \(j\) \((j > 1)\), there must be a final edge \((k, j)\) in a shortest path from 1 to \(j\). Thus, the shortest path lengths must satisfy the following equations, referred to as Bellman's equations (Bellman, 1958):

1. \(D_1 = 0\)
2. \(D_j = \min_{j \neq k} \{D_k + a_{kj}\} \forall j = 2, 3, ..., n\)

Bellman's equations solve the single-source shortest-paths problem in the general case, in which edge weights may be negative, given a weighted, directed graph with no negative-weight cycles. Equation 2 implies a system of \(n - 1\) inequalities and, for a fixed \(j\) and \(k \neq j\):

\[D_j \leq D_k + a_{kj}\]

Also, for fixed values of \(D_k\) and \(k \neq j\), the correct value of \(D_j\) can be determined by a simple LP model by maximising \(D_j\) subject to inequalities (3), thus satisfying equation 2:

\[\text{maximise } D_j\]
\[\text{subject to } D_j \leq D_k + a_{kj} \forall k \neq j\]

Next, the equivalence between Bellman's equations (1-2) and the LP model (4-5) is demonstrated through an illustrative example. In the case of metabolic networks, edge weights are always equal to one (i.e. \(a_{ij} = 1\)) in order to indicate a connection from enzyme \(i\) to enzyme \(j\). Consider the following example with four enzymes:

![Figure 3.7: Bellman's equations and LP in shortest path computation.](Image)

Enzyme A (shaded) is considered here as the source enzyme (i.e. \(D_A = 0\)). The minimal pathway distances from enzyme A to the other three enzymes (given next to the nodes) must be determined. According to Bellman's equation (2):

\[D_B = D_A + 1 = 0 + 1 = 1\]

By applying equation (5) for enzyme B:

\[D_B \leq D_A + 1 \Rightarrow D_B \leq 1\]

Then, by simply maximising \(D_B\) subject to (7), an optimal value of 1 is obtained, which is actually the same as in (6). By applying the same procedure for enzyme C, a pathway distance \(D_C = 2\) is obtained. Finally, by substituting the values of \(D_B\) and \(D_C\) from above, Bellman's equation (2) for enzyme E produces:

\[D_E = \min \{D_B + 1, D_C + 1\} = \min \{1 + 1, 2 + 1\} = 2\]

The same result can be obtained by first using inequality (5):

\[D_E \leq D_B + 1 \Rightarrow D_E \leq 2\]
\[D_E \leq D_C + 1 \Rightarrow D_E \leq 3\]

and then maximising \(D_E\) subject to (9) and (10).
Chapter 3. Methods

102

The above LP model can be solved using Bellman’s equations (Bellman, 1958), which defines a system of inequalities. Constraint 3.2 incorporates pathway information related to reaction connectivity, circularity and reaction directionality, facilitated by the use of parameter \( L_{ij} \) (for reversible reactions \( L_{ij} = L_{ji} = 1 \), however for irreversible reactions \( L_{ij} = 1 \) and \( L_{ji} = 0 \)). Constraint 3.3 assigns the initial value of zero to enzyme \( i^* \) to denote it as the source enzyme, while constraint 3.4 requires all \( D_i \) variables take positive values.

Finally, unbounded solutions can be avoided by adding:

\[
D_i \leq T \quad \forall i
\]  

where \( T \) is an appropriately large number. It should be noted that if \( D_i \) equals \( T \) at the final solution, it can be concluded that there is no path connecting the \( i^* \) source enzyme with enzyme \( i \) in the pathway under consideration.

The algorithm was implemented within the General Algebraic Modeling System (GAMS) software (Brooke et al., 1998), using the CPLEX 6.5 LP solver (a refinement of the basic simplex method of Dantzig (1963)) for solving LP problems such as the one in hand.

Unfortunately, by the time the collaboration with Simeonidis, Bogle and Papageorgiou came to fruition, the vast majority of the Chapter 5 analyses had already been performed. Nevertheless, section 5.3.4 describes analyses performed using LP derived pathway distance calculations. The LP method is very elegant and more work is currently being performed with it for whole metabolism network reconstruction and analyses (Lazaros Papageorgiou, personal communication).

At this point, the first of two major foundations of the work presented in the next two chapters has been described: the source of the \( E. \ coli \) SMM dataset and of related ancillary data — this can be thought of the “metabolic context” of the analyses. The second major foundation of the work is the “evolutionary context”: the identification of evolutionary relationships (i.e. homology) between proteins involved in the SMM datasets which is described below.

3.3 Structural assignments for genomes

If the complexity of the interactions within SMM networks is set aside for a moment, SMM can be thought of as being performed by the concerted action of a
number of proteins. In this "bag of proteins", certain enzymes will be homologous (i.e. share a common evolutionary ancestor). Identifying such homologues is one of the requirements for analysing pathway evolution. Pairwise comparison of protein sequences is the simplest way of detecting homology — proteins with detectable similarity are probably homologous, proteins with a high percentage of sequence identity having diverged only recently from the common ancestor. Below a certain level of similarity (around 30%), homology between proteins with a distant common ancestor may not be detected (Brenner et al., 1998). Two main strategies are used to detect such distantly related homologues: comparison of the three-dimensional structure of proteins which are often conserved even in the absence of detectable sequence similarity (Murzin, 1998) and profile-based sequence comparison methods (Park et al., 1998). Naturally, the former technique can only be used when the structure of both proteins to be compared is available. A further issue is the existence of multi-domain proteins composed of two or more evolutionary units capable of independent duplication and recombination.

The task is, therefore, to identify the domain make-up of SMM proteins and to define which of these units are evolutionarily related, grouping proteins with homologous domains into the same superfamily. Domains identified in proteins of known atomic structure are classified in databases such as the CATH (Pearl et al., 2001) and SCOP (Lo Conte et al., 2002) databases. By considering sequence, structure and functional similarities, these databases cluster domains thought to be evolutionarily related. Such structure classification databases are described in Section 1.3.4.

To identify the structural make-up of a set of proteins such as the SMM enzymes of E. coli, profile-based sequence comparison methods are used in conjunction with structural classification databases (Saqi & Sternberg, 2001; Teichmann et al., 2001a; Rison et al., 2002). The former are used to identify structural domains in proteins for which no structure is available and homology between domains is validated by reference to structural classification databases. Profile-based search methods are briefly discussed below.

### 3.3.1 Profile-based search methods

A profile can be thought of as a consensus primary sequence representation of a set of aligned sequences. Such a profile will highlight highly conserved positions in the alignment (such as key catalytic residues) as well as variable regions, such as loop regions, where insertions and deletions are common. The term “profile” was
coined by Gribskov et al. (1987) but Taylor (1986) and Barton & Sternberg (1990) describe similar concepts calling then “consensus templates” and “flexible sequence patterns” respectively. The profile scores each column in the alignment and the cost of substituting a residue at each of these positions, as well as the cost of an insertion or a deletion, can be summarised in a matrix. Such a matrix is the Position Specific Score Matrix (PSSM) used in the PSI-BLAST protocol described below (Altschul et al., 1990). The underlying model represented by a profile can also be elegantly described by a mathematical construct known as a hidden Markov model (HMM). Hidden Markov models are used by the SAM-T99 protocol, itself an update of the original SAM-T98 program (Karplus et al., 1998).

Naturally, query sequences can then be searched against profiles, be they PSSMs, HMMs or other. This can be repeated at a larger scale, scanning a query sequence against a library of profiles (see for example Schaffer et al., 1999). Profile-based search methods can detect evolutionary relationships that are undetectable by traditional pairwise sequence comparison methods (Park et al., 1998).

3.3.1.1 Position Specific Score Matrices and PSI-BLAST

PSI-BLAST is an iterative profile building and database searching package. PSI-BLAST starts from a single query sequence which is searched against a database using the BLAST algorithm (Altschul et al., 1990). Resulting significant matches are aligned to the original query sequence and the resulting multiple alignment is used to derive a profile. The profile is then used to search the database. Subsequent significant matches (i.e. matches scoring higher than a predetermined cut-off) are included in the profile to generate a new profile. This process is iterated until either no new matches are identified in the queried database (convergence) or a specified number of iterations is exceeded (see Figure 3.8).

3.3.1.2 Hidden Markov Models and SAM-T99

Hidden Markov models are a general form of probabilistic-model for sequences of symbols, and can be thought of as generative models that generate or emit sequences (Durbin et al., 1998). HMM are composed of states which “emit” symbols and transitions between states which have an associated probability. For example, each column in the multiple alignment can modelled by three states, match, insertion and deletion. The match state models the distribution of amino acids allowed in the column, the insertion state allows for the insertion of an amino acid after the
Figure 3.8: The PSI-BLAST procedure. PSI-BLAST begins with (a) a BLAST scan of the query sequence against an (often non-redundant) sequence database. (b) This search identifies matching sequences some of which (red and blue lines) match with a score higher than a predetermined cut-off. Others (black lines) match with a score below the cut-off (lines in the dashed boxes represents matches identified in subsequent iterations). (c) If necessary, significant matches may be omitted from the iteration (e.g. blue match). (d) The remaining sequences are aligned to the query protein (in the first iteration) and (e) a new PSSM is generated. After the first iteration, (f) PSSMs are scanned against the sequence database. As the PSSM is updated, new sequences (g) are “pulled-in” by the search (the sky blue line in one iteration, the green ones on the next). (h) The PSI-BLAST procedure ends when no new matches are identified between iterations (the search is said to have converged) or when a fixed number of iterations have been performed. The final alignment may be used to generate a PSSM which can be stored in a library of PSSM for subsequent scanning.
match state and the deletion state allows for the absence of any amino acid at this position (Eddy, 1998). Insert states can transit to themselves and therefore gaps of various lengths can be modelled. Match and delete states can only transit to the subsequent column or, in the case of a match state, to an insert state (see Figure 3.9). The probability of a particular sequence, including gaps, being emitted can thus be calculated. Conversely, if a profile-HMM has been generated, sequences can be aligned to it with dynamic programming algorithms (Durbin et al., 1998; Eddy, 1998). The strong probabilistic framework into which HMM can be assimilated (such as the use of Bayesian statistics) can be exploited to validate the results of HMM based searches (Durbin et al., 1998; Birney, 2000).

An HMM method for finding remote homologues of protein sequences was developed by Kevin Karplus and colleagues at the University of California, Santa Cruz. At the core of the method are profile-HMM building and HMM-scanning modules collected into the Sequence Alignment and Modeling (SAM) package (Karplus et al., 1998). Associated with the SAM package are wrapper scripts. Historically, these were developed to permit the use of SAM in the Critical Assessment of Structure Prediction (CASP) “competitions” (see http://predictioncenter.llnl.gov/) and so were called T98 and T99 for target98 and target99 respectively. Whilst the SAM modules are regularly updated (the current release at the time of writing this thesis is version 3.2), the wrapper scripts have either not been updated or not been renamed and the method is generally referred to as SAM-T99.

SAM-T99 begins with a single target sequence (known as a seed) and builds a
Chapter 3. Methods

HMM from the sequence and homologues found using the HMM for database search (Karplus et al., 1998). The process is iterative; first the seed is searched against a non-redundant sequence database such as NRDB. This search is traditionally performed using the WU-BLASTP implementation of BLAST and produces two sets of putative homologues: very close homologues \( (E < 0.00003) \) and possible homologues \( (E < 500) \). This initial stage has two advantages: the WU-BLASTP is an effective way of detecting homologues from a single sequence (more so than a single (seed) sequence HMM) and the NRDB is culled down to a size appropriate for a time-consuming profile-HMM based search (Karplus et al., 1998). The SAM-T99 then proceeds with four iterations of selection, training and alignment, i.e. finding significant matches to the HMM in the database, generating a new HMM from selected sequences, and finally aligning the training sequences back to the resulting HMM. For the first of these four iterations, the search HMM is derived from the seed (trivial) single sequence alignment and searched against the database of very close homologues identified by the WU-BLASTP. The threshold for selection of significant matches is strictly set to ensure only strong sequence matches are kept and the alignment produced at the end of the first iteration cycle is kept specific. On subsequent iterations, the input alignment is the output from the previous iterations and the search set is the larger set of possible homologues (Karplus et al., 1998). During these iterations, the selection threshold is progressively dropped to “pull-in” increasingly distant homologues of the seed. The final alignment produced can be converted into a profile-HMM suitable for inclusion into a library of models such as the SUPERFAMILY library of profile-HMMs for SCOP superfamilies (Gough et al., 2001; Gough & Chothia, 2002).

3.3.2 SCOP assignments

3.3.2.1 The SUPERFAMILY library

The SUPERFAMILY resource was used to assign putative structural domains to \( E. coli \) SMM proteins (Teichmann et al., 2001a; Gough & Chothia, 2002). SUPERFAMILY is derived from three resources: the SCOP structural classification of proteins database (Lo Conte et al., 2002), the ASTRAL database of domain sequences (Brenner et al., 2000), and the SAM-T99 iterative HMM building procedure (Karplus et al., 1998). The SCOP database identifies and classifies structural domains in proteins of known structure deposited in the PDB. The ASTRAL database uses these domain definitions to obtain the sequences representing such domains
from the PDB entries. The sequences may be used as SAM-T99 seeds. SUPERFAMILY is a library of HMM models built from some of these seeds (Gough et al., 2001). The seeds used in SUPERFAMILY are based on the sequences found for each SCOP superfamily in ASTRAL, filtered to remove any with greater than 95% homology. As a consequence, each SCOP superfamily is represented by one or more HMM models in SUPERFAMILY, depending on how many structures there are with less than 95% sequence identity in a given superfamily (Gough et al., 2001).

3.3.2.2 Assignment of SCOP superfamilies to genome sequences

A single or a collection of query sequences can be scanned against the SUPERFAMILY library. The data used in Chapter 4 were generated before the SUPERFAMILY resource was fully implemented. For these data, all HMM assignments were obtained and, for a particular protein, all assignments from different SCOP superfamilies overlapping by more than 30 residues were rejected. Now, the SUPERFAMILY library has been exploited to assign putative structures for all proteins of all fully sequenced organisms (Gough & Chothia, 2002, and http://www.supfam.org). For these assignments, a different domain clash resolution procedure is used: "for a full sequence, the regions are assigned one by one, beginning with the highest scoring and adding each subsequent non-conflicting lower score in turn. A conflict is defined as an overlap of 20% or more" (Gough et al., 2001).

3.3.3 CATH assignments

The Gene3D database was used to obtain CATH structural assignments for enzymes in the *E. coli* SMM (Buchan et al., 2002). Much of the terminology associated with the CATH database and discussed in this chapter is defined in the Introduction (see section 1.3.4.1). The Gene3D database is derived from data produced by the DomainFinder sequence search and analysis protocol used to populate the CATH-Protein Family Database (CATH-PFDB) (Pearl et al., 2002). Even though Gene3D and CATH-PFDB are derived from similar data using similar methodologies, they have different finalities. CATH-PFDB extends the CATH database by reliably integrating gene sequences from GenBank into CATH structural families (Pearl et al., 2000, 2002). The CATH database uses stringent structural and functional cut-offs for inclusion of a structural domain into a CATH superfamily; likewise, CATH-PFDB uses stringent cut-offs when associating a sequence domain with a CATH superfamily. Conversely, Gene3D is a database of pre-calculated structural assign-
ments to gene sequences and whole genomes aiming for maximal coverage, albeit at the cost of lower stringency (Buchan et al., 2002). Both CATH-PFDB and Gene3D use the same "raw ingredients": (i) representative sequences (Nreps) from every CATH-S95 family (i.e. for each cluster of sequences with 95% or more sequence identity within a CATH superfamily, the sequence derived from the structure with the best resolution) and (ii) a non-redundant library of genomic sequences. Both CATH-PFDB and Gene3D make use of the PSI-BLAST iterative database search algorithm (Altschul et al., 1997), and either generate data for, or exploit, the associated IMPALA profile scanning protocol. However, Gene3D exercises more leniency when "cleaning-up" putative structural assignments produced by the DomainFinder protocol.

CATH-PFDB and Gene3D, and the DomainFinder and DRange protocols used to generate them, are described below. The exploitation of these databases to generate the structural assignments used in the analyses performed in Chapter 5 is also described.

3.3.3.1 CATH-PFDB and DomainFinder

The CATH-PFDB population protocol is illustrated in Figure 3.10. CATH-S95 representative sequences (Nreps) are used as query sequences in a PSI-BLAST search of the GenBankCATHnr database (Figure 3.10a). GenBankCATHnr is a nonredundant sequence database comprising all the sequences in the GenBank nonredundant database (NRDB100, which only keeps one copy of any GenBank sequences with 100% identity (Benson et al., 2000)) and sequences derived from structural data deposited in the PDB (Westbrook et al., 2002) that have been classified in the CATH database (Pearl et al., 2002). There are two results from these searches: (i) the identification of GenBank hits to the query sequences (Figure 3.10b) and (ii) PSSM profiles (Figure 3.10c). The latter are automatically generated by the PSI-BLAST programme. As PSI-BLAST is an iterative search, the PSSM it generates changes at each iteration (unless the search converges). Only the PSSMs generated during the final PSI-BLAST iteration are stored in the CATH-IMPALA PSSM library (Figure 3.10d). For these PSI-BLAST runs, the initial matrix is BLOSUM62, the maximum number of iterations allowed is 20 and the E-value cut-off for inclusion in the next iteration is set to $5 \times 10^{-4}$ (Pearl et al., 2002).

It is theoretically possible to address PSI-BLAST drift during the iterations. As more distant relatives are pulled, the profile is modified and can probe more distant regions of sequence space. If the profile moves too far (e.g. by hitting a similar but
Figure 3.10: Populating the CATH Protein Family Database (CATH-PFDB). (a) CATH-S95 Nreps are searched against a non-redundant database using PSI-BLAST. The search identifies matching sequences (b) and (c) incidentally generates PSSMs suitable for inclusion in the CATH-IMPALA library (d). (f) matches are "inverted" to gather all Nrep hits to a given GenBank sequence. These matches are processed using the DomainFinder protocol (g). Suitable sequences are extracted (h), clustered (i) and assimilated into the CATH-PFDB (j). See text for full details.
Figure 3.11: The DomainFinder algorithm. The algorithm resolves multiple sequence matches on a single GenBank NRDB sequence (labelled GI) into distinct structural domains. The first and second GenBank NRDB sequence regions correspond with structural domains from CATH superfamily A and B respectively. The third region is matched by query sequences from different CATH superfamilies and requires further manual assessment. The “collapsed” assignments are assigned two sets of boundaries: consensus (most restrictive) and extreme (least restrictive) (adapted from Bray, 2001).

non-homologous sequence), the chances of spurious matches are increased (Pearl et al., 2002). One way to assess this drift is to compare the structure of query Nreps with the structure of any CATH-S95 “hit” in GenBankCATHnr (Figure 3.10e). If the query and target Nrep are in the same CATH family, the profile is not drifting. If a member of a different CATH superfamily is identified, the query and the cross-hit target can be structurally compared with the Sequential Structure Alignment Program (SSAP) algorithm (Taylor & Orengo, 1989). If the SSAP score is low, suggesting possible drift, the PSI-BLAST can be rerun with stricter E-value constraints. Alternatively, if the cross-hit has a high SSAP score, the two CATH superfamilies may be considered for possible merging.

The GenBank hits to CATH-S95 representative sequences are then “inverted” (Figure 3.10f): for each GenBank hit, a list of matching Nreps is derived. This list is known as a ‘hits.Hlevel’ file and is processed by DomainFinder (David Lee, computer programme and Pearl et al. (2002)). Sets of Nreps hits from the same superfamily are clustered to define the “consensus” and “extreme” ranges of putative domains whilst clashing assignments are rejected, as illustrated in Figure 3.11.

Putative domains and their consensus and extreme ranges are collected into “hits.Hlevel.ranges” files (Figure 3.10g). The extreme ranges are used to extract appropriate subsequences from the query GenBank sequences — one subsequence
per validated domain (Figure 3.10h). These subsequences are incorporated into the CATH superfamily associated with the clustered Nreps. In keeping with the CATH architecture, they are further clustered into sequence families with no less than 35% identity, 60% identity, 95% identity and 100% identity (i.e. the S35 (also known as S for Sequence), S60, S95 (also known as N for ‘Near identical’, hence Nreps) and S100 (also known as I for Identical) sequence families of CATH) using the Needleman and Wunsch based HOMOL algorithm (Orengo et al., 1997) and are finally integrated within the CATH-PFDB (Figure 3.10i and j). Domains overlapping by more than 10 residues are not incorporated into the CATH-PFDB (David Lee, personal communication).

3.3.3.2 CATH-PFDB and CATH-IMPALA

IMPALA (Integrating Matrix Profiles and Local Alignments) is designed for the procedure of comparing a single query sequence with a database of PSI-BLAST-generated PSSMs. The advantage of scanning against a library of PSSM is principally one of speed; scanning a small number of sequences against a relatively small profile database as opposed to scanning the same sequences against GenBankCATHnr using PSI-BLAST represents a considerable time economy (Schaffer et al., 1999; Pearl et al., 2002). One of the products of the CATH-PFDB population procedure is the generation of a PSSM for each CATH-S95 Nrep (Figure 3.10c). This collection of PSSM profiles constitutes the CATH-IMPALA library (Figure 3.10d). The CATH-IMPALA profiles provide slightly less coverage than equivalent PSI-BLAST searches at a set error rate (Pearl et al., 2002), because when searching the GenBankCATHnr database with PSI-BLAST hits identified at all iterations are considered, whereas the PSSMs stored in CATH-IMPALA are derived from final iteration hits alone. The difference in coverage was, however, calculated to be <1% (Pearl et al., 2002). Furthermore, IMPALA is calibrated to produce a performance similar to that of PSI-BLAST by careful selection of an appropriate E-value cut-off; Schaffer et al. (1999) suggest scaling the PSI-BLAST cut-off by a value proportional to the change in size between the original data set searched and the IMPALA library, but the value used by Pearl et al. (2002) was empirically derived by benchmarking and an IMPALA cut-off of 5 × 10^{-12} was found to give the same error rate as the PSI-BLAST cut-off of 5 × 10^{-4}. A new version of CATH-IMPALA is released every time the CATH-PFDB is updated and users can use IMPALA to run a query sequence against CATH-IMPALA online at http://www.biochem.ucl.ac.uk/bsm/cath_new/Impala/.
3.3.3.3 Gene3D and DRange

Gene3D is a database of pre-calculated structural assignments to gene sequences and whole genomes (Buchan et al., 2002). The Gene3D population process is illustrated in Figure 3.12.

The beginning of the Gene3D procedure is similar, if not identical, to the CATH-PFDB procedure: a library of protein sequences (e.g. an organism’s complete proteome) is searched against GenBankCATHnr or CATH-IMPALA using PSI-BLAST or IMPALA respectively (Figure 3.12a) with the aim of producing ‘hits.Hlevel’ files (Figure 3.12b).

In practice, neither of these two routes are used when populating Gene3D. Instead, to synchronise CATH-PFDB and Gene3D releases and to minimise computational time requirements, the ‘hits.Hlevel.ranges’ file produced whilst populating CATH-PFDB (Figure 3.10g) is also used by Gene3D (Figure 3.12c). Notwithstanding, Figure 3.10 is illustrated as is to show the principle behind Gene3D: the assignment of structural domains to protein sequences. In comparison, the principle behind CATH-PFDB can be thought of as the assignment of protein segments to structural families.

However, Gene3D was not fully implemented when structural assignments for the analyses presented in Chapter 5 were being generated. For these data, the hits.Hlevel file (Figure 3.12b) was indeed produced by searching the whole E. coli proteome against the CATH-IMPALA library using IMPALA (see Section 3.3.3.2).

The Gene3D population steps subsequent to the generation of the hits.Hlevel.ranges files are specific to this database. Assignments are cleaned-up using the DRange package (Figure 3.10d) and the resulting assignments stored in the Gene3D database (Figure 3.10e). The DRange package is composed of three modules: Collapse, MultiParse and CleanAssign (Buchan et al., 2002). These three modules are used to verify domain assignments. The first two of these modules were authored by Daniel Buchan and the CleanAssign module by Stuart Rison. The “clean-up” procedure is a triage procedure distinguishing between probably correct and probably incorrect assignments. The procedure is more flexible than that employed when populating CATH-PFDB and certainly much more lenient since the latter disregards any two domains overlapping by more than ten residues.

The Collapse module Domain assignments are first filtered using the Collapse module. The DomainFinder procedure used to convert the hits.Hlevel file to the
Figure 3.12: Populating the Gene3D Database. (a) Query sequences are scanned against either a sequence database (GenBankCATHnr) using PSI-BLAST or against a PSSM library (e.g. CATH-IMPALA) using IMPALA. Search results (b) are processed by DomainFinder to generate hits.Hlevel.ranges files (c). These are “cleaned-up” by the DRange package (d) and final assignments assimilated in the Gene3D database (e).
Chapter 3. Methods

Figure 3.13: DomainFinder’s cautious assignment of consensus regions. Several Nreps (indicated in black) have hit a region of a gene. The DomainFinder algorithm has attempted to merge these into a consensus region, but one of them is considered by DomainFinder to be too small to belong with the others and a second consensus, made from only one Nrep hit, is built. For the purposes of Gene3D, it is sufficient that the smaller domain is merged into the larger region (adapted from Buchan et al., 2002).

Figure 3.14: DRange: The Collapse module. Boxes, shown in white, represent Nreps regions assigned to the gene and the ‘New assignment’ boxes, in black, represent the possible outcome of collapsing the initial assignments. The Collapse module seeks to allow cases A and B without allowing case C (chaining) (adapted from Buchan et al., 2002).

hits.Hlevel.ranges file requires there to be a notable overlap between individual Nrep hits to a gene for these to be collapsed into one range (Figure 3.11). If a gene is only very distantly related to an Nrep, it is possible it will only match a small fragment of this Nrep. This means it will have limited overlap with other Nreps hitting the same region and might not be clustered into the same range even though all these Nreps belong to the same CATH superfamily. This scenario is illustrated in Figure 3.13.

Since the aim of Gene3D is to assign structural domains to proteins rather than rigorously extend the CATH database with sequence information, it is clear that the smaller of the two consensus ranges illustrated in Figure 3.13 could be subsumed into the larger domain. However, such a collapsing is not always valid, as shown in Figure 3.14.

To distinguish between valid collapses (Figure 3.14, cases A and B) and invalid collapses (Figure 3.14, case C), the Collapse module uses the following criteria: any
Figure 3.15: DRange: The MultiParse module. Genes are indicated as white boxes and domains as coloured boxes. The multi-domain protein is labelled with the two domains identified within it. Because the multi-domain match represents a global hit, it is assumed that the query gene has a similar pattern of domains; as a result, assignments to H families 1 and 2 are kept, whereas the assignment for H family 3 is lost (adapted from Buchan et al., 2002).

two regions to be merged must overlap by at least 60% of their lengths and extreme ranges must not extend beyond 20% of the length of the larger domain (Buchan et al., 2002). Furthermore, to avoid chaining (Figure 3.14, case C) — usually the consequence of a repeated sequence motif — the final merged region must not be more the 130% the size of the largest pre-merging domain.

The MultiParse module The MultiParse module is used when a query multi-domain protein has unresolved (clashing) structural assignments and a significant hit to a so-called CATH Class 5 sequence. CATH Class 5 sequences are multi-domain protein sequences derived from structures fully integrated into the CATH database (see Section 1.3.4.1). The structural domain make-up of the Class 5 sequences is thus empirically determined and this information can be exploited to resolve assignment clashes. A significant similarity of the query protein to the whole of the Class 5 sequence strongly suggests the query protein contains the same domains as the multi-domain protein and clashing assignments can therefore be discarded. This is illustrated in Figure 3.15.

The CleanAssign module The CleanAssign code performs the great majority of the structural assignment triage. The role of CleanAssign is to decide whether overlaps represent a cross-assignment (i.e. a gene region where two different CATH homologous superfamilies have been assigned) or an acceptable overlapping of domains from different superfamilies. The code was the result of several meetings between the author and Christine Orengo (CATH group leader), David Lee (Do-
mainFinder author), Frances Pearl (CATH database curator) and Daniel Buchan (Gene3D database creator and curator and DRange co-author).

At the core of the CleanAssign module are a simple overlap algorithm and a decision tree. Each query protein submitted has an associated set of one or more assignments as described in the hits.Hlevel.ranges file. If the protein only has a single assignment, there are no possible clashes and the assignments are entered into the Gene3D database. If the protein has multiple assignments described in the hits.Hlevel.ranges file, then for any pair of domains, three scenarios are possible:

1. the domains do not overlap;
2. only the extreme ranges of the domains overlap;
3. the consensus range of one of the domains overlaps with the consensus and/or extreme range of the other.

Domain pairs with no overlap are left unchanged. CleanAssign implements a decision tree to deal with scenarios 2 and 3 (Figure 3.16).

Each domain pair with overlap is flagged depending on its exit point out of the decision tree. It is therefore possible to vary the level of stringency of CleanAssign in two ways:

1. by rejecting all domains that go beyond a certain distance in the decision tree (e.g. keeping only pairs which do not have overlap between their consensus ranges, the equivalent of going no further than the first rhombus in Figure 3.16);
2. by relaxing the actual overlaps cut-offs described in the second to last and last steps of the decision tree.

Often, domains clash only at the H level of the CATH classification, that is, they have the same fold (CATH T level), but belong to different superfamilies. As such, cross-hits could be due to distantly related superfamilies having undetected or unconfirmed homology; such domains are not eliminated, but the assignment for the region of the gene covered by the domains is only made to the fold level. Such fold level assignments were not used in the dataset exploited in Chapter 5, but are included in Gene3D.
Chapter 3. Methods

Figure 3.16: DRange: The CleanAssign module. The module implements a decision tree to resolve clashes between assigned domains from different superfamilies. Clauses are increasingly lenient along the tree, but failure to accommodate even the final condition means clashing domains are rejected.
Chapter 3. Methods

119

![Diagram of structural assignment process]

**Figure 3.17**: The validation of *E. coli* K-12 MG1655 structural assignments using the DRange protocol. As each DRange module is applied, the number of validated assignments is culled down to 2448 assignments (adapted from Buchan et al., 2002).

### 3.3.3.4 CATH-PFDB and Gene3D

Gene3D is the repository for structural assignments verified using the DRange protocol and is available online at [http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/](http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/). This protocol is applied to all complete genomes released and, in May 2002, Gene3D included whole genome structural assignments for 66 genomes. For illustrative purposes, the number of structural assignments after each step of the validation procedure for the *E. coli* K-12 MG1655 is shown in Figure 3.17.

Figure 3.17 shows that Gene3D incorporated 2448 validated structural assignments. In contrast, only 1689 segments derived from *E. coli* sequences were incorporated into the equivalent release of CATH-PFDB (James Bray, personal communication). This a direct consequence of the great difference in stringency operated by the two protocols. Figure 3.18 illustrates these differences in stringency by comparing the outcome of DomainFinder and DRange on four example sets of domains.

For the purposes of large scale analyses, Gene3D therefore represents a more comprehensive resource with an extended coverage of the *E. coli* genome. This is why it forms the basis of the structural dataset used in Chapter 5.
Figure 3.18: Domain validation for CATH-PFDB and Gene3D. “Clean-up” criteria in CATH-PFDB are stricter than those for Gene3D. All hits/domains belonging to the same CATH superfamily are shown in the same colour. The first column shows unambiguous assignments and CATH-PFDB and Gene3D store the same assignment. In the second column, the problem described in Figure 3.13 occurs. CATH-PFDB rejects both assignments since it cannot confidently assign boundaries. Gene3D subsumes the smaller domain into the larger one using the Collapse module. In the third column, the extreme ranges of the two assigned domains overlap. CATH-PFDB rejects both domain if the extreme ranges overlap by more than 30 residues. The CleanAssign module clears both assignments into Gene3D. Finally, the last column shows two domains with substantial consensus overlap. CATH-PFDB rejects the assignments, Gene3D usually rejects them as well, although the MultiParse module may occasionally clear the domains.

3.3.4 Extending the homologous family repertoire with sequence families

The rationale for the identification of structural domains was to define the sets of homologous proteins within the SMM proteins of *E. coli*, but there are three reasons why structural assignments might be missed:

1. **Currently undetermined structures and/or non-inclusion of structural domains into structural databases:** Although it is generally thought that there are only a limited number of protein folds in nature (Chothia, 1992), new folds are still being discovered suggesting that more remain to be identified. By definition, an as yet unsolved new fold cannot be incorporated into structural databases such as CATH or SCOP. Such folds will therefore remain undetected.

2. **Domains undetected by the procedure:** There is always a trade-off between accuracy and coverage. To minimize the incidence of false positives, strict E-value cut-offs are applied by both structural assignment procedures. False negatives (i.e., the rejection of valid domains) may result from these cut-offs.
3. **Domains eliminated in clean-up procedure**: In addition to E-value cut-offs, the assignment procedures (i.e. the resolution of clashes between domains) may eliminate valid assignments.

Structural similarity can exist even in the absence of detectable sequence similarity, but detectable sequence similarity is a strong indicator of homology. Sequence similarity can therefore be used to identify homology by comparing regions of proteins or whole proteins with no structural assignments.

### 3.3.4.1 Automated determination of sequence domains

Automatic sequence domain determination methodologies have three steps: (i) perform an all against all comparison of sequences, (ii) cluster similar sequences, (iii) validate these clusters and if necessary sub-cluster them. The all against all comparison can be performed with any sequence similarity method, be it a pairwise method such as the FASTA algorithm (Pearson & Lipman, 1988), or a profile method such as PSI-BLAST (Altschul *et al.*, 1997). Problems can arise owing to the multi-domain nature of proteins. Clusters must therefore be carefully constructed and validated; often steps (ii) and (iii) are iterated (Park & Teichmann, 1998). Several automatic domain detection algorithms have been developed such as DOMAINER (Sonnhammer & Kahn, 1994), DIVCLUS (Park & Teichmann, 1998), MKDOM (Gouzy *et al.*, 1999) and SYSTERS (Krause *et al.*, 2000).

The DIVCLUS package (Park & Teichmann, 1998) was used to extend the homology repertoire of both the SCOP and the CATH based assignments with additional sequence families. To extend the SCOP assignments, unassigned proteins subsequence at least 75 residues long, and proteins wholly structurally unassigned, were collected into a sequence library (Teichmann *et al.*, 2001a). First, an all against all library search was performed using FASTA, and DIVCLUS sequence families were derived. Later, the unassigned sequences and subsequences were scanned against a library combining all *E. coli* sequences and the NRDB100 using PSI-BLAST. The generated FASTA and PSI-BLAST sequence families were combined where appropriate. Likewise, unassigned proteins subsequence at least 75 residues long and proteins wholly structurally unassigned were collected from the CATH assigned set. In this instance, only the PSI-BLAST against the combined *E. coli* and NRDB100 sequence library was performed (Rison *et al.*, 2002).

Although careful manipulation of DIVCLUS can maximise the quality of the output sequence families, the procedure is not perfect, especially when dealing with large
clusters (Park & Teichmann, 1998). The DIVCLUS algorithm can flag suspicious clusters; for both the SCOP and CATH related dataset, suspicious clusters were therefore manually verified and occasionally modified.

3.3.4.2 Assigning sequence families to structural families

The detection of remote homologues can sometimes be clarified by the use of intermediate sequence libraries (ISLs). ISLs expand structural databases by integrating genomic sequences into structural database. Examples of ISLs include HSSP (Holm & Sander, 1999), PDB-ISL (Teichmann et al., 2000) and the CATH-PFDB (Pearl et al., 2002). Such databases extend the structural databases' sequence space coverage and can be exploited as such: if sequence A is detectably homologous to sequence B and sequence B is associated with a particular structural family, then sequence A can be assigned to this structural family even in the absence of direct evidence of homology between sequence A and the structural family. Teichmann et al. (2000) demonstrated that pairwise searching of the PDB-ISL detected almost as many homologies (as validated with the SCOP database) as previous analyses using PSI-BLAST, but with considerable time economy since the PDB-ISL was searched only with the FASTA pairwise comparison algorithm.

Neither PDB-ISL nor CATH-PFDB were used herein, but the same principle was exploited. When wholly unassigned and unassigned regions were PSI-BLASTed against the combined NRDB and E. coli library and DIVCLUS clusters were generated, the final clusters sometimes contained E. coli proteins for which a structural assignment existed. If fragments, or whole sequences within a cluster, have homology to the matching region(s) of the structurally assigned E. coli protein(s), the structural assignment is transferred to the sequence family unless two or more different structural assignments clashed (see Figure 3.19).

3.3.5 Extending superfamilies into hyperfamilies

DomainFinder identified 16,000 significant regions in protein sequences that matched more than one homologous superfamily in CATH (cross-hits). This corresponded to 92 homologous superfamilies being linked via a match to a common sequence. Such cross-hits could be the consequence of invalid PSI-BLAST drift or motif matching (small proteins matching large structures containing repetitive secondary structures). On the other hand, cross-hits could suggest very distant homology between sequences classified in different CATH superfamilies. Many of the cross-hits where
homologous superfamilies belonged to the same fold family (i.e. had the same CAT number, but differed in the II-level of their CATH classification) arose in highly populated fold groups known as superfolds (Orengo et al., 1994). In such superfolds, function is highly variable (Todd et al., 1999, 2001), and therefore superfamilies remain distinct in the absence of convincing functional similarities. One such superfold is the TIM barrel (Orengo et al., 1994). Currently, there are 28 homologous superfamilies with a TIM barrel fold in the CATH database (release 2.4), yet it is known that many of these are functionally related (Copley & Bork, 2000). Whilst the CATH database—and indeed the SCOP database—conservatively segregate superfamilies, a more lenient approach is to collect superfamilies with strong evidence of homology into “hyperfamilies”. In practise, hyperfamily relationships are stored in the CATH database in a “neighbour table” (Pearl et al., 2002).

In the analyses presented in Chapter 5, structural assignments were made using the CATH database and proteins assigned to certain superfamilies were collected into two hyperfamilies. One hyperfamily collected certain superfamilies adopting the TIM barrel fold, the other hyperfamily collected superfamilies adopting the Rossmann fold (Rossmann et al., 1974). The former is described in Pearl et al. (2002), the second was defined following personal communications from Frances Pearl, UCL. Both hyperfamilies are described in Table 3.2.

### 3.4 Final datasets

In the preceding sections, the methodologies and databases used to generate the dataset used in Chapters 4 and 5 are described. This chapter ends with a summary
Chapter 3. Methods

Table 3.2: CATH hyperfamilies. Two hyperfamilies were created to combine distinct CATH superfamilies with strong evidence of evolutionary relationships (* this superfamily shares the common nucleotide binding domain with, and provides a link between, the Rossmann-fold NAD(P)-binding domain and the FAD/NAD(P) binding domain).

<table>
<thead>
<tr>
<th>CATH superfamily</th>
<th>Superfamily description</th>
<th>CATH hyperfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.40.50.300</td>
<td>Nucleotide binding Rossmann domain</td>
<td></td>
</tr>
<tr>
<td>3.40.50.720</td>
<td>NAD(P) binding Rossmann like domain</td>
<td></td>
</tr>
<tr>
<td>3.40.50.1130</td>
<td>Nucleotide binding domain*</td>
<td>77.10.10.1</td>
</tr>
<tr>
<td>3.50.50.40</td>
<td>FAD/NAD(P)-binding domain A</td>
<td></td>
</tr>
<tr>
<td>3.50.50.60</td>
<td>FAD/NAD(P)-binding domain B</td>
<td></td>
</tr>
<tr>
<td>3.20.20.20</td>
<td>Dihydropteroate (DHP) synthase</td>
<td>77.10.20.1</td>
</tr>
<tr>
<td>3.20.20.90</td>
<td>Triosephosphate isomerase (TIM)</td>
<td></td>
</tr>
<tr>
<td>3.20.20.120</td>
<td>Enolase superfamily</td>
<td></td>
</tr>
<tr>
<td>3.20.20.130</td>
<td>FMN-dependent oxidoreductase and phosphate (PP) binding enzymes</td>
<td></td>
</tr>
<tr>
<td>3.20.20.170</td>
<td>Aldolase class II</td>
<td></td>
</tr>
<tr>
<td>3.20.20.120</td>
<td>Quinolic acid phosphoribosyl (QAPR) transferase 1</td>
<td></td>
</tr>
</tbody>
</table>

of the generated dataset. For ease of reference, datasets generated for the analyses presented in Chapter 4 are prepended with ANAT (e.g. ANAT_PATHWAYS) and the datasets generated for the analyses presented in Chapter 5 are prepended with CONTEXT (e.g. CONTEXT_HOMOL).

3.4.1 Small Molecule Metabolism pathway datasets

Two SMM datasets were generated. In the first dataset (used in Chapter 4) pathways were not merged. For the second dataset, the pathways were merged as described in Section 3.2.3. These differences reflect two changes in philosophy: (i) in Chapter 4, pathways are thought of as distinct “bags of genes”, whereas in Chapter 5 pathways are considered as ordered sets of metabolic steps, and (ii) in Chapter 4, pathways are thought of as distinct biological entities, whereas in Chapter 5 pathways are merged to create as large networks as possible. The pathway dataset generated for these two chapters (respectively ANAT_PATHWAYS and CONTEXT_PATHWAYS) are described in Tables 3.3 and 3.4.

3.4.1.1 The ANAT_PATHWAYS dataset

The ANAT_PATHWAYS dataset is composed of 105 pathways. The pathways are listed in Table 3.3. The pathways were not stored in npl format; rather, a list of
Table 3.3: The ANAT_PATHWAYS dataset. Pathways are listed by their EcoCyc identifier. Further details on these pathways can be obtained from http://ecocyc.org:1555/ECOLI/new-image?type=PATHWAY&object=PATHWAYID-PWY replacing PATHWAYID with one of the listed identifiers.

all the genes in each pathway was obtained by “screen scraping” (Stein, 2002) the HTML source code of their related WWW page in EcoCyc. Npl files were not necessary since each pathway in this dataset was considered as a “bag of genes” and their relative connectivity ignored. The data were manually corrected following personal communications from Monica Riley and members of her research group.

The 105 pathways in ANAT_PATHWAYS have considerable overlap. For example, pathways GLYCOLYSIS/E-D and GLYCOLYSIS/TCA/GLYOX-BYPASS have the whole of glycolysis in common. Genes “duplicated” by virtue of appearing in a fragment of pathway common to two different pathways in ANAT_PATHWAYS are described as “virtual homologues” (Teichmann et al., 2001a). 581 distinct genes
compose these 105 pathways, but only 569 have known sequence (i.e. 12 of the genes have been identified only from their genetic or biochemical characteristics, but their activity has not been linked to an \textit{E. coli} gene or protein sequence).

### 3.4.1.2 The CONTEXT PATHWAYS dataset

In the CONTEXT PATHWAYS dataset, certain pathways were merged (see Section 3.2.3 and Rison \textit{et al.} (2002)). Initially, 102 EcoCyc pathways, composed of 738 reaction frames, were considered. The final dataset (Table 3.4) contains 82 pathways composed of 619 reaction frames. Of the original 102 EcoCyc pathways, 68 were left unchanged by the merging procedure, one was deleted (as it was found to be represented entirely in other pathways) and the remaining 33 were merged into 14 pathways accounting for the 82 (68 + 14) final pathways. The merging procedure was iterative. At each iteration an all against all comparison of pathways was performed and all overlapping pathways were flagged. For pathways overlapping more than one other pathway, the two largest pathways (in terms of numbers of reaction frames) were merged first. Iterations were repeated until no pathways overlapped by more than two contiguous reaction frames. Thus, some of the final merged pathways were finalised after a single iteration while others, such as the example given below, took five iterations.

The largest merged pathway was created by merging the EcoCyc pathways GLYCOLYSIS/TCA/GLYOX-BYPASS, GLYCOLYSIS/E-D, ANARESP1-PWY, FERMENTATION-PWY, GLUCONEO-PWY and GLYCOL-GLYOXDEG-PWY (respectively Glycolysis/Tricarboxylic Acid Cycle/Glyoxylate bypass, Glycolysis and Entner-Doudoroff, Anaerobic respiration, Fermentation, Gluconeogenesis and Glycol metabolism and degradation). The total number of frames in the six individual pathways was 81, the final number in the merged pathway was 42, illustrating the large overlap between individual pathways.

The CONTEXT PATHWAYS pathways were compared to the 89 SMM pathways identified in the metabolic pathway section of the KEGG database (Kanehisa & Goto, 2000). The 14 pathways created from the merger of two or more of the original EcoCyc pathways tended to be similar in size, and sometimes substantially larger, than their KEGG equivalent; the majority of the other pathways were smaller than their KEGG equivalent. However, the KEGG pathways are composite pathways, combining reactions occurring in a number of different organisms into one representation (Kanehisa & Goto, 2000). When only considering the portions of the KEGG pathways predicted by KEGG curators to occur in \textit{E. coli}, CON-
Table 3.4: The CONTEXT_PATHWAY dataset. Pathways 69 to 82 are the results of the merger between two or more EcoCyc pathways. Further details can be obtained from http://www.biochem.ucl.ac.uk/~rison/EcoliSMM/Pathways/index.html.
TEXT_PATHWAYS pathways appeared to be of a similar or larger size than their KEGG equivalent.

Whilst the merging procedure ensured that no two pathways in the CONTEXT_PATHWAYS dataset overlapped by more than two contiguous reaction frames, even if the pathways had been merged to completion (i.e. until no two distinct pathways in the final set shared a reaction frame), the product would not have been a single network representing all of \textit{E. coli}'s SMM. This is because certain pathways are only connected by a common metabolite frame, rather than a reaction frame, and only connectivity between reaction frames was considered. A measure of the remaining level of overlap is that 33 of the 581 distinct reaction frames in the 82 final pathways are found in more than one pathway. Nevertheless, the merges represent a transition from the traditional representation of SMM as distinct pathways towards a network representation (Jeong \textit{et al.}, 2000; Küffner \textit{et al.}, 2000).

Much of this thesis deals with the enzymes of the \textit{E. coli} SMM. It was therefore necessary to assign enzymes to each reaction frame. Of the 581 reaction frames considered, 59 had no known genes associated with them, the remaining 522 reaction frames accounted for 594 distinct genes. These genes encode for all the SMM enzymes considered in Chapter 5.

3.4.2 Structural and sequence assignments

\textit{E. coli} SMM proteins were assigned structural domains where possible. In addition, they were assigned sequence families where appropriate. This clusters SMM proteins into structure based superfamilies and sequence based families, both of which are indicators of homology. In other words, the process of assignment served to identify which \textit{E. coli} SMM enzymes were evolutionarily related. The relevant datasets are therefore identified as ANAT_HOMOLOGY and CONTEXT_HOMOLOGY. The ANAT_HOMOLOGY dataset is based on the SCOP structural classification database, the CONTEXT_HOMOLOGY dataset is based on the CATH structural classification database. Both datasets are summarised in Table 3.5.

The combined exploitation of all these datasets is presented in the next two chapters. The datasets describe metabolic relationships (the _PATHWAYS datasets), evolutionary relationships (the _HOMOL datasets), as well as additional contexts such as genomic relationships (as described by the chromosomal localisation of SMM genes) and functional relationships (as defined by EC number).
A. Methods, structural databases and domain “clean-up”.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Methods</th>
<th>Structural database</th>
<th>Domain “clean-up”</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANAT_HOMOL</td>
<td>SUPERFAMILY + PSI-BLAST and DIVCLUS</td>
<td>SCOP v1.53</td>
<td>&lt;30 residue overlap</td>
</tr>
<tr>
<td>CONTEXT_HOMOL</td>
<td>Gene3D + PSI-BLAST and DIVCLUS</td>
<td>CATH v1.7</td>
<td>DRange</td>
</tr>
</tbody>
</table>

B. Structural and sequence families, query sequences and assigned sequences.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Str. fams</th>
<th># Seq. fams</th>
<th># Query seqs</th>
<th># (%) Assigned query sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANAT_HOMOL</td>
<td>202</td>
<td>11</td>
<td>581</td>
<td>510 (88%)</td>
</tr>
<tr>
<td>CONTEXT_HOMOL</td>
<td>220</td>
<td>117</td>
<td>594</td>
<td>480 (81%)</td>
</tr>
</tbody>
</table>

Table 3.5: Homology datasets. A. Details of the methods and resources used. B. Details of the final assignments obtained (Str. fams, Structural families; Seq. fams, Sequence families; Query seqs, number of proteins for which assignments were sought; Assigned query sequences, number (percentage) of query sequences assigned to at least one structural or sequence family).
Chapter 4 mainly explores the evolution and structural anatomy of *E. coli* SMM. Chapter 5 focuses on the relationships between these contexts and the way they can shed light on the evolution and control of SMM.
Chapter 4

Analysing SMM — The Evolution and Structural Anatomy of Small Molecule Metabolism

The small molecule metabolism (SMM) of E. coli form an ideal dataset to answer general questions about how a large set of related proteins and pathways are structured and have evolved; the next two chapters of this thesis are dedicated to the analysis of E. coli’s SMM.

This chapter discusses: (i) the domain structure of the SMM 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 multi-domain proteins. Taken together, these descriptions form what can be called the structural anatomy of the SMM.

The elucidation of the structural anatomy of E. coli SMM also permits the analysis of:

1. the distribution of family members within and across pathways;
2. the types of features that can be conserved in protein families;
3. the nature of the homologues that are found within pathways;
4. the nature of the homologues that have been recruited across pathways;
5. the extent to which the families that form the SMM are unique to these pathways.

These observation offer insight into the evolution of E. coli SMM.
4.1 Pathways and the domain structure and homology of \textit{E. coli} SMM proteins

ANAT\_PATHWAYS and ANAT\_HOMOL, the two datasets used in this chapter, are described in detail in Chapter 3. The 105 pathways or superpathways considered varied in size from one to 37 genes; three-quarters of the pathways contained between two and ten genes. The SMM proteins were formed from the products of 581 genes. There were 12 the sequence of which is unknown at present: they have been identified only from their genetic or biochemical characteristics, but their activity has not been linked to an \textit{E. coli} gene or protein sequence. Thus, the number of different SMM proteins for which sequences were available was 569.

As described in Section 3.4.1.1, a number of the 105 pathways in ANAT\_PATHWAYS overlapped and a quarter of the SMM proteins were found to be active in more than one pathway. The description of the SMM in terms of separate pathways meant that the enzymes that occurred at nodes in the complex network appeared to be used repeatedly in different pathways. In the case of the SMM, 427 proteins were active in just one pathway; 96 proteins were active in two pathways; 32 in three pathways; 12 in four pathways; one in five; and one in six pathways. For the purposes of the analyses performed in this chapter, proteins active in more than one pathway were considered to perform as "virtual homologues" (see below).

The ANAT\_HOMOL dataset describes the domain make-up of \textit{E. coli} SMM proteins and, by extension, their evolutionary relationships (see Section 3.4.2). As described in the Methods section, hidden Markov models (HMMs) of SCOP domains were used to identify the evolutionary relationships of 695 domains in 487 SMM proteins. Four-fifths of the query proteins were completely, or nearly completely, covered by these assignments and one fifth were partially matched in that they also had an unassigned region of 75 or more residues. In addition, sequence matches were made between 27 domains from (i) the unmatched regions in four proteins partially matched by structural information and (ii) 23 other proteins. These were clustered into 11 sequence families.

Putting together the 487 \textit{E. coli} proteins whose domains were defined on the basis of structural information, and the 23 whose domains were defined by sequence comparisons, yielded information on the evolutionary relationships for 510 of the 581 different proteins that form the small molecule metabolic pathways in \textit{E. coli}, i.e. 88\% of the total number. In terms of pathways, 71\% of the 105 pathways
had a structural assignment for at least four-fifths of the enzymes, and 44% had a structural assignment for every single enzyme.

4.2 Domain coverage of *E. coli* SMM proteins

The matches made to SMM proteins by the HMMs and the sequence comparisons gave either the exact number of domains of which the proteins were composed, or allowed an estimate of this number.

As shown in Table 4.1A, of the 510 matched sequences, there were 271 where a single domain very largely covered the whole of the sequence, i.e. the sequence had fewer than 75 residues unmatched at the N- or C-termini. In most of these cases, the unmatched sections were much shorter than 75 residues. There were another 128 SMM proteins fully covered by two, three, four, five or six domains.

The remaining 111 matched proteins were partly covered by between one and four domains, i.e. they had an unmatched region of 75 residues or more, indicating the presence of one or more unmatched domains (Table 4.1A). A rough estimate of the number of domains in the unmatched sequences was made knowing that the average size of a domain in the SCOP database (Murzin et al., 1995) is 175 residues. It was assumed that unmatched regions of 75-260 residues corresponded to one domain; unmatched regions of 260-440 corresponded to two domains, etc. Using the same procedure, a rough estimate could also be made of the number of domains in the 59 SMM proteins of known sequence without any assigned domains. The results of these calculations (Table 4.1A) showed that overall, approximately half (54.8%) of SMM proteins contained one domain, nearly a third (29.0%) contained two domains and just under one-sixth (16.2%) contained three to six domains.

In the ANAT_HOMOL dataset, there were 695 structural domains distributed into 202 families, and 27 sequence domains distributed into 11 families, for a total of 722 domains. If the speculative results shown in Table 4.1A are taken into account, the total number of predicted domains in the dataset is 928.
### A. The number of domains in SMM proteins

<table>
<thead>
<tr>
<th>Number of domains (n)</th>
<th>Number of sequences completely matched by n domains</th>
<th>Number of sequences partly matched by n domains</th>
<th>Partially matched sequences: estimated number with n domains</th>
<th>Unmatched sequences: estimated number with n domains</th>
<th>Total: all proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>271</td>
<td>77</td>
<td>N/A</td>
<td>41</td>
<td>312</td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>26</td>
<td>55</td>
<td>14</td>
<td>165</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>5</td>
<td>36</td>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total no. proteins</strong></td>
<td><strong>399</strong></td>
<td><strong>111</strong></td>
<td><strong>111</strong></td>
<td><strong>59</strong></td>
<td><strong>569</strong></td>
</tr>
</tbody>
</table>

### B. Number and size of protein families

<table>
<thead>
<tr>
<th>Family size (n)</th>
<th>Number of families of size n</th>
<th>Family size (n)</th>
<th>Number of families of size n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>53</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4.1:** Domains and families in the ANAT_HOMOL dataset.
4.3 Protein families that form the *E. coli* SMM

4.3.1 Families of protein domains

As described in Chapter 3, the domains identified in the SMM proteins can be clustered into families on the basis of their evolutionary relationships. The 695 domains with structural information belonged to one of 202 different families; the 27 domains clustered by sequence comparisons gave another 11 different families. In total, the 722 domains identified in the 510 SMM proteins came from 213 different families, giving an average size of 3.4 members per family.

The sizes of individual families have an exponential character: there are few large families and many small families (Table 4.1B). The total membership of the largest 33 families, which have between 6 and 53 members, is slightly larger than the total membership of the other 180 families, which have between 1 and 5 members.

4.3.2 Domain combinations in *E. coli* SMM proteins

The domain data suggested that close to half of the SMM proteins were made of combinations of domains (see Table 4.1A). The patterns of domain combinations in these multi-domain proteins were investigated. For each family, the number of other families from which its members drew combination partners were tallied (Table 4.2A).

There were 57 families whose members always appeared in isolation, i.e. they only ever formed one-domain proteins.

Members of another 141 families formed combinations with other domains of known identity or with homologues of themselves, although some members also occurred in isolation. In other words, the majority of the members of these 141 families occurred in combination with one or more other domains in multi-domain proteins. There were also 15 families which occurred in proteins with unidentified additional domains (i.e. regions which had not been assigned sequence or structural domains).

Of the families that combined with other domains, the large majority combined with domains from only one or two other families, but a few families were more versatile (Table 4.2A). The family that made the largest number of different kinds of combinations, the Rossmann NAD-binding domain, combined with domains from 12 different families. The next 11 largest families combined with partners from
A. The number of different families from which partners are drawn.

<table>
<thead>
<tr>
<th>The number (n) of different families from which partners are drawn</th>
<th>Number of families whose members can be linked to n different families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Adjacent regions of unknown domain type</td>
<td>15</td>
</tr>
<tr>
<td>Total families</td>
<td>213</td>
</tr>
</tbody>
</table>

B. Protein families that have many partner families.

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

Table 4.2: Domain combinations in enzymes of the SMM of E. coli. A. A few of the families are versatile, but most only form combinations with one or two other families. B. The 12 most versatile families are listed and a description of the orientation in which domain pairs are found is given.
between 3 to 6 different families (Table 4.2B). All the domain pairs observed are listed in Table 4.3.

When domains from two families combined, they always did so in the same N-to-C orientation in 106 of the 110 possible different pairwise domain combinations. There were only four exceptions to the rule. Three exceptions involved three proteins that combined in an ABA or ABAB fashion, and hence the families A and B occurred next to each other in both ways within one sequence (see Table 4.3). In the fourth exception, the C-terminal domain inaconitate hydratase I was identified at the N-terminus ofaconitate hydratase B and vice-versa (as illustrated in Figure 4.5a).

### 4.3.3 Whole-proteins families

The previous two sections have been concerned with the families formed by domains and the combinations that the domains make with each other.

However, it is likely that many of SMM proteins in *E. coli* were produced by the simple duplication of whole proteins that have one or more domains. The 399 SMM proteins completely matched by structure or sequence domains (see Table 4.1A) were examined to estimate the extent to which they had arisen by the simple duplication of whole proteins.

A whole-protein family is therefore defined as a group of sequences with the same series of domains. The analysis identified 265 single-domain proteins that belonged to 59 families; 55 two-domain proteins in 17 families and 16 three-domain proteins in six families. The remaining 63 proteins not assigned to a whole-protein family were singlets: they did not have duplicates and therefore were not considered part of a family as such.

Thus, 336 proteins formed 82 whole protein families; these 336 accounted for 84% of the completely matched sequences and 254 (i.e. 336–82) of them were produced by simple gene duplications. This makes the assumption that whole-protein families were produced exclusively by duplication and ignores the possibility that the same set of domains were independently recruited to form homologous proteins. The independent recruitment scenario is however thought to be unlikely (Apic *et al.*, 2001; Bashton & Chothia, 2002).
Table 4.3: Domain pairs in *E. coli* SMM. 110 different domain pairings were observed. In four cases (indicated in bold), the domains occurred in both possible orientations; in all other cases, only one N-to-C orientation of domains was observed, accounting for the 114 pairs shown in this table. The number of times each domain pair was observed (No. times pair observed) is given (excluding virtual homologues) as is the number of pathways such a combination was observed in (No. pwys pair seen in; includes virtual homologues). Where three digit codes are given, the domain family belonged to a SCOP superfamily, in other cases, domain families were sequence families.
### 4.4 The distribution of family members within and across pathways

#### 4.4.1 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 4.1).

Description of the distribution of homologues was complicated by the quarter of SMM proteins (146 out of 581) active in more than one pathway; see above and Table 4.4. This meant that a family could participate 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 as having \( n - 1 \) "virtual homologues". The sequence matching calculations described above identified 201 domains in 131 of the 146 proteins that functioned in more than one pathway. By counting the number of times these domains are used in different pathways, it was observed that the 201 domains had 304 virtual homologues (see Table 4.4).

To take these virtual homologues into account when describing the distribution of domains within and between pathways, they were added to the total number of true domains to give a total of 1026 effective domains (i.e. 304 virtual homologues + 722 sequence and structural domains). These effective domains come from 213 families, therefore 813 (i.e. 1026 effective domains - 213 domain families) are either true homologues or virtual homologues of the other 213.

Examination of where the 813 homologues occurred showed that 506 were in different
Figure 4.1: Domain family size and number of pathways in the ANAT datasets. Family size is given in number of domains (re-counting proteins (virtual homologues) that occur in multiple pathways). 69 families are found only in one pathway, with 47 families of size one, 20 families of size two, one family of size three, and one family of size four. All the other 144 families are present on more than one pathway. (a) The data represented as a 2-D graph. Datapoints are colour-coded according to the number of times the combination of family size and number of pathways was observed: blue, once; purple: 2-6 times; red: 20+ times. (b) The same data represented in 3-D. The 2-D representation allows the easy identification of evenly distributed families (i.e. families where each member belongs to a different pathway), these lie along the x – y line. The 3-D representation shows that the majority of families are small and distributed in few pathways. The NAD(P)-binding Rossmanns family was omitted from the plots for clarity. The family has 71 members distributed between 41 pathways.
Chapter 4. Evolution and Structural Anatomy of SMM

pathways and the remaining 307 were in the same pathway as another member of the family or a virtual homologue. Thus, in the families that had more than one member, most members were involved in recruitment across pathways.

This can also be seen in the data shown in Figure 4.1, 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, were represented in more than one pathway and most of them (123) had between 50 and 100% of their members in different pathways (e.g. 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 were limited to one pathway. All these families were small: 67 had one or two members, and altogether they had 94 members. Of the homologous pairs of enzymes within pathways, over half were isozymes.

4.4.2 The distribution of the members of whole-protein families

The families likely to be formed by duplication of whole proteins are discussed above (see Section 4.3.3). The distribution of members of these families within and between pathways is described here. The proteins that are completely matched by multiple domains are either two or three-domain proteins. Among the 55 completely-matched two-domain proteins with homologues, there were thirteen homologues within pathways and 25 across pathways (assuming one protein from each of the seventeen two-domain families as the precursor). Out of the sixteen three-domain proteins that belonged to six three-domain families, there were four proteins that had 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.

4.5 Types of conservation within families

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. Three different types of conservation can be defined:
Chapter 4. Evolution and Structural Anatomy of SMM

1. conservation of chemistry, which occurs when homologues 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 Commission (EC) number of the reaction catalysed (Enzyme Nomenclature Committee, 1992, see also Table 2.6) 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 (Karp et al., 2002b). 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), the duplication is classified as conserving chemistry. This assumption was made because homologues were being examined, as opposed to considering only EC number when comparing possibly non-homologous proteins. There were exceptions to the connection between conservation of chemistry and EC number in homologous families, so that if EC numbers were not conserved, a more detailed inspection of the reactions and the proteins was required; sometimes chemistry is conserved while EC numbers are not (Todd et al., 2001). If substrates or cofactors were similar, the two enzymes were classified as conserving their main substrate-binding site, or a cofactor or minor substrate-binding site. If two enzymes belonged to the same family and catalysed the same reaction at the same point in the same pathway, they were considered to be isozymes.

4.6 Homologues within pathways

4.6.1 Pairs of homologues that form consecutive steps in pathways

A possible interpretation of the Horowitz “retrograde” theory of pathway evolution implies that duplications within pathways of enzymes in consecutive reactions are to be expected, since the substrates of such enzymes are usually similar (see Chapter 1 and Horowitz, 1945, 1965). In fact, of 445 consecutive pairs of enzymes that occurred in the SMM, only 26 (6%) contained domains from the same family. The range considered could be extended to triplets, since enzymes within a triplet are again
likely to have similar substrates. The SMM pathways contained 340 triplets of consecutive enzymes. Amongst these there were 37 (11%) cases where the first and third enzymes had contained homologous domain, as depicted in Figure 4.2a. Thus, homologous domains in proteins carrying out consecutive reactions were not only uncommon, they were less common than in enzymes separated by two metabolic steps.

Triplets of reactions in which all three enzymes have at least one domain from the same family were very rare: only two out of 340 triplets exhibiting this property were identified. In fact, the two triplets were actually one quadruplet of consecutive ligases in the peptidoglycan biosynthesis pathway. A triplet of $\beta\alpha_8$ barrels in three consecutive reactions in tryptophan biosynthesis was also identified. However, in this instance, the first two barrels are part of the same protein. These cases are discussed below.

These results suggest that gene duplications that conserve substrate binding properties have played a very minor role in the formation of consecutive steps in the SMM. Instead, recruitment usually takes 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, there were only six examples where the ligand-binding site of the main substrate was conserved and the catalytic mechanism changed (Table 4.5). Two of these examples occurred within the same pathway; conservation of the main substrate-binding site in consecutive enzymes therefore occurred in only five of the 106 pathways.

Two of the six were well-known examples: the $\beta\alpha_8$ barrels trpC and trpA in tryptophan biosynthesis (Wilmanns et al., 1991), and hisA and hisF in histidine biosynthesis (Lang et al., 2000). TrpC is a bifunctional enzyme consisting of two $\beta\alpha_8$ barrels, one of which is N-(5'phosphoribosyl)anthranilate isomerase and the other indole-3-glycerolphosphate synthase. TrpA is the $\alpha$-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 histidine and tryptophan biosynthesis examples are discussed extensively in the "Introduction" (see Section 1.1.5).

The four other cases listed in Table 4.5 are more complex than the two described above. In three of the four cases, the first EC number was conserved, so the reactions were not as different as those catalysed by trpC/trpA or hisA/hisF. Also, none of the genes were close to each other on the E. coli chromosome. In two of the four cases, the enzymes were not consecutive, but either "parallel" (see Figure 4.2b) or
Figure 4.2: Duplication within and across pathways. Superfamilies are indicated by different shapes (ovals and rectangles) with different members of the same superfamily distinguished by primes.
### Table 4.5: Conservation of the main substrate-binding site with change in reaction catalysed within a pathway.

These examples are the only detected cases of enzymes which 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.

<table>
<thead>
<tr>
<th>Superfamily and pathway</th>
<th>Genes</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphoenolpyruvate/pyruvate ((\beta\alpha)_8) barrels in fermentation</td>
<td>pykF/pykA, ppc</td>
<td>PykA/PykF and Ppc both have phosphoenolpyruvate as their substrate and belong to different EC classes</td>
</tr>
<tr>
<td>Ribulose-phosphate binding ((\beta\alpha)_8) barrels in tryptophan biosynthesis</td>
<td>trpA, trpC</td>
<td>Consecutive enzymes in different EC classes</td>
</tr>
<tr>
<td>Ribulose-phosphate binding ((\beta\alpha)_8) barrels in histidine, purine and pyrimidine biosynthesis</td>
<td>hisA, hisF</td>
<td>Consecutive enzymes in different EC classes</td>
</tr>
<tr>
<td>Phosphoribosyltransferases (PRTases) in histidine, purine and pyrimidine biosynthesis</td>
<td>prsA, purF and prsA, pyrE</td>
<td>Consecutive pairs of enzymes in EC class 2 (transferases)</td>
</tr>
<tr>
<td>dUTPase domains in deoxypyrimidine nucleotide/nucleoside metabolism</td>
<td>dcd, dut</td>
<td>Consecutive enzymes in EC class 3 (hydrolases)</td>
</tr>
<tr>
<td>Inosine monophosphate dehydrogenase ((\beta\alpha)_8) barrels in nucleotide metabolism</td>
<td>guaB, guaC</td>
<td>Enzymes one step apart in EC class 1 (oxidoreductases)</td>
</tr>
</tbody>
</table>
had a pathway distance of 2 (i.e. were separated by one intervening enzyme) (see Figure 4.2a).

In fermentation, the pyruvate kinase isozymes pykA and pykF act on phosphoenolpyruvate (PEP) in an EC class 2 reaction (transfer reaction). Phosphoenolpyruvate carboxylase (ppc) also uses PEP as a substrate in an EC class 4 reaction (lyase reaction). These “parallel” enzymes belong to the same family of $(\beta\alpha)_8$ barrels. There were three phosphoribosyltransferase enzymes in histidine, purine and pyrimidine biosynthesis that were related: amidophosphoribosyl transferase (purF) and orotate phosphoribosyltransferase (pyrE) both act on the substrate 5-phosphoribosyl-1-pyrophosphate (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 were two members of the inosine monophosphate dehydrogenase $(\beta\alpha)_8$ barrel family in nucleotide metabolism: IMP dehydrogenase (guaB) and GMP reductase (guaC), which are separated by the GMP synthase (guaA) catalysed reaction.

### 4.6.2 Homologous pairs of enzymes with the same substrate or product

Enzymes are said to form a “parallel pair” when two enzymes produce the same or similar products, so that they are both succeeded by the same enzyme (Figure 4.2b). Parallel pairs of enzymes share a homologous domain in nine out of 56 cases; in two of the cases, only one out of many domains is shared. An example of this type of scenario is given in Figure 4.3: fucA and rhaD in fucose and rhamnose catabolism formed a homologous parallel pair. The seven cases are described in Table 4.6A.

Conversely, there are also cases where pairs of different enzymes have the same or similar substrates but produce different products (Figure 4.2b). Again, these are instances of parallel pairs. In such cases, the two enzymes can either carry out related reactions or quite different reactions. Homology between enzymes was identified in eight out of 48 such cases, as described in Table 4.6B.

The assessment of substrate similarity used to identify parallel pairs was performed manually based on graphical substrate representations found in EcoCyc rather than based on a more formal assessment such as one derived from the comparison of Simplified Molecular Input Line Entry System (SMILES) strings encoding these substrates (Weininger, 1988).
A. Homologous pairs that produce the same product.

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

B. Homologous pairs that act on the same substrate.

<table>
<thead>
<tr>
<th>Preceding enzyme(s)</th>
<th>Homologous pair</th>
<th>Homologous domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same substrate, similar reactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aroC</td>
<td>trpDE, pabAB</td>
<td>trpE/pabA: anthranilate synthase, aminodeoxy-isochorismate synthase/lyase subunit, trpD/pabB: class I glutamine amidotransferase</td>
</tr>
<tr>
<td>Same main substrate, different cofactor/minor substrate, similar reactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glcB/aceB</td>
<td>maeA, maeB</td>
<td>Amino acid dehydrogenase domain and NAD(P)-binding Rossmann domain, Two aspartate/ornithine carbamoyltransferase domains</td>
</tr>
<tr>
<td>carAB</td>
<td>pyrLB, argF, argI</td>
<td></td>
</tr>
<tr>
<td>Same main substrate, different reactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eno</td>
<td>pykA and pykF, ppc, purF, pyrE</td>
<td>Phosphoenolpyruvate/pyruvate domain, Phosphoribosyltransferase domain</td>
</tr>
<tr>
<td>prsA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preceding enzyme with multiple substrate specificity: different substrates, similar reactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uxaC</td>
<td>uxaB, uxuB, gpt, apt</td>
<td>NAD(P)-binding Rossmann domain, P-loop nucleotide triphosphate hydrolase domain, Two actin-like ATPase domains</td>
</tr>
<tr>
<td>deoD</td>
<td>lyxK, rhaB</td>
<td></td>
</tr>
<tr>
<td>rhaA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.6:** Homologous pairs of enzymes that act on the same substrate or produce the same product.
Chapter 4. Evolution and Structural Anatomy of SMM

### Table 4.7: Pathways with high levels of duplication

<table>
<thead>
<tr>
<th>Pathway</th>
<th>p-val.</th>
<th>Dup. level</th>
<th>No. Ds in Pwy</th>
<th>No. Fs in Pwy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide metabolism</td>
<td>0</td>
<td>0.35</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Histidine, purine and pyrimidine biosynthesis</td>
<td>0</td>
<td>0.36</td>
<td>42</td>
<td>27</td>
</tr>
<tr>
<td>KDO, peptidoglycan and lipid-A precursor biosynthesis</td>
<td>0</td>
<td>0.37</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>Phosphatidic acid and phospholipid biosynthesis</td>
<td>0</td>
<td>0.40</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Aerobic respiration, electron donors reaction list</td>
<td>0</td>
<td>0.42</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>Anaerobic respiration, electron donors reaction list</td>
<td>0</td>
<td>0.56</td>
<td>27</td>
<td>12</td>
</tr>
<tr>
<td>Anaerobic respiration, electron acceptors reaction list</td>
<td>0</td>
<td>0.58</td>
<td>31</td>
<td>13</td>
</tr>
<tr>
<td>Gluconeogenesis</td>
<td>0.01</td>
<td>0.30</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>Colanic acid biosynthesis</td>
<td>0.01</td>
<td>0.38</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Glycogen catabolism</td>
<td>0.01</td>
<td>0.33</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Polyisoprenoid biosynthesis</td>
<td>0.01</td>
<td>0.50</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Overall, there was little bias towards homology between consecutive enzymes expected to have some similarity by virtue of binding similar substrates (see above). For parallel pairs, the fraction of homologous enzymes is somewhat higher: 13% and 17% respectively for the two scenarios described above.

### 4.6.3 Pathways with a high proportion of homologues

The existence of homologues within pathways is of special interest. Homologues might be expected because of the similarities of substrates in consecutive steps of pathways. This would support the retrograde model of pathway evolution (Horowitz, 1945).

A simple statistical test was used to determine if a particular pathways had a significantly high level of domain duplications. In this test, the domains in our set—amplified to include virtual homologues, but excluding isozymes—were distributed across the 105 pathways at random 10,000 times, and the observed duplication levels were compared to the random distribution. Thus an expectation value (p-value) was derived for the observed duplication level for each pathway.

11 out of 105 pathways had a high level of duplication unlikely to have occurred by chance (i.e. a p-value of 0.01 or less); these are listed in Table 4.7.

Three of the 11 sets of enzymes were labelled as pathways, but in fact they
Figure 4.3: Fucose and rhamnose catabolism. This pathway illustrates both serial recruitment of enzymes and a (boxed) example of "parallel" enzymes. 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.
were simply lists of similar single redox reactions for several electron donors or acceptors in aerobic and anaerobic respiration (pathways 5, 10 and 11 of the ANAT_PATHWAYS, Table 3.3). The occurrence of many iron-sulphur or nickel-iron centres in these three was therefore expected. In the other eight pathways, the reasons for the high proportion of homologues were varied and are described below.

4.6.3.1 Phosphatidic acid and phospholipid biosynthesis

PlsB and PlsC, both acyltransferases, are members of the same sequence family. Cts and PssA, transferases of a similar type, both belong to the SCOP “phospholipase D/nuclease” superfamily and are also significantly similar at the sequence level. Therefore, this pathway contained two families; in each family, sequences are related at the sequence level and the catalytic mechanism is conserved.

4.6.3.2 Colanic acid biosynthesis

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

4.6.3.3 Nucleotide metabolism

Most of the reactions in this pathway involve the transfer of phosphate groups; there are three members of the nucleotide triphosphate hydrolase superfamily, Adk, Gmk and PurA, as well as three members of the purine and uridine phosphorylase superfamily, Amn, XapA and 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, chemistry is conserved.
4.6.3.4 3-deoxy-D-manno-octulosonate, peptidoglycan and lipid A-precursor biosynthesis

The four consecutive ligases MurC, MurD, MurE and MurF all have catalytic domains the homology of which is apparent at the sequence level, as well as a glutamate ligase domain. These four consecutive enzymes have similarities in their catalytic mechanism as well as acting on similar substrates. There are three acetyl/acyltransferases, GlmU, LpxA and LpxD, which share a trimeric \(\beta\)-helix domain. GlmU also has a sequence domain in common with KdsB, the CPM-KDO synthetase, which probably harboured its N-acetylglucosamine-1-phosphate uridyltransferase activity. A further sequence family with conservation of chemistry encompassed LpxB, DktA and MurG, which are all transferases with similar 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.

4.6.3.5 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.

4.6.3.6 Glycogen catabolism

Two pairs of the seven enzymes in glycogen catabolism share similarity detectable at the sequence and structural level: MalS and MalZ. Both have a glycosyltransferase domain. GlgP and MalP, both phosphorylases, have a \(\beta\)-glucosyltransferase and glycogen phosphorylase domain in common, as shown in Figure 4.4.

4.6.3.7 Gluconeogenesis

Four proteins with homologous Rossmann domains, and therefore exhibiting conservation of co-factor binding, were identified in this pathway. In addition, there is a sequence family grouping two malic enzymes, one NADP\(^+\)-linked (MaeB) and one NAD-linked (SfcA).
Types of Duplication
- Conservation of chemistry with close conservation of the substrate-binding site
- Conservation of chemistry with less conservation of the substrate-binding site
- Internal duplication

Domain Families
- Glycosyltransferases
- Phosphoglucomutase, first 3 domains
- α-amylases, C-terminal β-sheet domain
- [glycosyltransferase & glycogen phosphorylase]

**Figure 4.4:** Glycogen catabolism. The glycogen catabolism pathway contains two duplications with conservation of catalytic mechanism. One is in the consecutive enzymes encoded by glgP and malP (with close conservation of substrate-binding as well as catalytic mechanism), the other duplication occurs for enzymes AmyA and MalZ (separated by MalP and with less conservation of the substrate-binding site). There are also internal duplications, where several members of the same domain family occur in one polypeptide sequence (MalS and pgm) and isozymes (MalS and AmyA).
4.6.3.8 Histidine, purine and pyrimidine biosynthesis

There were two examples of conservation of the substrate-binding site with change in catalytic mechanism: consecutive enzymes phosphoribosylpyrophosphate synthase (PrsA) and amidophosphoribosyl transferase (PurF), which are also related to pyrE, and consecutive enzymes hisA and hisF. In both cases, the relationship is detectable at the sequence level, and homologues carry out different types of reactions. There were several families with conservation of a minor substrate binding site: four proteins contain a glutathione synthetase ATP-binding domain and a biotin carboxylase N-terminal domains (CarB, PurK, PurT, PurD) and four proteins contain a class I glutamine amidotransferase domain (PurL, HisA, CarA, GuaA). In the latter case, GuaA and CarA are sequence related. In addition, the six-domain protein CarB shares a methylglyoxal-synthase-like domain with PurH.

4.6.4 Overview of highly-homologous pathways

Many examples of conservation of chemistry, which occur when there is a requirement for the same type of catalysis several times within a pathway, were observed in the eight pathways described in detail. Conservation of cofactor-binding domains was also common. However, homologues that conserved the main substrate-binding site and change the catalytic mechanism were rare.

There are two types of duplications within pathways that have not been discussed extensively: duplications within polypeptide chains and isozymes. These types of duplications were almost as common as conservation of cofactor or chemistry, but by definition cannot take place across pathways, and so were not been considered in detail. However, isozymes are considered in detail in Chapter 5.

4.7 Homologues in different pathways

4.7.1 Conservation of function across pathways

There were 115 families with members in more than one pathway. In 40 of these families, there was fairly close conservation of catalytic mechanism with the first two or more EC numbers being shared in all members of the family. In another 13 families, the cofactor-binding function was conserved.
Within the remaining 62 families, enzymes catalysed various reactions and bound various substrates, therefore conserving 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 (Todd et al., 2001). 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 (Neidhart et al., 1990; Petsko et al., 1993; Murzin, 1993; Babbitt & Gerlt, 1997; Todd et al., 1999, 2001); again, these modified homologues have different EC numbers.

4.7.2 The extent of serial recruitment of proteins across pathways

The observation that recruitment of individual domains and domain combinations across pathways is widespread was broadened by considering the extent to which consecutive proteins were recruited from one pathway to another. For instance, serial recruitment might be expected if a segment of a genome, such as an operon, were duplicated, and the duplicated enzymes were recruited to form a new pathway. By looking at whether there are homologous doublets or triplets of consecutive enzymes in different pathways, as illustrated in Figure 4.2c, evidence for such recruitment can be uncovered.

There were 445 different sets of two consecutive reactions involving enzymes of known sequence the ANAT_PATHWAYS set of pathways. In some cases, a reaction is catalysed by one or more polypeptide chains. Also, a pair of reactions might be carried out by different regions of one polypeptide chain.

Pairs of consecutive enzymes were considered to be homologous if both the first and second reaction catalysing enzymes-pairs had at least one homologous domain in the enzyme(s) catalysing that reaction. This notion is difficult to explain textually, but simple to represent graphically as is done in Figure 4.2c. The unprimed oval and rectangle are both in one pathway, the primed oval and rectangle in an other. In each pathway the oval and rectangle enzyme are consecutive. The reaction catalysed by the primed and unprimed enzymes are usually similar but, because they are in different pathways, not identical.

Since certain enzymes were used in more than one pathway, the proteins involved had to be non-identical. Excluding pairs sharing only one of the common nucleotide-
binding "battery domains" (Rison & Thornton, 2002, and Chapter 6) such as Rossmann domains, P-loop nucleotide triphosphate hydrolases and PLP-dependent transferases, and also cases where only one domain was shared between multi-domain proteins with different domain architectures, only a small number of genuine candidates for serial recruitment remained. These genuine cases of homologous pairs of consecutive enzymes are described below, in Figures 4.3 and 4.5, and in Table 4.8.

The serial recruitment of enzymes from one pathway to another would be supported if, in a homologous pair of consecutive enzymes, the genes encoding the enzymes involved in the first pathway were close to each other on the chromosome, and the genes involved in the second pathway were also close to each other on the chromosomes. Such an observation would constitute additional evidence for recruitment by duplication of a segment of the chromosome.

Only two such cases were identified; one of these is shown in Figure 4.3: the rhaB and rhaD genes were found to be separated by one gene on the *E. coli* chromosome, and the fucK and fucA genes found to be separated by two genes (Table 4.8). These genes were also homologous to araB and araD in arabinose catabolism and located in the same way, although for clarity the latter were not included in Figure 4.3. Genes araB and araD are separated by one gene on the chromosome (see Table 4.8).

The second case involved the two pairs of enzymes in Figure 4.5b: the enzymes in carnitine metabolism and menaquinone biosynthesis were found to be 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 (see Table 4.8).

In the example given in Figure 4.5a, one of the enzyme pairs is adjacent while the other is not. The enzymes in Figure 4.5a are in leucine biosynthesis and the TCA cycle. Isopropylmalate isomerase 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 biosynthesis, 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, even the isozymes acnA and acnB (see Table 4.8), are scattered around. This example of pathway duplication is mentioned by Huynen & Snel (2000), 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 were not in the ANAT_PATHWAYS dataset, as there is no experimental evidence for their activity in the potential methyl-citrate cycle in *E.*
Chapter 4. Evolution and Structural Anatomy of SMM

(a) Leucine biosynthesis and the TCA cycle

(b) Carnitine metabolism and menaquinone biosynthesis

**Figure 4.5:** Homologous pairs of consecutive enzymes. These two examples are cases in which the two first and two second reactions in each distinct pathway are catalysed by enzyme(s) with homologous domains. In leucine biosynthesis, LeuC and LeuD are subunits of one enzyme, while AcnA and AcnB in the TCA cycle are isozymes.
Table 4.8: Homologous pairs of consecutive enzymes and serial recruitment. The certain features of the genes discussed in Section 4.7.2 are summarised. Pos. refers to the position of the genes on the *E. coli* chromosome, regardless of their orientation (Ori.) Thus caiD and caiC, with positions 36 and 37 respectively, are side by side on the *E. coli* chromosome. The pathway position (P. pos.) is also listed, i.e. whether the gene(s) catalyse the first of the two consecutive reactions (flagged I) or the second (flagged II). Homologous genes are colour-coded accordingly and the SCOP superfamily(ies) they belong to listed in the SF column. Homologous pairs of consecutive enzymes are grouped in reaction sets. Set 1 is illustrated in Figure 4.5b, set 2 (with the exception of araB and araD) is illustrated in Figure 4.3, and Set 3 is illustrated in Figure 4.5a.
Potential duplications of consecutive enzymes were identified within nucleotide metabolism and ubiquinone biosynthesis. In these, neither pair of enzymes was close to each other on the *E. coli* chromosome. A similar instance was identified with the multifunctional multi-domain enzyme AdhE involved in fermentation, which contain 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 the ones described above suggests that in general, recruitment of domains, whether within or across pathways, is not ordered with respect to a chain of consecutive reactions. In general, individual proteins or domains and not sets of consecutive enzymes are recruited to pathways. Since so few pairs of homologous consecutive enzymes as a fraction of the possible pairs were identified, it is not surprising that no homologous triplets were found. There seems to be little order in the recruitment of domains in the construction of metabolic pathways. 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. These notions are explored in detail in Chapter 5.

### 4.8 *E. coli* SMM pathways and horizontal gene transfer

Horizontal gene transfer is defined as the transfer of genetic information from one genome to another, specifically between two species (Li & Graur, 1991). This transfer of genetic material may include the transfer of one or more functional genes from a different species. This has implications with regards to the work presented herein. Should many of the genes involved in the SMM of *E. coli* originate from a different species, the detected patterns of pathway evolution would be "muddied". The extent of horizontal gene transfer in the development of *E. coli* metabolism is therefore considered in this section.

By testing whether the GC content in the first and third codon positions were atypical when compared to the entire genome, Lawrence & Ochman (1998) identified *E. coli* genes potentially horizontally transferred within the last $10^8$ years. They found 755 candidates in *E. coli*, the product of at least 234 lateral transfer events. The age of the acquired genes was estimated by pooling acquired genes on the basis
of their GC content at each position, and performing reverse amelioration analysis on each pool, i.e. estimating the number of G/C to A/T mutations required for each pool to reach base compositions at all codon positions matching those typically observed in a bacterial species (Lawrence & Ochman, 1997).

Only sixteen of the putative horizontally transferred genes were part of the ANAT dataset, indicating that this set of enzymes has probably not been affected by recent horizontal transfer on a large scale. All but one of these sixteen proteins had a structural or sequence assignment, and all of the proteins with an assignment were members of families that contained domains from other SMM proteins not themselves candidates for recent horizontal transfer.

The exception was glyoxalase II (gloB), which has the only metallo-hydrolase domain in \textit{E. coli} small molecule metabolism. There were two cases where a sequence of multiple genes may have been recently horizontally transferred: thiogalactoside acetyltransferase (lacA) and \(\beta\)-galactosidase (lacZ), as well as five genes (gfl, rfbC, rfbA, rfbD and rfbB) in O-antigen biosynthesis (i.e. the synthesis of a repeat unit composed of four sugars which are attached to lipids in the outer membrane).

Based on the degree of amelioration, whilst the oldest horizontally transferred genes in the \textit{E. coli} genome were acquired nearly 100 million years ago, most of the transferred DNA has a relatively recent origin in the \textit{E. coli} chromosome, and the average age of horizontally transferred genes is 6.7 million years ago Lawrence & Ochman (1998). It therefore seems probable that much of the SMM of \textit{E. coli} was established before acquisition of horizontally transferred genes.

## 4.9 Homologues of SMM proteins in other functional categories

The 695 domains in the SMM proteins that are homologous to proteins of known structure belong to 202 families. The HMMs for these families were matched to all \textit{E. coli} proteins, and 134 of these families were found to have additional members outside small molecule metabolism. In all, the 134 families had 1517 members in \textit{E. coli}. Of these, 577 members were in SMM proteins and 1039 were in proteins outside the SMM. This means that most of the constituents of SMM proteins, nearly 85%, belong to families whose members have been recruited not only within and between the SMM, but also from and to many other physiological roles or functional classes that have been described for \textit{E. coli} (Riley, 1998a).
The 68 families whose members were only found in the SMM were all small and contained 118 members all together.

4.10 Discussion

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 fifty years (Bridges, 1935; Lewis, 1951; Zuckerkandl, 1975; Jacob, 1977; Petsko et al., 1993; Todd et al., 2001). The idea of homologues forming pathways followed Horowitz' argument for retrograde evolution of pathways, and an example was discussed by Wilmanns et al. (1991) in the tryptophan biosynthesis pathway. Yčas (1974) and Jensen (1976) suggested recruitment across pathways as a mechanism of pathway evolution. The importance of such recruitment for the formation of pathways that have evolved recently has been described for the mandelate pathway by Petsko et al. (1993) and for a pathway that degrades a xenobiotic pesticide by Copley (2000). On the basis of the distribution of 38 homologous \((\beta\alpha)_8\) barrel structures in central metabolism, Copley & Bork (2000) have also argued that recruitment plays a significant role in the formation of metabolic pathways.

This chapter gives a detailed description of the extent and roles of these different mechanisms in the formation of 510 of the 581 proteins that form the ANAT_PATHWAYS dataset. 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 SMM proteins were determined. The domains belonged to one of 213 different families which had between 1 and 51 members, and on average, 3.4 members. Domains in almost 70% of the families underwent 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, were widely distributed across different pathways. The presence of homologues within pathways was less common: of the 106 pathways, only eleven had a significant number of homologous domains. Even in these cases, it was common for homologous enzymes to conserve catalytic or cofactor binding properties and it was very rare for them to be close in a pathway, to conserve substrate recognition and to change their catalytic mechanism. Similarly, recruitment of family members across path-
ways involved conservation of catalytic mechanism and cofactor-binding domains much more than conservation of substrate recognition with a change in chemistry. This suggests that it is more difficult to evolve a new catalytic mechanism than a new substrate-binding site (Petsko et al., 1993; Babbitt & Gerlt, 1997). There was very little order in this process of recruitment of enzymes; few examples of serial recruitment of consecutive enzymes from one pathway to another were observed.

There were 134 families whose members formed over nearly 85% of the SMM proteins and which also had members outside small molecule metabolism. 68 families, whose members formed just over 15% of the SMM proteins, only occurred in these pathways.

Though during evolution central metabolism has been modified by losses, substitutions and innovations (see Dandekar et al., 1999; Huynen et al., 1999, 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 (LCA) (Lazcano & Miller, 1999).

Overall, the results reported in this chapter suggest that, even in the LCA, the functional domains that form the repertoire of proteins in an organism must have had an extensive “mosaic” character (Teichmann et al., 2001b). Most proteins were formed by families whose members have a function that can be used repeatedly, or can be easily modified for related uses. Only a minority of proteins were formed by small families whose members have a functional role that is required in only one or a few instances, and which cannot be easily modified to perform other roles.

The next chapter continues to explore these concepts but, rather than considering SMM as a collection of individual pathways, it looks at SMM as a network. In this network, the distance separating two enzymes is used as a metric, and correlated to a number of other contexts to gain a quantitative understanding of recruitment in SMM.
Chapter 5

Analysing SMM — Pathway Distance, Genome Distance, Homology and Function

In the previous chapter, the structural and evolutionary anatomy of SMM pathways in *E. coli* was described. This investigation gave a comprehensive picture of the pattern of protein domain organisation, both within *E. coli* metabolic genes, and within and between different metabolic pathways.

In that chapter, each pathway in the EcoCyc database was considered as a separate entity. Comparing the distribution of domain family members within and across pathways, it was observed that recruitment of domains across pathways is more common than recruitment within pathways. When considering domain families with more than one member, the majority of families had over twice as many members distributed across as within pathways. Furthermore, pairs of consecutive enzymes exhibiting conservation of substrate binding with a change in catalytic mechanism, a pattern consistent with retrograde evolution (Horowitz, 1945, 1965), were rarely observed. Rather, the patterns provided support for a non-local recruitment patchwork model of pathway evolution. Similar observations were made by Tsoka & Ouzounis (2001).

The analysis of aspects of the genome other than the predicted amino acid sequence of the proteins encoded by the genes has been described as the “context of a gene” (Huynen & Snel, 2000). In this chapter, three contexts are considered: the genome (i.e. the relative location of SMM enzyme genes on the *E. coli* chromosome), metabolism (i.e. the relative location of enzymes within the SMM network) and the evolutionary context.
Chapter 5. Pathway Distance, Genome Distance, Homology and Function

Much work has already been done regarding the spatial organisation of genes in bacteria. Tamames et al. (1997), considering *Haemophilus influenzae* and *E. coli*, observed that functionally related genes (as classified within simplified scheme derived from GenProtEC's gene classification scheme (Riley, 1998a)) were neighbours more often than functionally unrelated genes. A strong correlation between genomic clustering and function was also detected when considering a large number of genomes (Overbeek et al., 1999, 1998), in particular for genes in close proximity not just in one, but in many genomes. Recently, the concepts presented by Overbeek et al. (1998, 1999) were generalised and implemented within a function prediction algorithm which connects genes likely to share functional similarity (in particular involvement in common metabolic pathways) by analysing orthology and genomic localisation of genes (Kolesov et al., 2001). Such correlations are, however, strongly dependent on phylogenetic distance (Huynen & Snel, 2000; Tamames, 2001; Dandekar et al., 1998).

In this chapter, the work presented in the previous chapter is extended; patterns of domain distribution and recruitment within the *E. coli* SMM are explored further. A trinity of contexts (evolutionary relationships of genes, genomic location of genes and metabolic environment of enzymes) is explored, rather than considering the SMM as a “bag of genes” (Huynen & Bork, 1998). Furthermore, as much as possible, SMM is analysed as a single network rather than a collection of arbitrarily defined pathways. In particular, the use of the pathway distance metric described in Section 3.2.6 allows a “stepwise” analysis which detects hitherto unobserved patterns of recruitment as well as clarifies the metabolic range of SMM gene clustering. A large-scale analysis of isozymes (homologous enzymes participating in the same metabolic step) is also performed, and the inline reuse of enzymes (i.e. the reuse of the same enzyme at different locations in the SMM) is investigated.

From these data, certain properties of *E. coli* SMM are identified and their possible implications for the evolution of the SMM network and its regulation discussed.

5.1 Small Molecule Metabolism pathways

The CONTEXT_PATHWAYS dataset used in this chapter is described in details in Chapter 3. As discussed previously, many of the pathways separately described in EcoCyc possess a high level of overlap, i.e. stretches of the same reaction frames are found in both. In Chapter 4, genes found in reaction frames reused in different EcoCyc pathways were identified as “virtual homologues”. These “virtual homologues”
reflect the "inter-connectedness" of the pathways — which can more realistically be considered as a network. Here, such duplication were dealt with by merging pathways with three or more reaction frames in common (see Section 3.4.1.2).

Although the dataset is described extensively in Chapter 3, many of the properties of the dataset are summarised in Table 5.1 for ease of reference. Further information may also be obtained from http://www.biochem.ucl.ac.uk/~rison/EcoliSMM/index.html.

### 5.2 Structural annotation, sequence families and evolutionary relatedness

To investigate the relationship between pathway distance and evolutionary relatedness of SMM enzymes, it was necessary to describe the enzymes in terms of their structural domain composition. Information on the domain structure and evolutionary relationships of the proteins of known atomic structure is available from the CATH database (Orengo et al., 1997; Pearl et al., 2001). In CATH, structural domains in the Protein Data Bank (PDB) (Berman et al., 2000) are classified in a 4-level hierarchical scheme. Domains predicted to share a common ancestor on the basis of sequence, structure and functional similarities are assigned the same CATH number and belong to the same superfamily (these superfamilies are further subdivided into families on the basis of sequence identity, e.g. S95 sequence families contain members that are 95% or more sequence identical). Two proteins containing a domain classified in the same CATH superfamily can be considered to be evolutionarily related — at least with respect to that domain.

In this chapter, the CONTEXT_HOMOL dataset was used. This dataset describes the domain make-up and the homologies between the 594 genes in the CONTEXT_PATHWAYS dataset. The CONTEXT_HOMOL dataset was derived from the Gene3D database and by using sequence comparison methods to analyse whole genes and gene regions of greater than 75 residues for which no structural assignments could be made. Again, properties of the dataset are summarised in Table 5.1.

### 5.3 Context based analyses

The subsequent sections in this chapter describe the relationships between a number of SMM related contexts such as the metabolic, genomic, evolutionary and functional
Pathways:
- Number of original EcoCyc pathways considered: 102
- Number of final pathways: 82
- Number of 82 final pathways composed of 2 or more of the original 102 EcoCyc pathways: 14

Reaction frames:
- Total number of frames in 82 analysed pathways: 619
- Number of distinct frames: 581
- Number of these 581 frames used more than once: 33 (5.68%)

Gene assignments:
- Number of the 581 reaction frames with no known genes: 59 (10.15%)
- Total number of genes in 619 analysed reaction frames: 776
- Number of distinct genes: 594
- Number of these 594 distinct genes used more than once in the pathways: 79 (13.30%)

GenBank identifiers:
- Number of the 594 distinct genes assigned a GenBank PID: 586 (98.75%)

Structural (CATH) and sequence domain families:
- Number of CATH domain families: 220
- Number of sequence families pre-linkage to structural families: 137
- Number of sequence families post-linkage to structural families: 117
- Total number of families: 337

Domain family assignments:
- Number of the 586 genes with known PID assigned to one or more CATH domain families: 382 (65.19%)
- Number of the 586 genes with known PID assigned to one or more "sequence" domain families only: 98 (16.72%)
- Number of the 586 genes with known PID assigned to one or more CATH and/or "sequence" domain families: 480 (81.91%)

Genomic locations:
- Number of the 594 distinct genes with an identifiable chromosomal location: 584 (98.32%)

**Table 5.1:** An overview of data analysed in Chapter 5. These data are based on the CONTEXT_PATHWAYS and CONTEXT_HOMOL datasets described in Chapter 3.
5.3.1 Homology and pathway distance

Homologous pairs (i.e. enzyme pairs with at least one CATH or sequence domain in common) were tallied for each investigated pathway distance. The percentage of all positive pairs was then plotted for each pathway distance (Figure 5.1). Overall, 92 recruitment events (homologies in enzymes with 1-11 metabolic steps) were observed.

In order to ascertain the significance of these data, the probability (p-value) of observing these percentages by chance was calculated. An all-versus-all comparison of SMM enzymes with at least one structural or sequence assignment was performed, and all pairs sharing at least one sequence or structural domain were flagged. For each distance, a number of pairs equal to that considered (e.g. there are valid observed 660 pairs at pathway distance 2) were picked with no replacement from the all-versus-all collection. For each set of pairs picked, the percentage of positive pairs (i.e. having at least one domain in common) was calculated. The picking process

Figure 5.1: Homology and Pathway distance. At each pathway distance (x-axis), the percentage of enzyme pairs at that distance sharing homology in at least one domain (histogram, primary y-axis) is plotted. Observed percentages found by simulation to be statistically significant are in bold type. Only pairs where a structural and/or sequence assignment has been made to both proteins are considered. The number of such pairs is shown (line plot, secondary y-axis). The dashed line indicates the average percentage of homologous pairs expected if SMM enzymes were randomly distributed (≈1.7%).

contexts. Such analyses use a number of metrics; these metrics, in particular the pathway distance metric, are discussed in Chapter 3.
Chapter 5. Pathway Distance, Genome Distance, Homology and Function

<table>
<thead>
<tr>
<th>Pathway Distance</th>
<th>Number of pairs</th>
<th>Number of homologous pairs</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>660</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>491</td>
<td>19</td>
<td>$5.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>3</td>
<td>425</td>
<td>11</td>
<td>$6.2 \times 10^{-2}$</td>
</tr>
<tr>
<td>4</td>
<td>367</td>
<td>9</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>344</td>
<td>8</td>
<td>0.23</td>
</tr>
<tr>
<td>6</td>
<td>284</td>
<td>3</td>
<td>0.71</td>
</tr>
<tr>
<td>7</td>
<td>287</td>
<td>3</td>
<td>0.72</td>
</tr>
<tr>
<td>8</td>
<td>294</td>
<td>3</td>
<td>0.74</td>
</tr>
<tr>
<td>9</td>
<td>254</td>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>176</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>11</td>
<td>129</td>
<td>2</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 5.2: p-values for the observed percentages of homologous pairs. For each pathway distance analysed, the number of pairs considered is listed (pairs were only considered if at least one domain assignment was available for each enzyme) and the number of homologous pairs is also listed. Statistically significant p-values (cut-off: 0.075) are in bold; the homology percentage observed at these distances is not the consequence of a chance distribution of enzymes.

was repeated 500,000 times to derive the average random percentage of positive pairs, its standard deviation and the p-value for the experimental percentages. This procedure is similar to that used in Chapter 4 to identify pathways with a significant and unusually high level of duplication (see Section 4.6.3).

These p-values can be found in Table 5.2; they indicate that the observed percentage duplication for the conserved pathway distances is only significantly different from random at one, two, and three steps (significance cut-off: 0.075); that is, at these distances, the observed level of duplication is significantly higher than expected by chance. Overall, homologous enzymes within the metabolic neighbourhood are rare, accounting for at most 5% of the observed instances. Beyond three steps, the likelihood of homology does not appear to be dependent on pathway distance.

For each SMM protein pair, all the domains shared were considered; the 95 homologous pairs accounted for 113 domains. An attempt was made to classify the rationale for the duplication in these domains. Three possible rationales were considered:

1. Chemistry conserved (where commonalities in the catalytic process dominate)

2. Substrate conserved (identical or similar substrates)
Table 5.3: Domain conservation explanations. The 113 instances of domain duplication are classified, where possible, into one of three categories: chemistry conserved (where conservation of chemistry is the most salient feature), co-factor/minor substrate binding conservation (e.g. the duplicated domain is nucleotide binding domain) and substrate binding conserved (where the duplicated domains bind identical or similar substrates, but the homologous enzymes do not necessarily catalyse the same reaction). In 33 cases, it was not possible to classify unambiguously the recruitment.


Such distinctions are not always obvious to make; often conservation of chemistry implies a common substrate moiety and, likewise, the nature of the chemistry is often linked to the co-factors used (Todd et al., 2001). Therefore, not all instances of recruitment were classified. Furthermore, the sequence domain recruitments were not classified. The most common explanation for domain recruitment is conservation of the catalytic mechanism. This accounts for 39 of the 113 instances of domain duplication (34.5%). Conservation of co-factor binding comes second, accounting for 35 (31%) of the cases. The least common apparent cause of domain duplication is conservation of substrate binding, occurring in 6 instances (5.3%). For the remaining 33 domain duplications, it was not possible to classify unambiguously the instance. These results are summarised in Table 5.3.

5.3.2 Homology and gene intervals

Similarly to pathway distance, the relationship between gene intervals and gene homology was considered. A chromosomal localisation was assigned to the majority of 584 (98%) of the 594 distinct genes present in CONTEXT_PATHWAYS. Therefore, for the majority of enzyme pairs in these pathways, a gene interval (i.e. the number of genes on the E. coli genome separating the two
genes encoding the enzymes in the pair) was derived. This generates a discrete distribution of gene intervals. 4405 E. coli genes are identified in the “Gene Table” (http://www.genome.wisc.edu/pub/analysis/m52orfs.txt(Blattner et al., 1997)) which was used for determining genomic locations; therefore, the largest gene interval possible is 2201 (since only the minimal gene interval on the circular chromosome is considered).

Gene pairs were binned into four gene interval sets i.e. the set of gene pairs separated by zero to five genes, by six to 50 genes, by 51-500 genes, and genes separated by by more 500 genes. The percentage of homologous pairs was calculated for each of these bins. In this chapter, SMM 594 genes are analysed, so theoretically, there are a 176,121 possible gene pairs. However, only 584 genes had an identifiable genomic location (see Table 5.1) and only pairs in which both genes had at least one structural/sequence family assignment were considered. Therefore, the percentage of homologous pairs in the gene interval bins was plotted for a total of 124,750 pairs (Figure 5.2).

![Figure 5.2: Gene intervals and Homology](image)

**Figure 5.2:** Gene intervals and Homology. For each gene interval bin, the percentage of all pairs that are homologous are plotted (bars) and the actual number of observed homologous pairs is printed above each bar (bold type). The line plot shows the total number of pairs considered for each gene interval bin; the numbers of pairs are printed.

### 5.3.3 Gene intervals and pathway distance

For a large number of the 176,121 possible SMM genes pairs, no pathway distance was available (i.e. the two genes were further apart than the 11 steps considered
Figure 5.3: Gene interval and Pathway distance. Gene intervals are plotted against pathway distance for the 3495 gene pairs where both these measures are available.

or they lay in two distinct pathways). Nevertheless, for 3495 pairs, both pathway distance and gene intervals were available; data for these pairs are plotted in Figure 5.3.

The scatter plot in Figure 5.3 shows no obvious pattern, but binning revealed some trends. At each pathway distance, enzyme pairs were sorted into bins gathering pairs with a gene interval between zero and five genes, a gene interval of between six and 50 genes, a gene interval of between 51 and 500 genes, and finally, of pairs with a gene interval greater than 500 genes. The relative contributions of the first three bins at various pathway distance is illustrated in Figure 5.4. There is an evident correlation between pathway distance and the proportion of genes with low gene intervals (0 to 5 genes) (see Figure 5.4a). To determine to what extent this correlation was due to clustering of metabolic genes into operons, a list of *E. coli* operons was obtained from the RegulonDB database (Salgado *et al.*, 2001, and Gabriel Moreno-Hagelsieb, personal communication) and used to flag *E. coli* genes known or predicted to be part of operon structures. Figure 5.4b shows the subset of all pairs in which both genes are part of an operon and Figure 5.4c where both genes are predicted not to be part of an operon. In the former case, the trend observed in Figure 5.4a is more marked, whilst in the latter case it disappears. The influence of operons in the clustering of genes is obvious; a possible extension of this work would be to consider gene clustering in the absence of the “operon effect”, i.e. to determine if genes still cluster (albeit at larger gene intervals) when the signal from
operon structure is eliminated. Such an analysis might for example be performed by deriving an expectation model for operons (Craven et al., 2000).

The plot in Figure 5.4 can be "reversed", considering the relative contributions of genes at a given pathway distance for each gene interval bin, as shown in Figure 5.5.

5.3.4 Pathway distance, chromosomal distance and function

This section concerns data generated in the course of a collaboration with colleagues from the Department of Chemical Engineering of UCL. Again, the correlation of pathways distance with other metrics is considered, but in this section, pathway distance was measured using the linear programming based algorithm described in Section 3.2.6.2. Furthermore, the pathway dataset used was a subset of the HOMOL_PATHWAYS dataset containing only 68 of the latter's 82 pathways (eliminating mostly very short pathways, including some containing only one metabolic step). This dataset is referred to as the CHEMENG_PATHWAYS dataset.

5.3.4.1 Chromosomal distance and pathway distance

In section 5.3.3, the correlation between gene intervals (i.e. the number of genes separating two genes on the circular chromosome of E. coli) and pathway distance is illustrated. A similar correlation was investigated using the CHEMENG_PATHWAYS dataset and pathway distances computed using the linear programming based algorithm. However, instead of using gene intervals as a measure of the distance separating two genes, chromosomal distance was considered. Chromosomal distance simply measures the number of base pairs (bp) separating the two closest codons of the genes. Again, these distances were binned, this time for gene pairs separated by: 0-100 bp, 101-1000 bp, 1001-10,000 bp, 10,001-100,000 bp, 100,001-1,000,000 bp and more 1,000,000 bp (to a maximum of 2,319,607 bp). These data are plotted in Figure 5.6.

There is a clear correlation between pathway distance and genome distance. As pathway distance increases, the percentage of genes separated by short genome distances drops. For pathway distances of one, two, three, and four steps, gene pairs separated by at most 10,000bp (i.e. bins 0-100bp, 101-1,000bp, and 1,001-10,000bp) account for 14.8%, 12.2%, 4.1% and 2.1% of the pairs analysed respectively. For the other three distance bins, no clear trend is evident. These patterns confirm that SMM genes are "metabolically clustered" on the genome. Furthermore, the relatively
Figure 5.4: Pathway distance and Gene intervals. At each pathway distance (x-axis), the percentage of enzyme pairs with a gene interval of 0-5 genes (blue diamonds), 6-50 genes (pink square) and 51-500 genes (yellow triangle) is plotted for (a) all pairs, (b) pairs with both enzymes predicted to be in operons and (c) pairs with both enzymes predicted to be out of operons (operon prediction from (Salgado et al., 2001))
Figure 5.5: Gene intervals, Pathway distance and Homology. For each gene interval bin, the relative percentages of gene pairs with pathway distances 1, 2, 3 and 4-11 steps are plotted for (a) all pairs and (b) only pairs where the genes were found to be homologous. Numbers within bars represent the actual number of pairs observed.
high percentage of metabolic-gene pairs found within 100bp (a very short distance in a ~4.6Mbp long chromosome) suggests that this clustering is the consequence of prokaryotic operon structures in which co-regulated genes are rarely separated by longer distances (Salgado et al., 2000). An interesting feature of these results is that they “tease out” some of the details of the picture shown in Figure 5.5. In the latter figure, enzyme pairs are binned by gene interval. Assuming an average gene length of 1,000bp (a length thought to be uniform in bacterial genomes (Casjens, 1998)), the first of the gene interval bins is the approximate equivalent of a 0-5,000bp bin in terms of chromosomal distance. Here this distance is encompassed by three bins. The main “contributor” to the trend shown in Figure 5.5 are the genes within 0-100bp of one another, and therefore, in all likelihood, contiguous genes. The next chromosomal distance bin, 101-1,000bp, is nearly always the rarest (Figure 5.6). A possible explanation for this is that, since this distance just reaches that of an average gene, it represents an “impossible distance”: two genes will either be contiguous (and hence separated by 100bp or less) or separated by at least one gene (so separated by at least 1,000bp) — thus avoiding the 101-1,000bp bin. This is supported by the data observed for the 1,001-10,000bp bin, which is the second most populated of the “small” (< 100,001bp) bins. Overall, these data confirm the
“metabolic clustering” of *E. coli* SMM genes and support similar conclusions to that
drawn in Section 5.3.3. Furthermore, the “impossible distance” bin suggests that a
measure of gene separation based in gene interval is preferential.

5.3.4.2 Function and pathway distance

Reaction frames in the dataset described above were assigned an EC number by ref­
erence to the EcoCyc database. This is subtly different from other pathway-distance
based calculations herein, where the genes within reaction frames were considered.
The EC number assigned to the reaction frame was used because one reaction frame
can contain one or more proteins and each of these can be assigned one or more EC
numbers. However, the reaction frame is only ever assigned one EC number in Eco­
Cyc — the individual chemistries performed by the proteins involved in the reaction
frame are subsumed and, in effect, the net reaction catalysed in the reaction frame is
assigned an EC number. EC numbers classify reactions within a hierarchical 4-level
scheme (e.g. the reaction catalysed by glyceraldehyde-3-phosphate dehydrogenase
has EC number 1.2.1.12) (Enzyme Nomenclature Committee, 1992, and Table 2.6).
The level to which two enzymes have identical EC numbers can therefore be used as
a measure of the similarity of the function they perform (Martin *et al.*, 1998; Todd
*et al.*, 2001); enzymes with identical EC numbers perform the same biochemical
function, enzymes with only the first EC level in common share only very gen­
eralised functional similarity (e.g. both oxidoreductases). Finally, enzymes with
completely different EC numbers often share little or no functional commonalities.
Therefore, the number of matching EC levels (none, one, two, three or four) is used
as the functional similarity metric. Hierarchy must however be respected: EC num­
bers 1.2.1.12 and 2.2.1.12 classify different reactions (respectively an oxidoreductive
reaction and a transfer reaction). Because they differ in the first digit, regardless of
their identity at the second, third and fourth level, they are considered to have no
matching EC levels.

The EC numbers assigned to the reaction frames in all reaction frame pairs were
compared and the level of EC number conservation determined. The results are
plotted with respect to pathway distance in Figure 5.7.

No obvious correlation between EC number and pathway distance could be estab­
lished. Furthermore, the data show that conservation of EC number is relatively
rare at all distances (the percentage of enzyme pairs with at least two EC levels in
common exceeds 10% only once), although admittedly if conservation of at least one
EC level is considered, then over one in five pairs at pathway distance one conserve
Chapter 5. Pathway Distance, Genome Distance, Homology and Function

Figure 5.7: Pathway distance and enzyme function. At each pathway distance (x-axis), the percentage of enzyme pairs with all \((L1+L2+L3+L4)\), 3 or more \((L1+L2+L3)\), 2 or more \((L1+L2)\), 1 or more \((Level1)\) or no EC levels matching is plotted. The \(L1+L2+L3+L4\) is a subset of the \(L1+L2+L3\) set (which in turn is a subset of \(L1+L2\), etc.), therefore, only the percentages for the sets “None” (no conservation of EC number) and Level1 (one or more EC levels conserved) add up to 100%.

Some function. It is known that the relationship between EC numbers and pathways is complex, with pathways requiring a number of enzyme types to perform their task (Tsoka & Ouzounis, 2001). These preliminary data would suggest that enzymatic chemistries are varied along the substrate conversion routes and that these transformations are not performed in “blocks” of similar catalysis. However, a recent publication suggests clustering of enzymes belonging to the same metabolic class is a common feature of metabolic networks in many organisms: the findings of Alves et al. (2002) and their implications regarding the findings presented in this section are discussed in Chapter 6.

5.4 Homology within reaction frames

Herein, isozymes are defined as homologous proteins which perform in the same catalytic step (reaction frame) in \textit{E. coli} metabolism. This definition is not without weakness; first, co-involvement in the same EcoCyc reaction frame is sufficient to constitute “performing in the same catalytic step”, even if the two proteins have
very different specific functions (e.g. two members of a complex performing different tasks). Secondly, the definition requires there to be detectable homology between the two putative isozymes. To a certain extent, this second requirement balances the first problem, since homologous proteins often, although not always, perform similar chemistries, and in all likelihood would do so if found in the same reaction frame. Nevertheless, such a definition was necessary to perform a large-scale analysis of EcoCyc reaction frames. In addition, two kinds of isozymes were distinguished: "complete isozymes", where the enzymes in question are detected to have identical domain make-up, and "partial isozymes", where the proteins have one or more, but not all, domains in common.

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

One reaction frame may contain more than one set of homologues, because a reaction frame can contain more than one gene product. For example, a reaction frame could contain genes A, B, C and D. If A and B are homologues, and C and D are homologues, but no member of the first set is homologous with a member of the second set, than the reaction frame contains two distinct sets of isozymic proteins: AB and CD. Furthermore, as described above, isozymes can be complete or partial; even within one set of homologues, some members of the set may be complete homologues, whilst others may be only partial homologues. Finally, some completely homologous sets may contain proteins of varying sizes. For example, proteins A and B can be flagged as complete isozymes by virtue of having the same domain make-up, but protein B might be 50 or more residues longer than protein A — suggesting an unidentified additional domain in protein B.

The 95 (complete and partial) isozyme pairs cluster into 59 sets of homologous proteins (with 5 instances of reaction frames containing more than 1 set of distinct isozymes). Where possible, one or more of nine rationales for presence of isozymes was assigned to each set. The nine reasons are listed in Table 5.4 and the 59 sets of isozymes identified are described in detail in Table 5.5. To illustrate the scenarios presented above, a couple of examples selected from Table 5.5 may be considered. Nitrate reductases NarG, NarZ and NapA are all homologous. NarG and NarZ...
are complete isozymes, having the same domain make-up, whilst NapA is a partial homologue to both NarG and NarZ, since NapA contains two domains not detected in NarG or NarZ. The homology for this set of proteins is therefore described as “C{NarG, NarZ}/P” in Table 5.5. NuoM, NuoN and NuoL are all subunits of NADH dehydrogenase 1. They have an identical domain make-up, but NuoM is 509 residues long, NuoN 425 residues long and NuoL 613 residues long. The homology for this set of proteins is therefore described as “C{104, 188}” in Table 5.5; all members of the set are homologues and differences in size, relative to the largest protein (here NuoL) are given.

Most commonly (13 cases), the isozymes had different preferred substrates or minor substrates. For example, AsnA and AsnB both catalyse transamination of aspartate to asparagine, but AsnA uses NH$_3$ as the amine “donor”, whilst AsnB uses glutamine; FabA and FabZ have different-length preferred fatty acid chain substrates. The isozymes were often active during “different conditions” (11 instances), for example the fumarases FumA and FumB are active during aerobiosis and anaerobiosis respectively. In nine cases, the isozymes have different roles, commonly one isozyme was catabolic and the other biosynthetic (e.g. Alr and DadX). In seven cases, the isozymes were part of the same enzymatic complex or constituents of separate (but functionally related) complexes. These cases were difficult to explain unambiguously, although between homologous complexes, homologous polypeptides often performed similar roles (see for example the formate dehydrogenase constituents FdnG, FdoG, FdhF, FdoH, FdnH, Fdni and Fdol in Table 5.5). Different regulation accounted for seven sets of isozymes. For example, the aldolases AroF, AroG and AroH are all subject to different feedback control. Different kinetics (six sets), alternative cellular localisations (three sets), different co-activity (two sets) and different heterogenous groups (one set) were also observed.
<table>
<thead>
<tr>
<th>Rationale</th>
<th>No. instances</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>substrate</td>
<td>13</td>
<td>Isozymes have different “preferred” substrates or have one substrate in common, but differ in another (usually minor) substrate.</td>
</tr>
<tr>
<td>conditions</td>
<td>11</td>
<td>Isozymes are active under different external conditions (e.g. aerobic/anaerobic; growth media).</td>
</tr>
<tr>
<td>different roles</td>
<td>9</td>
<td>Although both enzymes could theoretically catalyse the same metabolic step, one of them (usually through different substrate preference) is the effective enzyme, with the other performing a similar, but distinct, role. This includes isozymes where one enzyme is anabolic and the other catabolic.</td>
</tr>
<tr>
<td>complex</td>
<td>7</td>
<td>Isozymes are active in the same complex or are constituents of separate (but functionally related) complexes.</td>
</tr>
<tr>
<td>regulation</td>
<td>7</td>
<td>Isozymes activity regulated internally (e.g. constitutive/induced expression; different allosteric regulation).</td>
</tr>
<tr>
<td>kinetics</td>
<td>6</td>
<td>Isozymes with different physico-chemical properties (e.g. optimal pH; $K_m$).</td>
</tr>
<tr>
<td>localisation</td>
<td>3</td>
<td>Isozymes have different cellular localisations.</td>
</tr>
<tr>
<td>different co-activity</td>
<td>2</td>
<td>Isozymes are both multifunctional and share only one activity in common. Either can perform the catalysis for this common reaction.</td>
</tr>
<tr>
<td>heterogeneous group</td>
<td>1</td>
<td>Isozymes use a different heterogeneous group (e.g. metal cation).</td>
</tr>
<tr>
<td>unknown</td>
<td>14</td>
<td>No clear rationale could be identified.</td>
</tr>
</tbody>
</table>

**Table 5.4:** The 9 rationales identified in the 59 set of isozymes described in Table 5.5. One set of isozymes can have more than one rationale associated with it, so the total number of instances (No. instances) exceeds 59. For 14 sets of isozymes, no rationale could be identified.
Table 5.5: Isozymes in *E. coli* SMM

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. accA, accD</td>
<td>C</td>
<td>complex</td>
<td>acetyl-CoA carboxyltransferase</td>
<td>AccA and AccD form the α-2-β-2 complex of acetyl CoA carboxyltransferase.</td>
</tr>
<tr>
<td>1. entE, entF</td>
<td>P</td>
<td>complex</td>
<td>enterobactin synthase complex</td>
<td>EntE and EntF are part of the enterobactin synthase complex and have similar, but distinct, substrates.</td>
</tr>
<tr>
<td>1. hcaC, hcaE</td>
<td>P</td>
<td>complex</td>
<td>dioxigenase</td>
<td>HcaE forms the large α-subunit of 3-phenylpropionate dioxigenase. HcaC is a ferredoxin.</td>
</tr>
<tr>
<td>1. nuoG, nuoI, nuoF</td>
<td>C{nuoI, nuoF}/P C{104, 188}</td>
<td>complex</td>
<td>NADH dehydrogenase</td>
<td>A cluster of 13 genes encode NADH dehydrogenase 1. NuoFGI are 4Fe-4S proteins. NuoLMN may be involved in proton translocation.</td>
</tr>
<tr>
<td>1. fdoG, fdoG, fdoH</td>
<td>C{fdoG, fdoG}/P C</td>
<td>complex</td>
<td>formate dehydrogenases</td>
<td>There are three formate dehydrogenases (FDHs) in <em>E. coli</em>: FDH-N, FDH-O and FDH-F. FDH-N is composed of FdnGHI; FDH-O of FdoGHI. The G, H, and I genes are respectively the active site subunits, electron transfer subunits and cytochrome subunit of FDH-N and FDH-O. FDH-N is used during nitrate respiration, FDH-O when shifting from aerobiosis to anaerobiosis. FdhF is linked to a hydrogenase complex and also contains an active site.</td>
</tr>
</tbody>
</table>
### Table 5.5: continued

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. hyaB, hyfG, hycE, hybC</td>
<td>C{hyfG, hycE}/P</td>
<td>complex</td>
<td>hydrogenases</td>
<td>There are four hydrogenase complexes in <em>E. coli</em>, active under different conditions. HyaB, HyfG, HycE and HybC are all hydrogenase “large-subunits”. HyfA, HyfH, HycF and HycB all have 4Fe-4S domains and HyaA, Hyfl, HycG, HybO are all hydrogenase “small-subunits”.</td>
</tr>
<tr>
<td>2. hyfA, hyfH, hycF, hycB</td>
<td>C{hyfA, hycB}; C{hyfH, hycF}/P</td>
<td>complex</td>
<td>conditions</td>
<td></td>
</tr>
<tr>
<td>3. hyA, hyfl, hycG, hybO</td>
<td>C{hya, hybO}; C{hyf, hycG}/P C{146,193}</td>
<td>complex</td>
<td>localisation</td>
<td></td>
</tr>
<tr>
<td>4. hyfB, hyfD, hyfF</td>
<td>complex</td>
<td>conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. narG, narZ, napA</td>
<td>C{narG, narZ}/P</td>
<td>complex</td>
<td>nitrate reductases</td>
<td>Nitrate reductase A is composed of NarGHI, nitrate reductase Z of NarZYZV and periplasmic nitrate reductase is composed of NapABCDFGH. NarG, NarZ and NapA contain the site of actual nitrate reduction. NarH and NarY are electron transfer subunits and NarI and NarV are the cytochrome B like subunits.</td>
</tr>
<tr>
<td>2. narH, narY</td>
<td>C \ C</td>
<td>complex</td>
<td>localisation</td>
<td></td>
</tr>
<tr>
<td>3. narI, narV</td>
<td>complex</td>
<td>conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. aceB, glcB</td>
<td>C{190}</td>
<td>complex</td>
<td>malate synthases</td>
<td>GlcB is most active in cells grown on glyoxylate. AceB is active in the glyoxylate bypass.</td>
</tr>
<tr>
<td>1. acnA, acnB</td>
<td>C</td>
<td>complex</td>
<td>conditions</td>
<td>aconitases</td>
</tr>
<tr>
<td>1. fumA, fumB</td>
<td>C</td>
<td>complex</td>
<td>conditions</td>
<td>fumarases</td>
</tr>
<tr>
<td>1. glpA, glpD</td>
<td>P</td>
<td>complex</td>
<td>glycerol-3-phosphate dehydrogenases</td>
<td>GlpA forms part of the GlpAB catalytic dimer of glycerol-3-phosphate dehydrogenase. GlpD is aerobic, GlpAB is anaerobic.</td>
</tr>
</tbody>
</table>

*continued on next page*
Table 5.5: continued

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. speC, speF</td>
<td>C</td>
<td>conditions different roles regulation</td>
<td>ornithine decarboxylases</td>
<td>SpeF is degradative and inducible, especially at low environmental pH. SpeC is biosynthetic and constitutively expressed.</td>
</tr>
<tr>
<td>1. sodA, sodB</td>
<td>C</td>
<td>conditions heterogenous group</td>
<td>superoxide dismutases</td>
<td>SodA complexes with manganese (Mn) and is aerobic. SodB complexes with iron (Fe) and is both aerobic and anaerobic.</td>
</tr>
<tr>
<td>1. aroK, aroL</td>
<td>C{66}</td>
<td>conditions kinetics</td>
<td>shikimate kinases</td>
<td>AroK has a higher $K_m$ than aroL. The enzymes are differently repressed by tyrosine and tryptophan.</td>
</tr>
<tr>
<td>1. treA, treF</td>
<td>C</td>
<td>conditions localisation</td>
<td>trehalases</td>
<td>TreA is periplasmic, TreF is cytoplasmic. TreA is active under conditions of high osmolarity.</td>
</tr>
<tr>
<td>1. cysK, cysM</td>
<td>C</td>
<td>conditions substrate</td>
<td>acetylserine lyases</td>
<td>CysK is acetylserine lyase A. CysM is acetylserine lyase B and can use thiosulphate instead of sulphide ($H_2S$). CysM is required for efficient cysteine biosynthesis during anaerobic growth.</td>
</tr>
<tr>
<td>1. pheA, tyrA</td>
<td>P</td>
<td>different co-activity</td>
<td>chorismate mutases</td>
<td>PheA acts as both a chorismate mutase and phrenate dehydratase whilst TyrA acts as a chorismate mutase and a phrenate dehydrogenase. Both are succeeded by TyrB, which turns the product of the former into L-phenylalanine and the latter into L-tyrosine.</td>
</tr>
<tr>
<td>1. relA, spoT</td>
<td>C</td>
<td>different co-activity</td>
<td>ppGpp synthases</td>
<td>RelA is a ppGpp synthase and a GTP pyrophosphokinase. SpoT is a ppGpp synthase and a ppGpp pyrophosphohydrolase.</td>
</tr>
</tbody>
</table>

*continued on next page*
<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. tdcB, ilvA</td>
<td>C{185}</td>
<td>different roles</td>
<td>threonine dehydratases</td>
<td>IlvA is biosynthetic, TdcB is catabolic.</td>
</tr>
<tr>
<td>1. entC, menF</td>
<td>C</td>
<td>different roles</td>
<td>isochorismate synthases</td>
<td>EntC is the enterobactin synthesis-specific isochorismate synthase and catalyses a reversible reaction. MenF is the menaquinone synthesis-specific isochorismate synthase and catalyses an irreversible reaction.</td>
</tr>
<tr>
<td>1. alr, dadX</td>
<td>C</td>
<td>different roles</td>
<td>alanine racemases</td>
<td>DadX (catabolic) is induced; Alr (biosynthetic) is constitutive.</td>
</tr>
<tr>
<td>1. sdaA, sdaB, sdhY</td>
<td>C{sdaA, sdaB}/P</td>
<td>different roles regulation substrate</td>
<td>L-Serine/L-Threonine deaminases</td>
<td>SdaA and SdaB are L-serine and L-threonine deaminases; SdhY is only an L-threonine deaminase.</td>
</tr>
<tr>
<td>1. argD, astC</td>
<td>C</td>
<td>different roles</td>
<td>transaminases</td>
<td>AstC (catabolic) has a higher affinity for succinylornithine than for acetylornithine. ArgD is anabolic.</td>
</tr>
<tr>
<td>1. epd, gapA</td>
<td>C</td>
<td>different roles</td>
<td>dehydrogenases</td>
<td>GapA is the effective glyceraldehyde-phosphate dehydrogenase (GAPDH) with some possible erythrose-4-phosphate dehydrogenase (EPDH) activity. Epd is mainly involved in PLP biosynthesis as an EDPH but has low level GAPDH activity.</td>
</tr>
<tr>
<td>1. gltA, prpC</td>
<td>C</td>
<td>different roles</td>
<td>citrate synthases</td>
<td>PrpC is a methylcitrate synthase with only minor citrate synthase activity. GltA is the effective citrate synthase.</td>
</tr>
</tbody>
</table>

continued on next page
<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pflB, tdcE</td>
<td>C</td>
<td>different roles</td>
<td>pyruvate/2-keto-</td>
<td>PflB's principal substrate is pyruvate, TdcE's principle substrate is 2-ketobutyrate, but both can use the other's main substrate.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>substrate</td>
<td>butyrate formate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lyases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. cadA, ldcC</td>
<td>C</td>
<td>kinetics regulation</td>
<td>lysine decarboxylases</td>
<td>CadA is the most active decarboxylase. It is also more thermostable and has a low optimum pH. LdcC is weakly expressed, less active and thermostable, but has a broad pH range with a higher optimum pH.</td>
</tr>
<tr>
<td>1. pykA, pykF</td>
<td>C</td>
<td>kinetics regulation</td>
<td>pyruvate kinases</td>
<td>PykF is remarkably stable. PykA shows only limited cooperativity among phosphoenolpyruvate binding sites.</td>
</tr>
<tr>
<td>1. gpt, hpt</td>
<td>C</td>
<td>kinetics substrate</td>
<td>phosphoribosyl-</td>
<td>Hypoxanthine is the main substrate for hpt, guanine the main substrate for gpt, but both enzymes can use the other's favoured substrate.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transferases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. pdxK, pdxY</td>
<td>C</td>
<td>kinetics substrate</td>
<td>pyridoxine/</td>
<td>There are two distinct activities: pyridoxal kinase (PL) and pyridoxine kinase (PN). PdxK, pyridoxal kinase, has high PN and moderate PL activity. PdxY, pyridoxal kinase 2, has low PN and high PL activity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pyridoxal kinases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. glpQ, ugpQ</td>
<td>C{111}</td>
<td>localisation</td>
<td>glycerolphosphoryl</td>
<td>GlpQ is periplasmic. UgpQ is cytoplasmic. They act on different ranges of phosphodiesterases.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>substrate</td>
<td>phosphodiesterases</td>
<td></td>
</tr>
<tr>
<td>1. aroF, aroG, aroH</td>
<td>C</td>
<td>regulation</td>
<td>2-dehydro-3-deoxy-</td>
<td>These three aldolases have different feedback control and account for different percentages of aldolase activity: AroG (80%), AroF (20%) and AroH (1%).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>phosphohethonate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>aldolases</td>
<td></td>
</tr>
</tbody>
</table>

*continued on next page*
<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. cls, ybhO</td>
<td>C{73}</td>
<td>regulation substrate</td>
<td>cardiolipin synthases</td>
<td>YbhO can use different substrates, however it does not seem to have in vivo activity.</td>
</tr>
<tr>
<td>1. asnA, asnB</td>
<td>C</td>
<td>substrate</td>
<td>transaminases</td>
<td>Both AsnA and AsnB catalyse transamination of aspartate to asparagine. AsnA uses NH₃ as the amine &quot;donor&quot;, whilst AsnB uses glutamine.</td>
</tr>
<tr>
<td>1. fabA, fabZ</td>
<td>C</td>
<td>substrate</td>
<td>β-hydroxyacyl-ACP dehydrases</td>
<td>FabZ has broad substrate specificity acting on short to long fatty acid chains; FabA acts mainly on intermediate length fatty acid chains.</td>
</tr>
<tr>
<td>1. fabB, fabF</td>
<td>C</td>
<td>substrate</td>
<td>acyltransferases</td>
<td>FabB is active in fatty acid elongation, whilst FabF, used in membrane phospholipid synthesis, is not.</td>
</tr>
<tr>
<td>1. ackA, tdcD</td>
<td>C</td>
<td>unknown</td>
<td>acetate/propionate kinases</td>
<td></td>
</tr>
<tr>
<td>1. agaY, gatY</td>
<td>C</td>
<td>unknown</td>
<td>tagatose 1-6 bis-phosphate aldolases</td>
<td></td>
</tr>
<tr>
<td>1. aldA, aldB</td>
<td>C{63}</td>
<td>unknown</td>
<td>aldehydegenases</td>
<td>AldB function predicted by homology to AldA.</td>
</tr>
<tr>
<td>1. argF, argI</td>
<td>C</td>
<td>unknown</td>
<td>ornithine carbamoyltransferase</td>
<td>Trimers of identical and non-identical chains encoded by duplicate genes ArgI and ArgF produce active ornithine carbamoyltransferase. ArgI and ArgF are found at different loci.</td>
</tr>
<tr>
<td>1. ddiA, ddiB</td>
<td>C{58}</td>
<td>unknown</td>
<td>D-alanine-D-alanine ligases</td>
<td></td>
</tr>
</tbody>
</table>

continued on next page
Table 5.5: continued

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale(s)</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. garR, glxR</td>
<td>C</td>
<td>unknown</td>
<td>tartronate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>semialdehyde</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reductase</td>
<td></td>
</tr>
<tr>
<td>1. gntK, idnK</td>
<td>C</td>
<td>unknown</td>
<td>gluconokinases</td>
<td></td>
</tr>
<tr>
<td>1. gpmA, gpmB</td>
<td>C</td>
<td>unknown</td>
<td>phosphoglycerate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mutase</td>
<td></td>
</tr>
<tr>
<td>1. ilvB, ilvI</td>
<td>C</td>
<td>unknown</td>
<td>acetohydroxy-</td>
<td></td>
</tr>
<tr>
<td>2. ilvN, ilvH</td>
<td>C{67}</td>
<td>syntheses (AHAS)</td>
<td>butanoyl synthetic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AHAS)</td>
<td></td>
</tr>
<tr>
<td>1. metL, thrA, lysC</td>
<td>C{metL, thrA}/P</td>
<td>unknown</td>
<td>aspartate kinases/dehydro-</td>
<td>ThrA and MetL are both aspartate kinases and homoserine dehydrogenases. LysC acts as an aspartate kinase only.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>genases</td>
<td></td>
</tr>
<tr>
<td>1. rfbA, rffH</td>
<td>C</td>
<td>unknown</td>
<td>dTDP-glucose</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>pyrophosphorylases</td>
<td></td>
</tr>
<tr>
<td>1. rfbB, rffG</td>
<td>C</td>
<td>unknown</td>
<td>dTDP-glucose</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4,6-dehydratases</td>
<td></td>
</tr>
<tr>
<td>1. talA, talB</td>
<td>C</td>
<td>unknown</td>
<td>transaldolases</td>
<td></td>
</tr>
<tr>
<td>1. tktA, tktB</td>
<td>C</td>
<td>unknown</td>
<td>transketolases</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5: Isozymes in *E. coli* SMM. Isozymes are homologous proteins found within the same reaction frame. We identified 59 such sets of isozymes. Sets of homologous genes are numbered. Where possible, one or more explanations for the presence of homologues within one frame are given. The "homology" of each set is also described. Sets flagged C are completely homologous (i.e. the same domains have been identified in all proteins in the set). Sets flagged P are partially homologous (i.e. they have one or more domains in common—but not all). Certain sets have mixed homologies, with some of the proteins in the set completely homologous, and others only partially homologous. In such cases, the completely homologous proteins are listed in curly-braces. Finally, some completely homologous sets have proteins of varying sizes, where size differences are greater than 50 residues (suggesting unidentified domain(s)); the size differences relative to the longest protein are listed within the curly-braces. See section 5.4 for further detail.
5.5 Inline reuse

Enzymes are sometimes used at two or more different metabolic steps within a pathway. Experimentally, this equates to an EcoCyc reaction frame used more than once in the SMM network. This is not the same as the "virtual homologues" described in Chapter 4, which are a consequence of the arbitrary splitting of the SMM network into overlapping pathways. "Inline reuse" literally means the same gene product is used more than once in the SMM network: one enzyme catalyses several distinct steps in different parts of the network. For example, the enzyme DeoD phosphorylates a number of different purine nucleoside during nucleotide metabolism. Between each of these phosphorylation steps, one or more other enzymes modify the bases.

The occurrence of such reuses at each pathway distance can be tallied. By definition, no inline reuse can occur at pathway distance one (enzymes catalysing successive steps in metabolic pathways), since consecutive EcoCyc reaction frames catalysed by the same enzyme were merged (see Section 3.2.3). Each appearance of a reused enzyme within a network is therefore separated by one or more intervening metabolic steps, referred to as intervening frames (IFs). The tally for reuses can be found in Table 5.6. Inline reuse was only observed at pathway distances two, three and four. Details of the inline reuses are given in Table 5.7.

<table>
<thead>
<tr>
<th>Steps</th>
<th>Number of inline recruitments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5+</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 5.6: Inline reuse in E. coli SMM. The number of inline reuses at each distance is listed. Inline reuses are only observed for distances two, three and four. By definition, there can be no inline reuse of enzymes side by side (pathway distance one), as identical enzymes found in two adjoining EcoCyc reaction frames were merged into a single frame.

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

The case of the gene product of lpdA warrants special attention. LpdA is a dihydrolipoyl dehydrogenase and is a subunit in both the pyruvate and α-ketoglutarate
dehydrogenase complexes. Whilst the overall chemistries performed by these complexes is different, in both cases, the dihydrolipoyl dehydrogenase subunit re-oxidises dihydrolipoamide, a cofactor used in the reactions catalysed by the other subunits of the complexes. In a sense, LpdA therefore acts independently from the rest of the complex and performs the same chemistry on the same substrate at each point of recruitment (Voet & Voet, 1995).
Table 5.7: Instances of inline reuse.

<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Pathway</th>
<th>Intervening gene(s)</th>
<th>#IF (^1)</th>
<th>Recruitment type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>dgoA</td>
<td>Galactonate catabolism</td>
<td>dgoK</td>
<td>1</td>
<td>MF</td>
<td>DgoA is a multifunctional enzyme; it first catalyses the dehydration of D-galactonate, DgoK then phosphorylates the product and the product of the phosphorylation is then lysed by DgoA, acting this time as an aldolase.</td>
</tr>
<tr>
<td>metL/thrA</td>
<td>Homoserine biosynthesis</td>
<td>asd</td>
<td>1</td>
<td>MF</td>
<td>MetL is a bifunctional enzyme performing two non-consecutive reactions, first the phosphorylation of L-aspartate, then, after the dehydrogenase Asd, MetL oxidises L-aspartate-semialdehyde to homoserine. ThrA is an isozyme of MetL, similarly bifunctional and catalysing the same steps as described for metL.</td>
</tr>
<tr>
<td>tktA/tkB</td>
<td>Pentose phosphate pathway</td>
<td>[talA, talB]</td>
<td>1</td>
<td>MS</td>
<td>TktA catalyses the major transketolase activity in <em>E. coli</em>. In this pathway, it acts both on ribose-5-phosphate and xylulose-5-phosphate, producing the substrates for the next reaction catalysed by transaldolases talA and talB, which in turn produce one of the two substrates for the second transketolase reaction listed in the pentose phosphate pathway. TktB catalyses the minor transketolase activity in <em>E. coli</em>; it is an isozyme of TktA and performs identical reactions to those listed for TktA.</td>
</tr>
<tr>
<td>relA</td>
<td>ppGpp metabolism</td>
<td>gppA</td>
<td>1</td>
<td>MS</td>
<td>GTP pyrophosphokinase catalyses the synthesis of guanosine 5'-triphosphate 3'-diphosphate (pppGpp) as well as guanosine 3',5'-bispyrophosphate (ppGpp) by transferring the pyrophosphoryl group from ATP to GTP or GDP respectively. Phosphatase GppA catalyses the transition from pppGpp to ppGpp.</td>
</tr>
</tbody>
</table>

*continued on next page*
Table 5.7: continued

<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Pathway</th>
<th>Intervening gene(s)</th>
<th>#IF (^1)</th>
<th>Recruitment type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>deoD</td>
<td>Nucleotide metabolism</td>
<td>add</td>
<td>1</td>
<td>MS</td>
<td>DeoD is a ubiquitous purine nucleoside phosphorylase multiply recruited within nucleotide metabolism. DeoD catalyses the generalised reaction purine nucleoside + orthophosphate = purine + α-D-Ribose 1-phosphate. In this instance of reuse, DeoD phosphorlyases hypoxanthine and deoxyadenosine, with deaminase Add acting in between.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>deoD</td>
<td>1</td>
<td>MS</td>
<td>Add (deoxyadenosine deaminase/adenosine deaminase) and DeoD mutually bracket one another (i.e. the chain deoD, add, deoD, add occurs in the nucleotide metabolism pathway). Functions of deoD and add are described above.</td>
</tr>
<tr>
<td>hisB</td>
<td>Histidine biosynthesis</td>
<td>hisC</td>
<td>1</td>
<td>MF</td>
<td>HisB encodes a single polypeptide possessing the two enzyme activities: histidinol-P phosphatase and imidazolglycerolphosphate dehydratase. The intervening enzyme, HisC, acts as a histidine-phosphate aminotransferase.</td>
</tr>
<tr>
<td>ndk</td>
<td>Pyrimidine ribonucleoside/tide metabolism</td>
<td>pyrG</td>
<td>1</td>
<td>MS</td>
<td>Ndk is a nucleoside diphosphate kinase with broad substrate specificity: the terminal phosphate of a nucleoside-triphosphate is transferred to a nucleoside-diphosphate. In the first such reaction, UDP is phosphorylated to UTP. UTP is converted to CTP by the CTP synthase PyrG. In turn CTP acts as the nucleoside-triphosphate donor to ADP.</td>
</tr>
</tbody>
</table>

\(^{1}\) Number of intervenes
<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Pathway</th>
<th>Intervening gene(s)</th>
<th>#IF</th>
<th>Recruitment type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>udk</td>
<td>ditto</td>
<td>cdd</td>
<td>1</td>
<td>MS</td>
<td>Uridine kinase Udk phosphorylates both uridine and cytidine. Cytidine deaminase Cdd catalyses the conversion of cytidine to uridine.</td>
</tr>
<tr>
<td>purB</td>
<td>nucleotide metabolism</td>
<td>purH; purA</td>
<td>2</td>
<td>MS</td>
<td>Adenylosuccinate lyase PurB catalyses the removal of fumarate from 5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole and from succinyl-AMP to form AICAR and AMP, respectively. PurH is a bifunctional AICAR transformylase and IMP cyclohydrolase. PurA is an adenylosuccinate synthase. PurH and PurA convert AICAR to succinyl-AMP. (AICAR: aminoimidazole carboxamide ribonucleotide).</td>
</tr>
<tr>
<td>ubiG</td>
<td>Ubiquinone synthesis</td>
<td>ubiH; ubiE; ubiF</td>
<td>3</td>
<td>MS</td>
<td>UbiG catalyses both the O-methylation reactions involved in ubiquinone synthesis. These take place 3 metabolic steps apart. UbiH, UbiE and UbiF catalyse the intervening hydrolysis, methyltransferase and hydroxylase steps respectively.</td>
</tr>
<tr>
<td>ndk</td>
<td>Deoxy-pyrimidine nucleotide/side metabolism</td>
<td>dut; thyA; tmk</td>
<td>3</td>
<td>MS</td>
<td>Ndk's role in pyrimidine nucleotide/side metabolism is described above. It plays a similar role in deoxypyrimidine nucleotide/side metabolism, catalysing the transformation of dUDP to dUTP in one case and from dTDP to dTTP in the other. The intervening enzymes, a pyrophosphatase, a synthase and a kinase convert dUTP to dTDP via dUMP and dTMP.</td>
</tr>
</tbody>
</table>

continued on next page
Table 5.7: Instances of inline reuse. For each instance, the gene recruited, the pathway in which the recruitment occurs, the number of intervening frames between the two occurrences of the recruited gene and the intervening genes are listed, as well as some details (obtained from EcoCyc (Karp et al., 2000)) concerning the recruitment event. Genes square-bracketed together occur in the same reaction frame. The type of recruitment is also indicated (MF: Multifunctional enzyme, MS: Multiple substrate specificity, ID: Identical reaction).

<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Pathway</th>
<th>Intervening gene(s)</th>
<th>#IF†</th>
<th>Recruitment type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpdA</td>
<td>Glycolysis and TCA</td>
<td>gltA; [acnA, acnB]; icdA</td>
<td>3</td>
<td>ID</td>
<td>LpdA is the dihydrolipoamide dehydrogenase subunit of the pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase complexes. As part of the first complex, it is involved in the formation of acetyl-CoA from pyruvate and as part of the second, of succinyl-CoA from 2-oxoglutarate, but in both instances it perform the same chemistry on the same substrate (Voet &amp; Voet, 1995). These steps are connected by TCA enzymes citrate synthase (GltA), aconitases A and B (AcnA/B) and isocitrate dehydrogenase (IcdA).</td>
</tr>
</tbody>
</table>
Chapter 5. Pathway Distance, Genome Distance, Homology and Function

5.6 Discussion

5.6.1 Recruitment of homologous proteins

Homology within pathway distances 1-11 is essentially localised to the shortest of these distances, and overall recruitment of homologous proteins is rare within this range. Even at pathway distance 2, the distance at which recruitment of homologous proteins is most likely, under 5% of the possible enzyme pairs share one or more domains (see Figure 5.1). Of the 3711 enzymes pairs considered (i.e. all the pairs at distances 1-11), only 95 (2.56%) show homology. However, as shown in the previous chapter, recruitment is a common feature in SMM pathways. It therefore follows that much of the homology observed in Chapter 4 is the consequence of recruitment from distances greater than 11 steps, from other pathways, or indeed from non-SMM genes.

Nevertheless, 95 homologous pairs were observed within pathways. These have a bias for short distances with pathway distances 1, 2 and 3 accounting for two-thirds of the cases of homology (see Table 5.2). When homology does occur, the data show it is most likely at the shortest pathway distances. Two patterns emerge: at a global level, recruitment events from the metabolic neighbourhood are rare. Recruitment does take place, but it does so from the most suitable enzyme, not the “nearest” enzyme (Teichmann et al., 2001a). At a local level however, when considering only homology in the metabolic neighbourhood (pathway distances 1 to 11), instances are not uniformly distributed; rather, more occur at shorter distances. The simplest explanation for this observation would be that, for example, for two homologous enzymes A and B found 1 step apart, A was recruited from B or B was recruited from A (adjacent recruitment). However, in the absence of convincing phylogenetic information, the possibility that either A or B was in fact recruited from a homologue at some distance in the SMM network cannot be excluded. Assuming that the observed instances are adjacent recruitment, the drop in homology with increasing distance is consistent with the retrograde model of evolution (Horowitz, 1945, 1965).

Analysis of the data shows that substrate conservation is not the principal explanation for the observed recruitments in the CONTEXT_HOMOL pathways (See Table 5.3). Whilst the pattern of recruitment shown in Figure 5.1 is consistent with retrograde pathway evolution, the rarity of conservation of substrate binding is not. Furthermore, in the absence of phylogenetic data, there is no directional information to discriminate between forward or retrograde evolution.
5.6.2 Nearby pathway enzymes are clustered in the genome

It is known that gene separation can be used as an indicator of shared function (Tamames et al., 1997; Overbeek et al., 1998, 1999) and physical interaction (Dandekar et al., 1998). One possible conception of shared function is proximity in the SMM network, so it might reasonably be expected to observe a distinct trend when plotting pathway distance against gene interval. However, the plot shows a range of gene intervals at each pathway distance (see Figure 5.3). However, the process of binning the gene intervals reveals a clear trend: enzymes coded by nearby genes in the E. coli genome are more likely than distant ones to be close in a pathway (Figure 5.4). The correlation between pathway distance and gene interval when considering all pairs (Figure 5.4a) is strengthened when considering only “operon pairs” (Figure 5.4b) but disappears when considering “non-operon pairs” (Figure 5.4c); so it would appear that operons do account for this correlation. The pattern observed for cumulative percentages at each pathway distance was also investigated (data not shown). By pathway distance four, over 90% of pairs observed with gene interval 0-5 genes had already been encountered. By contrast, only by pathway distance eight was a similar percentage of the pairs with gene interval 51-500 genes observed and, for larger bins, the pathway distance was nine or greater. For SMM genes, not only is an “operon effect” being observed, but this is a short range effect, essentially only clustering genes found at pathway distances of four or less.

This theory was tested by considering the 845 known and predicted operons obtained from RegulonDB (Salgado et al., 2001). Only 104 of these contained at least one pair of SMM genes (i.e. two or more of the 594 genes in our 82 SMM pathways). In 81 of these (78%), all SMM gene pairs with a known associated pathway distance were less than 5 metabolic steps apart and in 72 cases (69%) all possible gene pairs were within five metabolic steps. That is, in nine of the cases, the operons included gene pairs for which no pathway distance was identified (i.e. genes in separate pathways, or at distances greater than 11 steps or containing non-SMM genes), but in 72 of the cases the operon was only composed of SMM genes within 5 steps of one another. The 81 operons obeying the “within five steps” rule account for 235 of the 594 SMM genes (40%). This percentage increases to 58% when considering only the 402 SMM genes known, or predicted, to be in an operon.

Interestingly, a similar “plateau” at pathway distance 4 was observed by Kolesov et al. (2001) and a median size of 3 “same-pathway” gene clusters was observed by Overbeek et al. (1999), with the latter considered an underestimate by the re-
The observation that, in prokaryotes, functionally related genes cluster and that these genes often participate in the same biosynthetic pathway is neither unexpected nor novel and this clustering is generally accepted to be the consequence of the operon gene organisation of prokaryotes. However, this relationship has not previously been quantitatively explored for the whole SMM of an organism and verified on a set of known and predicted operons. By correlating gene interval and pathway distance, the range of the clustering can be measured. Analysis of the *E. coli* genome suggests an average operon size of 3-4 genes (Salgado et al., 2001). It can be concluded that, in general, for *E. coli* SMM enzymes, operons cluster blocks of 3 to 4 genes all within a short (4 steps or less) pathway distance of one another. These operons are possibly co-regulated at a higher level in "uber-operons" (Lathe et al., 2000). This observation constitutes an important rationale for the often exploited use of genomic co-localisation in gene function prediction.

5.6.3 Genome distance, pathway distance and homology

Following the observations that SMM genes nearby on the chromosome often code for enzymes nearby in the SMM network and that enzymes nearby in the SMM are more likely to be homologous, the correlation of genome distance and homology (Figure 5.2) was investigated. Of the 590 enzyme pairs with a gene interval of 0 to 5 genes, 31 (5.25\%) were homologous, whilst for the other bins considered, the percentage of homologous pairs was approximately 2\%. Genes close by in the genome are more likely to be homologous than genes further apart, but homology is still rare. In other words, genes nearby on the genome are likely to be functionally related but not necessarily evolutionarily related.

The three contexts considered in this chapter (genome, metabolism and evolutionary relationship) are presented together in Figure 5.5. Three facts emerge from this investigation:

1. Enzymes close by in the SMM network are often encoded by genes close by in the genome (12\% of pairs of proteins four or less metabolic steps apart are encoded by genes separated by at most five genes).

2. Enzymes close by in the SMM network are more likely to be homologous than distant ones (2.9\% of pairs of proteins four or less metabolic steps apart are homologous compared to only 1.5\% for pairs of proteins separated by more than four metabolic steps).
3. Genes close by in the genome are more likely to be homologous than distant ones (5.2% of pairs of genes separated by five or less genes are homologous compared to 1.7% of pairs of genes separated by more than five genes).

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

5.6.4 Operons, inline-reuse, isozymes and regulation

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

Operons cluster functionally related genes. In the case of SMM genes, they ensure the coordinated presence of enzymes, as the absence of any one enzyme along a linear pathway would block it. Since SMM is a large network, it would not be feasible to place all SMM enzymes under the control of a single promoter. However, it is no more realistic to have all SMM enzymes under individual control. Observations herein suggest a compromise solution: the clustering of nearby (under 5 metabolic steps) pathway genes in “blocks” of 3-4 genes.

Inline-reuse of proteins can also be thought of as a form of co-regulation; expression of a single enzyme guarantees the catalysis of several steps. For multi-substrate (MS) reused enzymes, the catalysed steps are related by chemistry. The latter are classic examples of enzymes which have a broad specificity that have been utilised in the evolving cell (Jensen, 1976). In the case of multifunctional (MF) reuse, chemistries are different at each catalysed step but the fusing of two independently functional entities into one enzyme can be thought of as the ultimate co-regulation mechanism, a scenario known to occur commonly in *E. coli* SMM (Tsoka & Ouzounis, 2000).

Few of the isozymes described above are co-located within an operon structure and therefore within a short distance of one another (only 5 out of our 59 sets had all
isozymes in a set within 5 genes of one another). As described in the previous chapter, lateral gene transfer was not found to play a key role in this observation (Teichmann et al., 2001a). Adjacent genes would suggest a recent duplication event or strong evolutionary pressure to keep the genes nearby. It would appear that for the set of isozymes listed in Table 5.5, the duplication events are not recent and evolutionary pressure has acted to separate the genes to allow segregation of transcriptional control and/or future specialisation of the isozymes. SMM networks are ancient and have had a long time to be segregated and specialised. Nevertheless, a number of the observed instances of homology, in particular the isozymes with no clear rationale for duplication (e.g. argF and argI), could be awaiting functional and regulatory specialisation. It may also be the case that nearby isozymes are more common in recently evolved pathways (Copley, 2000).

The next chapter ties together many of the concepts discussed in the previous two chapters.
Chapter 6

Discussion

6.1 Towards a network representation of pathways

The analysis of SMM has been performed by a number of groups using a per-pathway approach (Tsoka & Ouzounis, 2001; Teichmann et al., 2001a; Saqi & Sternberg, 2001). This strategy has the advantage of simplicity; using the pathway definition of metabolic databases such as KEGG (Tsoka & Ouzounis, 2001; Saqi & Sternberg, 2001) or EcoCyc (Teichmann et al., 2001a) minimises the amount of pathway processing required and gives a familiar representation of metabolism. However, whilst each individual pathway can be considered a separate entity and distinctions such as domain recruitment between and within pathways can be made, SMM is a complex and complete network, and ignoring irreversible reactions, any metabolite in one part of the network is theoretically “synthesisable” from another. The division of the SMM network into distinct pathways is therefore arbitrary (Gerrard et al., 2001). A possible way to deal with this is to ignore these divisions, and instead consider SMM as a single network (Rison et al., 2002; Alves et al., 2002).

The CONTEXT_PATHWAYS dataset described in Section 3.4.1.2 and analysed in Chapter 5 is a step towards this network representation. Many overlapping pathways were merged in an attempt to minimise redundancy in the pathway dataset and maximise the size of the pathways analysed. This “halfway house” strategy can be extended to its logical conclusion, i.e. merging of the pathways into a single network. Indeed, this process need not be done by successive merges of pathways. An alternative approach is to collect all possible metabolic transitions and derive all possible connections between enzymes from them (see for example Küffner et al.)
The network properties of metabolism are well illustrated by Küffner et al. (2000) who collected enzymatic reactions from a large number of sources for many organisms. The number of paths from glucose to pyruvate with a maximum of 9 steps was found to approximate 500,000; a number which dropped to 80,000 paths when stoichiometry and thermodynamics were considered. The consideration of metabolism as a network rather than a collection of pathways is a logical progression in its analysis.

6.1.1 Dealing with ubiquitous compounds

One of the problems with a network approach is the presence of promiscuous metabolites such as ATP. The problem with such compounds is that their inclusion makes it possible for enzymes which have nothing in common but the use of a "promiscuous" metabolite to appear only one step apart (Alves et al., 2002). Alves et al. (2002) dealt with this problem by generating several connectivity matrices, one considering all metabolites as vertices of the network graph and other considering as vertices only those metabolites that are involved in fewer than 10, 50 or 100 reactions, thus excluding metabolites at different levels of promiscuity. Another approach to promiscuous compounds is that used by Schuster et al. (1999), who use the concept of an "elementary flux mode" as a mathematical tool to define and comprehensively describe all metabolic routes that are both stoichiometrically and thermodynamically feasible for a group of enzymes. Conceptually, this equates to a system in which by-products should not accumulate, and metabolites other than initial substrates should not be depleted. In their strategy, ubiquitous metabolites can be treated as external (pool) metabolites and only internal metabolites (i.e. metabolites participating in reactions of the model for which the formation rate is exactly balanced by the consumption rate) need be considered. Using this approach, a system of enzyme catalysed reactions can be decomposed into a set of elementary nodes useful for the systemic organisation and analysis of complex metabolic networks (Schuster et al., 2000). Another approach is that described by Karp & Paley (1994): production rules and heuristics are used to distinguish between main compounds (which lie along the "backbone" of the pathway, and are therefore shared between consecutive steps) and side compounds — very often, ubiquitous compounds are "sidelined", thus avoiding potentially spurious paths through them.
Chapter 6. Discussion

200

Figure 6.1: Attributes of exponential and scale-free network structures. (a) Representative structure of a network generated by the (random) Erdős and Rényi model. (b) The network connectivity can be characterised by the probability $P(k)$ that a node has $k$ links. For a random network, $P(k)$ peaks strongly at $k = \langle k \rangle$ and decays exponentially for large $k$ (that is $P(k) \approx e^{-k}$ for $k \gg \langle k \rangle$ and $k \ll \langle k \rangle$). (c) In the scale free network, most nodes have only a few links, but a few nodes called hubs (shaded) have a very large number of links. (d) $P(k)$ for a scale-free network has no well-defined peak, and for large $k$, it decays as a power law appearing as a straight line with slope $-\gamma$ on a log-log plot (from Jeong et al., 2000).

6.1.2 Network properties of SMM

A lot of interest has recently been expressed concerning the properties of biological networks in general and metabolic networks in particular (see for example Jeong et al., 2000; Wagner & Fell, 2001). Many networks may be represented as graphs, for example actor collaborations (a social network), WWW pages and hyper-references and national power grids. Traditionally, networks of complex topology have been described with the random graph theory of Erdős and Rényi (Barabasi & Albert, 1999). This model assumes that each pair of nodes in the network is connected randomly with probability $p$, leading to a statistically homogeneous network in which most nodes have the same number of links, $\langle k \rangle$; in particular, connectivity follows a Poisson distribution that peaks strongly at $\langle k \rangle$ (Jeong et al., 2000). In an alternative model, vertex connectivities follow a scale-free power-law distribution in which $P(k)$ (the probability of a vertex connecting to $k$ other vertices) follows $P(k) \approx k^{-\gamma}$ (Barabasi & Albert, 1999; Jeong et al., 2000). These two models are illustrated in Figure 6.1.

Analysis of the metabolic pathways of 43 organisms (Jeong et al., 2000) and of the metabolism of *E. coli* specifically (Wagner & Fell, 2001) have shown that these
belong to the class of scale free models, a property subsequently confirmed by others (Alves et al., 2002). An interesting feature of scale-free networks is that they are evolved using two properties: (i) the network continuously expands by the addition of new vertices that are connected to the vertices already present in the systems and (ii) the probability with which a new vertex connects to existing vertices is not uniform; there is a higher probability that it will be linked to a vertex that already has a large number of connections (Barabasi & Albert, 1999). This has bearings on the evolution of pathways, "if early in the evolution of life metabolic networks have increased in size by adding new metabolites, the most highly connected metabolites should also be the phylogenetically oldest" (Wagner & Fell, 2001) — a property which appears to hold true.

The power-law property can be paraphrased as "few with many, and many with few", so in the case of metabolic networks, there are a few highly connected substrates and many poorly connected ones. This property has been found to occur in many biological systems (Qian et al., 2001b), for example, protein fold distribution (a few folds with many representatives and many folds with few representatives (Qian et al., 2001a)); intermolecular interactions of different proteins (a few proteins interacting with many others, many proteins interacting with only a few others (Park et al., 2001)), and domain combinations in E. coli metabolic enzymes (a few domain families combining with many other families, many domain families combining with few partners (Chapter 4 and Teichmann et al. (2001a); Apic et al. (2001))).

Metabolic networks are also small-world networks (Watts & Strogatz, 1998; Barabasi & Albert, 1999; Jeong et al., 2000; Wagner & Fell, 2001), in that even though they may be sparsely connected and highly clustered, the typical separation between two vertices in the graph remains small. This principle is popularly called the "six degrees of separation" concept, known to many from the movie of the same name starring Will Smith and based on a play by John Guare. The concept is that any two individuals on the planet are connected, through a series of acquaintances and acquaintances of acquaintances, etc., involving no more than six individuals. Small-world networks are exploited by aficionados of the "six degrees of Kevin Bacon" game where the aim is to connect actor Kevin Bacon to any other actor using a chain of film co-stars (see http://www.cs.virginia.edu/oracle/).

Again, the small-world nature of metabolic networks has evolutionary implications. In such a network, perturbations would spread rapidly and may allow the metabolism to react to these promptly (Wagner & Fell, 2001). The scale-free and the small-world nature of metabolic networks make them robust and error-tolerant (Jeong et al.,
properties indispensable for survival and evolution.

6.2 Detecting evolutionary relationships in SMM pathways

An essential part of the analysis of SMM is the identification of homology between participating enzymes. Tsoka & Ouzounis (2001) clustered the metabolic proteins of *E. coli* into families on the basis of sequence similarity alone using the GeneRAGE package which can automatically cluster a large protein dataset (Enright & Ouzounis, 2000). GeneRAGE begins with a BLAST-based “all-versus-all” comparison and verifies BLAST assignments and putative multi-domain protein divisions using a Smith-Waterman alignment algorithm. GeneRAGE clustered 548 metabolic enzymes into 405 protein families of which 316 (57%) were single-member families. Sequence information was combined with a comparison of metabolic networks to derive a “phylogeny of pathways” (Forst & Schulten, 2001).

Copley & Bork (2000) investigated homology among the triosephosphate isomerase (TIM) (βα)_8 barrel superfamilies in SCOP and its implication for the evolution of metabolic pathways. They obtained the protein sequences for SCOP (βα)_8 barrels and detected homologies between them using PSI-BLAST. The ubiquity and diversity of (βα)_8 barrels makes their evolutionary relationships difficult to define, in particular with respect to distinguishing instances of convergent and divergent evolution. In the SCOP database, 23 superfamilies of (βα)_8 barrels are defined; within these, members probably have a common evolutionary origin, but the SCOP curators consider there is insufficient evidence to merge further any of these 23 superfamilies. Copley and Bork however, by using carefully validated PSI-BLAST searches, identified probable homology between six of these (βα)_8 barrels, all of which are phosphate-binding. To this canonical phosphate-binding extended superfamily, a further six SCOP superfamilies were linked on the basis of PSI-BLAST searches, structural alignments and careful analysis of key residues. As well as predicting homology between 12 of the 23 SCOP (βα)_8 barrel superfamilies, Copley and Bork derived a phylogeny, based on sequence, structure and function, for the members of these 12 superfamilies involved in central metabolism (i.e. glycolysis, TCA cycle, pentose phosphate pathway, amino acid biosynthesis and nucleotide biosynthesis).

In Chapter 4, a comprehensive investigation of *E. coli* SMM pathways was performed in order to define their structural anatomy (Teichmann *et al.*, 2001a,b). The
study investigated 581 genes involved in 106 EcoCyc SMM pathways. Structural assignments for the proteins encoded by these genes were obtained by scanning the proteins against a library of Hidden Markov Models (HMM) for SCOP domains—an assignment strategy now encapsulated in the SUPERFAMILY database (Gough & Chothia, 2002). Structurally unassigned regions were, when possible, further clustered into sequence families. This provided domain composition and evolutionary relationship information for 501 proteins (88% of the total number). SCOP was also used in a recent structural census of metabolic networks for *E. coli* (Saqi & Sternberg, 2001): SCOP domains sequences were integrated into a non-redundant protein sequence database and *E. coli* SMM proteins PSI-BLASTed against this database. 440 out of 660 proteins (71%) proteins had at least one match to a SCOP domain.

In Chapter 5, a conceptually similar database to SUPERFAMILY was used to identify the evolutionary relationships between 586 *E. coli* SMM enzymes (Rison et al., 2002). The Gene3D database is a database of structural assignments for whole genes and genomes in the CATH domain database (Buchan et al., 2002, 2003). Instead of HMM models for domains, Gene3D uses PSI-BLAST profiles for CATH domains. 382 (65.1%) proteins were assigned to at least one CATH superfamily. Again, structurally unassigned sequences were clustered using sequence comparison methods and an additional 98 enzymes were classified into a sequence family bringing the total number of evolutionarily mapped proteins to 480 (82%). A graphical overview of these assignments (inspired by Saqi & Sternberg (2001)) is shown in Figure 6.2.

Two groups make use of a combination of SCOP and PSI-BLAST. Saqi & Sternberg (2001) scanned *E. coli* SMM protein sequences against a (non-redundant) database of SCOP domain sequences. Using an *E*-value cutoff of 0.0005, hits matching at least half of the SCOP domain were clustered and the target of the best (lowest) *E*-value was taken as the assignment. For the detection of homologous pairs of enzymes, Alves et al. (2002) also used PSI-BLAST, considering enzyme pairs with an *E*-value smaller than 0.001 to be homologous in a study which considered not only *E. coli* but 11 other organisms. These assignments were extended by PSI-BLASTing enzyme sequences against the SCOP database to detect homologies missed by the sequence-alone method. The percentage of homologues that would have been missed without the use of SCOP is roughly between 1% and 8%, and is lower, on average, for Archea than for Bacteria or Eukaryota (Alves et al., 2002).

In many of these studies, the percentage of enzymes assigned a putative structure is high. This is probably because: (i) enzymes are “over-represented” in protein atomic structure databases and (ii) *E. coli* is a model organism and the studies of
Figure 6.2: Domain families in EcoCyc pathways. The 82 pathways of the CONTEXT_PATHWAYS set are ordered, from top to bottom, by the number of distinct domain families identified in their enzymes. A coloured square indicates that at least one member of the 337 domain families identified in the E. coli SMM enzymes has been detected in that pathway. These domains include “standard” CATH domains (classes 1-4); CATH hyperfamilies (class 77) which cluster distinct CATH superfamilies thought to be distant relatives (e.g. certain (βα)₈ barrels); and sequence families (class 88). Similar diagrams for SCOP assignments to KEGG and EcoCyc pathways can be found in Saqi & Sternberg (2001).

Tsoka & Ouzounis (2001) and Saqi & Sternberg (2001), as well as those presented in Chapters 4 and 5, considered only E. coli metabolism. The E. coli SMM protein repertoire was also analysed in terms of its suitability for comparative modelling, a procedure known to perform poorly below 35% identity. The distribution of percent identities of the alignments for E. coli genes with structural matches was bimodal, peaking at 10-20% and 90-100% (Saqi & Sternberg, 2001). This means that many SMM enzymes, even in well characterised organisms such as E. coli, will still prove challenging to model. Naturally, the most effective way of unequivocally detecting evolutionary relationships would be to solve the structures of all metabolic enzymes in all organisms, or at least of representative examples of all SMM enzymes – an aim which may be made easier to achieve by structural genomics initiatives (Erlandsen et al., 2000; Bonanno et al., 2001). For a defined set of proteins (i.e. metabolic proteins in a model organism), this should be achievable and already some “textbook pathways”, such as that of tryptophan biosynthesis, have at least one structural representative of each enzyme (Mayans et al., 2002).
6.3 Domain composition of SMM enzymes

6.3.1 Most common domains

Domains containing both α-helices and β-strands (α/β domains) form by far the largest proportion of domains in SMM enzymes, a trend maintained at the level of each pathway (Martin et al., 1998; Hegyi & Gerstein, 1999; Teichmann et al., 2001a; Saqi & Sternberg, 2001). This bias can be observed in Figure 6.2. The most common fold (i.e. topological arrangement of secondary structures) in the SMM enzymes is the (βα)s barrel (Saqi & Sternberg, 2001); the same census identified the three most commonly occurring superfamilies as the NAD(P)-binding Rossmann domain, the PLP-dependent transferase domain and the P-loop containing nucleotide triphosphate hydrolase domain. A very similar result was found in the ANAT_HOMOL dataset and a similar one in the CONTEXT_HOMOL dataset (see Table 6.1).

The thiamin diphosphate binding domains and PLP-dependent domains both bind cofactors, respectively thiamin diphosphate (or thiamin pyrophosphate, TPP) and pyridoxal 5'-phosphate (PLP). PLP is a cofactor involved mainly in transamination reactions (crucial in amino-acid synthesis), but it is also used in other reactions (e.g. by glycogen phosphorylase); TPP is a cofactor used in aldehyde transfer. The other domains are discussed below.

<table>
<thead>
<tr>
<th>Structural census (Saqi &amp; Sternberg, 2001)</th>
<th>ANAT_HOMOL (Teichmann et al., 2001a)</th>
<th>CONTEXT_HOMOL (Rison et al., 2002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 NAD(P)-binding Rossmann fold</td>
<td>NAD(P)-binding Rossmann-fold domains</td>
<td>Nucleotide binding hyperfamily</td>
</tr>
<tr>
<td>2 PLP-dependent transferases</td>
<td>Thiamin diphosphate-binding fold</td>
<td>[βα]s barrel hyperfamily</td>
</tr>
<tr>
<td>3 4Fe-4S ferredoxins</td>
<td>PLP-dependent transferases</td>
<td>PLP-dependent transferase-like</td>
</tr>
<tr>
<td>4 Thiamin diphosphate-binding fold</td>
<td>P-loop containing nucleotide triphosphate hydrolases</td>
<td>P-loop-containing nucleotide triphosphate hydrolases</td>
</tr>
<tr>
<td>5 P-loop containing nucleotide triphosphate hydrolases</td>
<td>FAD/NAD(P)-binding domain</td>
<td>Electron transport α-β plait</td>
</tr>
</tbody>
</table>

Table 6.1: The top 5 superfamilies detected in E. coli SMM enzymes in three different surveys. The first two surveys use the SCOP database, the last one uses the CATH database. The "FAD/NAD(P)-binding domain" superfamily is the sixth most common superfamily in the Saqi & Sternberg (2001) survey; the "4Fe-4S ferredoxins" superfamily is the sixth most common superfamily in ANAT_HOMOL.
6.3.2 Battery domains

The most obvious connection between the domains listed in Table 6.1 is their involvement in the "energetics" of the cell. ATP can be thought of as the cell's "energy currency", but cells have a second currency "reducing power" (Voet & Voet, 1995). Reductive power is "harnessed" in compounds such as NAD and NADP. Many of the Table 6.1 domains "handle" these currencies. The P-loop hydrolases, a SCOP superfamily and members of the CATH nucleotide-binding hyperfamily, are involved in the hydrolysis of nucleoside triphosphates (NTPs) such as ATP — a highly exergonic reaction coupled to other otherwise endergonic metabolic reactions. The NAD(P)-binding Rossmann domains exploit reducing power and the ferredoxins (electron transport \( \alpha-\beta \) plaits in CATH) transfer electrons from one enzyme system to another (e.g. during oxidative phosphorylation). These domains can therefore be thought of as "battery domains" (Rison & Thornton, 2002); "plugged" into a protein, they provide or store the energy involved in metabolic transformations. The need to energise metabolism accounts for their ubiquity (Saqi & Sternberg, 2001); these "battery domains" are very often associated with one or several other domains performing catalysis and/or substrate recognition (Chapter 4 and Teichmann et al. (2001a); Apic et al. (2001)). Unsurprisingly, ATP and NAD(P) are amongst the most frequently used (Ouzounis & Karp, 2000) and most connected (Wagner & Fell, 2001; Alves et al., 2002) metabolites in SMM.

6.3.3 Domain combinations

As shown in Chapter 4, approximately half the SMM proteins were composed of a single domain and half were multi-domain proteins (Teichmann et al., 2001a). In general, multi-domain members of a structural family combined exhibited a limited repertoire of domain combinations (i.e. most domain families were found to combine with only one or two other families). However, some versatile families (e.g. Rossmann NAD(P)-binders) combined with a large number of other domains. For proteins with identical domain composition the order of domains in the proteins was usually conserved. Thus, domain combinations obey a scale-free pattern in the SMM of \( E. \ coli \). Similar conclusions are drawn on a much larger scale by Apic et al. (2001), who investigated domain combinations in 40 genomes including Archea, Eubacteria and Eukaryota. A possible explanation for the conservation of domain order is given by Bashton & Chothia (2002), who analysed combinations of the classical Rossmann domain with other domains in protein with a solved structure. In most
cases, domain order is conserved because recombination of the domain has occurred only once during the course of evolution. In nearly all of the cases where domains are combined with Rossmanns in both orientations (i.e. found in combination both at the C- and N-termini of the protein), they have different geometrical relationships which give them different functional properties.

### 6.3.4 Domain fusions

The process of combining domains can be thought of as the ultimate form of coregulation (Chapter 5 and Rison et al. (2002)). Interestingly, it has been shown that metabolic enzymes are frequently involved in gene fusion events (Enright et al., 1999; Tsoka & Ouzounis, 2000). In addition, there is a structural bias to domain fusion, with α/β fold domains having a high-level of participation in such events (Hua et al., 2002).

A comparison of the SMM enzymes of *E. coli* and *S. cerevisiae* identified 20 cases of gene fusion or fission: two or more *E. coli* genes fused as one equivalent gene in *S. cerevisiae* and vice versa (Jardine et al., 2002). At most, the fused genes were two metabolic steps apart in metabolism (although sometimes over 10 genes apart in the genome). More cases of fusion than fission were detected (15 cases compared to five) suggesting that, in *S. cerevisiae*, gene fusion may compensate for the absence of operon structures (Zhang & Smith, 1998; Jardine et al., 2002).

### 6.4 Domain distribution

Copley & Bork (2000) found (βα)₈ barrel homologues to be widely distributed both within and between SMM pathways. Similarly, considering other homologous families, the data presented in Chapter 4 show that domains within the same family are widely distributed across pathways, although the presence of homologues within pathways is observed (Teichmann et al., 2001a). In both cases, recruitment usually involves conservation of catalytic mechanisms and/or co-factor binding. Conservation of substrate binding with modification of chemistry was rarely observed, a bias further supported by the data presented in Chapter 5 (Teichmann et al., 2001a; Rison et al., 2002).

In certain circumstances, the percentage of homologues observed was unusually high. Whilst for consecutive enzymes, which are expected to have some similarity in their substrate-binding site, there was little bias towards homology, homology was more
likely for so-called “parallel” enzymes (pairs of enzymes with the same substrate(s) or product(s)), as discussed in Chapter 4. Also, the serial recruitment of domains (i.e. the block recruitment of more than one consecutive enzyme to another pathway) was found to occur rarely, suggesting that novel pathways are not in general derived from block duplication of existing ones (Teichmann et al., 2001b).

Using their sequence families, Tsoka and Ouzounis investigated two mirror aspects of SMM: (i) functional versatility (i.e. the association of families with distinct reactions and pathways) and (ii) molecular diversity (i.e. the distribution of reactions and pathways across families). The authors found that 91% of the enzyme families spanned only one or two distinct Enzyme Commission (EC) numbers. This trend was even more pronounced when only the higher levels of the EC hierarchy were considered. A different picture of functional versatility of SMM enzymes was observed when they considered participation in a SMM pathway as a description of function: the distribution “widened” towards multifunctional families (i.e. families with members participating in more than one pathway). These correlations were “inverted” to investigate molecular diversity: 86% of reaction types were catalysed by a single enzyme family; however, only 12% of pathways spanned a single enzyme family. For Tsoka & Ouzounis (2001), these data suggested that functional versatility (as described by EC number) tended to be well conserved within families—a picture admittedly affected by the large number of single-member families in their dataset. The reverse relationship—the number of enzyme families spanned by a pathway—suggested that biochemical pathways only required a small number of different enzyme types to be effective, with one enzyme type multiply recruited. Saqi & Sternberg (2001) also found that the majority of families had only one to two members in the SMM repertoire, and occurred in only one or two networks, indicating specialisation for a specific biological context.

6.5 Context based analyses

Chapter 5 focused on several contexts: evolutionary propinquity (homology), genomic localisation, functional similarity and metabolic proximity (Rison et al., 2002). There has been a vast amount of research exploring the relationships between these contexts in the last few years, fuelled mainly by the increasing availability of complete genomes (for reviews, see Huynen & Snel, 2000; Rison & Thornton, 2002). Without doubt, context based analyses represent a notable development in the analysis of organisms, in particular the correlation between contexts. The correlation
between six pairs of contexts is discussed below.

6.5.1 The genomic and functional contexts

It has long been known that the clustering of genes on chromosomes is indicative of shared functionality (Tamames et al., 1997; Overbeek et al., 1998, 1999). Considering neighbouring gene pairs in *E. coli* and *H. influenzae*, Tamames et al. (1997) found that these showed significant conservation of function as described by using a generalised functional classification scheme derived from that of Riley (1993). Overbeek et al. (1998, 1999) described the use of “pairs of close bidirectional best hits” (PCBBH). Genes were considered to occur in a “run” if they all occurred on the same strand, and the gaps between adjacent genes were less than 300bp — genes found in the same “run” were said to be close. Two genes in two different genomes were considered “bidirectional best hits” (BBH) if a recognisable similarity existed between them and there was no gene within each genome with a higher similarity than the gene in the other genome. Thus, pairs of genes close in one genome, each making a BBH with a pair of close genes in another genome, formed a PCBBH. PCBBHs were found to be highly unlikely to be the result of a chance arrangement of genes, and to have functional significance by virtue of clustering genes participating in the same pathway (as determined by consulting the WIT database). The study also illustrated the advantage of considering multiple organisms in tandem, which both increased the number of PCBBHs identified, and generated higher resolution in the significance scores. The importance of using multiple organisms was also highlighted by Dandekar et al. (1998). Here, three sets of genomes at large phylogenetic distances were considered. When the level of identity between orthologous proteins of the three genomes compared was less than 50%, gene order conservation was by and large restricted to genes coding for proteins that show physical interaction (Huynen & Snel, 2000). The advantage of using distant organisms, and of requesting the presence of orthologues in three clades (a strategy reminiscent of the “Clusters of Orthologous Groups” (COGs) strategy (Tatusov et al., 1997)), was to minimise the risk that gene order conservation was due to recent horizontal transfer events (Casjens, 1998; Huynen & Snel, 2000). Thus, genome clustering is associated with function as described by a simplified functional classification scheme (Tamames et al., 1997), co-participation in the same metabolic pathway (Overbeek et al., 1999) and even physical interaction (Dandekar et al., 1998). The second of these options is explored specifically when the genomic and metabolic contexts are correlated.
6.5.2 The genomic and metabolic contexts

In Chapter 5, the relationship between the genomic and functional contexts is explored by quantifying the correlation between gene intervals and pathway distance (since the latter is a measure of co-participation in a pathway and hence shared function). As expected, there is a demonstrable relationship between the two contexts, with genes nearby in the genome far more likely to encode enzymes acting close-by in metabolism (over 70% of gene pairs separated by less than six genes encoded enzymes less than three metabolic steps apart).

The correlation between these two contexts was also investigated by researchers compiling the KEGG database (Ogata & Kanehisa, 1996; Ogata et al., 2000). Researchers identified correlated clusters called FRECs (Functionally Related Enzyme Clusters) using an elegant method based on the comparison of two graphs, one a metabolic graph (enzymes as nodes, common compounds as edges), and the other a genome graph (enzyme-encoding genes as nodes with two adjacent genes on the chromosome considered to be connected by a single-edge, ignoring the direction of transcription), as well as a matrix describing the correspondence between the edges of both graphs. Ten organisms were considered. In many, the incidence of FRECs (calculated as the ratio of enzymes in FRECs to the total number of enzymes considered) was high (e.g. 42% for *E. coli* or 34% for *Mycoplasma pneumoniae*), but for *S. cerevisiae*, the percentage was only 6% (as expected for an eukaryote, since eukaryotes do not in general coregulate the expression of their genes by clustering). More surprisingly, the percentage for bacterium *Synechocystis* sp. was 8%, suggesting that this species may have evolved a mechanism to compensate for the dispersion of functionally related genes on the genome (Ogata et al., 2000).

In Chapter 5, it was observed that the correlation between gene intervals and pathway distance was almost entirely due to operon structure. Ogata et al. (2000) also considered the influence of operon structures in the identification of *E. coli* FRECs by comparing their overlap with the known and predicted operons of *E. coli* (Blattner et al., 1997; Salgado et al., 2001). Again, FRECs showed a high level of overlap with operons, suggesting that most of the FRECs found in *E. coli* contain multiple enzyme genes that are likely to be co-regulated by such polycistronic transcription (Ogata et al., 2000).

Overbeek et al. (1999) considered three genes to be a conservative underestimate of the average size of a gene cluster because it was the median size of “same pathway” clusters they found. For Ogata et al. (2000), the average FREC size was three genes (although the median was two genes); by definition, these are clusters of genes
expressing metabolically-close enzymes. The SNAP (Similarity-Neighbourhood Approach) method of Kolesov et al. (2001) is a computational method for finding genes that are functionally related, but do not possess any noticeable sequence similarity. Orthologous genes in different genomes are connected by S-edges and adjacent genes in the same genome are connected by N-edges. Closed graphs alternating S-edges and N-edges, known as SN-cycles, are found to be likely to connect functionally related genes. The pathway coefficient \( (K_p) \) is a measure used by SNAP to assess the functional similarity of genes identified in an SN-cycle — it essentially calculates the proportion of SN-genes within a certain metabolic distance of one another \( (D_t) \). The \( K_p \) of SN-cycles plateaued at a \( D_t \) of 4, suggesting that SN-cycles clustered genes for enzymes within five metabolic steps of each other. In conjunction with the data presented in Chapter 5, these studies suggest that, in general, for \textit{E. coli} SMM enzyme-encoding genes, operons cluster three to four genes all within a short (four steps or less) pathway distance. Although between genomes there is great fluidity in the context of orthologous genes, Lathe et al. (2000) suggest that conservation is maintained at a higher order than that of operons and thus propose the existence of "uber-operons".

6.5.3 The genomic and evolutionary contexts

The relationship between gene intervals and homology was explored in Chapter 5. In the ANAT datasets, for 124,750 genes pairs, both gene interval and homology data were available. Genes close by in the genome (gene interval 0-5 genes) are more likely to be homologous than genes further apart, but homology is still rare. A similar pattern is witnessed using base-pair separation as a measure of the chromosomal interval between two \textit{E. coli} genes. Therefore, the combinations of contexts discussed so far indicate that, at least for SMM, whilst genes nearby on the genome are likely to be functionally related, they are not often homologous. Interestingly, Volkmutth & Alexandrov (2002), investigating the genomes of \textit{S. cerevisiae} and \textit{A. thaliana}, found that "nearby genes tend to share a fold more often than expected by chance alone", and perhaps more intriguingly still, "such evolutionary trace is, surprisingly, present in expression data: genes that are correlated in expression are more apt to share a fold than two randomly chosen genes."
6.5.4 The metabolic, evolutionary and functional contexts

In this section, the correlation between the metabolic context and the evolutionary and functional contexts is discussed. The relevant corpus consists essentially of two studies, that presented in Chapter 5 and the work of Alves et al. (2002). These two works are paralleled in this section and their salient findings are described.

Chapter 5 reveals that metabolically nearby enzymes are more likely to be homologous than distantly separated ones. This dependency was only statistically significant at short pathway distances (1-3 steps). Beyond that distance, the number of homologous pairs observed is not significantly different from that which might be expected by chance. Overall, homologous enzymes within a metabolic neighbourhood (1-11 steps) are rare, accounting for at most 5% of the enzyme pairs encountered. For the homologous proteins, the most common explanation for domain duplication was conservation of chemistry with conservation of co-factor binding a close second. In the same chapter, the relationship between pathway distance and conservation of function as described by EC number was investigated. No clear trend was observed. The highest percentage of enzymes pairs catalysing the same class of reaction (i.e. catalysing a reaction with the same first EC number) was found at pathway distance 1 (~24%). This percentage only rose above 20% again once more at pathway distance 11, although the number of pairs considered at this distance was very small and the high percentage likely to be spurious. Otherwise, the percentage remained between 10% and 18%.

Alves et al. (2002) are more emphatic in their conclusions based on the analysis of the SMM network of 12 organisms. Similarly to the findings presented in Chapter 5, they show that the percentage of homologues that are less than three steps away from each other in the SMM network is significantly higher than would be expected by chance alone in all the organisms they analyse. Furthermore, pairs of non-homologous enzymes in all species are, on average, further away from each other in the network than pairs of homologues. The distribution of homologous pairs catalysing various classes of reactions is also studied; transferases and synthases have a very high percentage of homologue pairs within two reaction steps of each other. The same is true, but to a lesser extent, for hydrolases and lyases. This suggest that homologous enzymes catalysing these four classes of reactions occur within three steps of each other significantly more often than expected if enzymes had been distributed randomly in the network. However, oxidoreductases and isomerases are spread throughout the metabolic network more evenly than the four other classes. Furthermore, in all species, Alves et al. found an association between similarity in
chemistry (as defined by the same first EC digit) and proximity in the network.

To summarise, Alves and colleagues state three key results:

1. The percentage of pairs of homologous enzymes that are less than three steps away from each other in the metabolic network is significantly higher than what is expected had the network evolved randomly.

2. There is often a clustering effect of enzymes belonging to the same class in metabolic networks (i.e. catalysing a reaction starting with the same EC number).

3. In several species, these two effects are linked and there is a particularly strong tendency for homologous enzymes with similar chemistry to be found less than three steps away from each other in the network.

### 6.5.5 Parallel investigations

The work presented in Chapter 5, and published in part in the Journal of Molecular Biology (Rison et al., 2002), is now considered alongside that of Alves et al. (2002). Both works consider metabolism as a network rather than a collection of separate pathways. Both works consider pathway distance and correlate it to enzyme homology and conservation of function. However, Alves et al. (2002) consider organisms other than *E. coli* and include data for a further 11 organisms including archaea and eukaryota.

Alves et al. (2002) redefine the retro-evolution and recruitment models of pathway evolution as respectively, local evolution (homologous enzymes close to each other in the reaction network, independent of the pathway they belong to) and long-distance evolution (homologous enzymes far from each other in the reaction network, independent of the pathway they belong to). A distinction between global and local recruitment is made in Section 5.6.1, but these are not directly associated with any theory of pathway evolution. Both studies observe local recruitment to be significantly higher than expected by chance. In Alves et al.’s study, this local recruitment is significant for enzymes less than three metabolic steps apart, whilst in Chapter 5 of this thesis, it is significant for enzymes less than four steps apart (although recruitment at pathway distance 3 is only significant with a high p-value cut-off of 0.075) (Rison et al., 2002).

For Alves et al., such localised evolution may be used by the cell to minimise metabolic disruption. Because the duplicate of an enzyme is likely (at least in the
early stages) to retain some partial original activity, it will disrupt the metabolism because there will now be two enzymes producing and consuming the reactants that had been produced and consumed by just one. If the duplicate is close to the original in the metabolic network, the disruption will be localised to this site. However, if the “old” enzyme is not close to the new enzyme in the network, there will be a disruption of the physiological state of another part of metabolism, further decreasing the fitness of the cell (Alves et al., 2002).

Alves et al. suggest that a functional clustering effect is observed by comparing the incidence of homologues catalysing the same class of reaction both within two steps and beyond. However, they use a fairly liberal definition of shared function: homologues catalysing a reaction with the same first EC number. Such a trend was not obvious when considering the data presented in Section 5.3.4.2. Since Alves et al.’s study shows that homologues are more likely than predicted by chance to be less than three steps away from each other and it is known that homologous enzymes are often functionally related (Todd et al., 2001), the results may merely suggest a clustering of homologues rather than “functional blocks of similar chemistry”.

The findings of Alves et al. (2002) and that presented in Chapter 5 are compared in Table 6.2.

6.5.6 The evolutionary and functional contexts

The idea of chromosomal and gene duplications as a general source for new genes was proposed more than 60 years ago (Bridges, 1935; Lewis, 1951). Hence, the repertoire of proteins from which to evolve novel functions is limited or, as stated eloquently by Zuckerkandl (1975): “for proteins, evolution soon became a family affair. For them, living systems rapidly turned into a nearly closed club with few new members admitted without family ties to old members.” Thus, the evolutionary and functional contexts are intimately linked. An enormous amount of work therefore concerns the so called “structure-function” paradigm; protein structure is the most evolutionarily informative of protein related knowledge and in essence structure-function analysis is about identifying how homologous protein can generate different functions (divergent evolution) (Murzin, 1998) and non-homologous protein perform the same function (convergent evolution) (Galperin et al., 1998).
The percentage of pairs of homologous enzymes that are fewer than three steps away from each other in the metabolic network is significantly higher than what is expected had the network evolved randomly.

There is often a clustering of enzymes belonging to the same class in metabolic networks. In particular, transferases and synthases — and to a lesser extent, hydrolases and lyases. Isomerases and oxidoreductases (the latter prime “users” of promiscuous NAD(P) and FAD) do not show this trend.

The percentage of enzyme pairs with at least two EC levels in common exceeds 10% only once (at pathway distance 1, with 10.1%); admittedly if conservation of at least one EC level is considered, then over one in five pairs at pathway distance 1 conserve some function (considering all possible reaction frame pairs, not just homologous enzyme pairs).

In several species these two effects [higher chance of homology for metabolically close enzymes; higher chance of function conservation for nearby enzymes] are linked; there is a particularly strong tendency for homologous enzymes with similar chemistry to be found fewer than three steps away from each other in the network.

No direct investigation.

Table 6.2: Comparison of the Alves et al. (2002) and the Chapter 5 studies. The first three row parallel general conclusions made in the studies, the next three rows compare the methodologies.
Chapter 6. Discussion

Furthermore, the mechanisms for the evolution of new functions must be considered, distinguishing between two notions: "chemistry-driven" evolution (Gerlt & Babbitt, 1998) and "substrate-driven" evolution (Farber & Petsko, 1990), concepts often associated with the retrograde (Horowitz, 1945) and patchwork (Yčas, 1974; Jensen, 1976) models of pathway evolution respectively.

The status quo remains that homology is intimately linked with functional similarity; Todd et al. (2001), surveying 468,000 homologous pairs (enzymes/enzymes and enzymes/non-enzymes) found that variation in EC number is rare above 40% sequence identity, and above 30%, the first three digits may be predicted with an accuracy of at least 90%. Below 30% sequence identity however, functional variation is significant and much closer inspection of enzyme structure and related literature is required to understand the molecular basis of these differences. Another observation is that, in general, substrate specificity is diverse across a superfamily, whilst the reaction chemistry is maintained (Todd et al., 2001). Alves et al. (2002) also demonstrate that homologues are more likely to share functional similarities than non-homologues in all of the 12 organisms' SMM they analyse.

A vast amount of literature concerns that relationship between the evolutionary and functional contexts (see references cited above and the reviews by Orengo et al., 1999; Gerlt & Babbitt, 2001; Grishin, 2001). These two contexts are not investigated directly in this thesis. Rather, the consequences of evolution of new protein functions are explored by investigating patterns of homology and function within the SMM system. Nevertheless, the data presented in Chapters 4 and 5 do support the importance of recruitment of chemistry in the evolution of function.
Chapter 7

Conclusions

7.1 Using functional annotation for research

Chapter 2 of this thesis is essentially “self-contained”, but some of the implications of the work are briefly discussed in this section. One of the main problems identified was that, although very similar, many of the functional classification schemes used in the annotation of genomes were not directly comparable (Rison et al., 2000). The second main problem was the mixing of semantically distinct conceptions of functions – the “apples and oranges” problem (Riley, 1998b; Rison et al., 2000; van Helden et al., 2001). To a great extent, the “apples and oranges” problem is solved by using much more rigorously defined classification systems such as the “Gene Ontology” (The Gene Ontology Consortium, 2001). The popularity of the “Gene Ontology” is waxing and it seems likely to become the de facto functional annotation scheme for genomes and to be integrated within a number of other resources such as SWISS-PROT and InterPro (Lewis et al., 2000, and http://www.geneontology.org/). This in turns offers a solution to the first of the problems described above. If the “Gene Ontology” is applied ubiquitously (and if necessary retroactively) to genome and proteome resources, functional comparisons between various systems (e.g. two different organisms) will be greatly facilitated.

Functional annotation schemes really come to the fore in large-scale biology, be it for the annotation of complete genomes (Xie et al., 2002), the validation of data obtained from high-throughput experiments (Eisen et al., 1998), or the interpretation of gene clusters (Tatusov et al., 1997; Enright et al., 2002). The “Gene Ontology” means that one of the most free-form and “unregulated” kind of biological data – the functional data – is increasingly suitable for computer-based analyses.
7.2 The evolution of metabolic pathways

The data presented in this thesis mostly support the growing body of evidence suggesting patchwork evolution as the prevailing pathway evolution strategy (Chapters 4 and 5 and Copley, 2000; Copley & Bork, 2000; Teichmann et al., 2001a; Tsoka & Ouzounis, 2001; Rison et al., 2002; Rison & Thornton, 2002).

Homologues are widely distributed within and between pathways. Nearly half of the sequence families identified in the *E. coli* SMM had members spanning more than one pathway (Tsoka & Ouzounis, 2001). Related \((\beta\alpha)_8\) barrels had a diffuse distribution in SMM pathways (Copley & Bork, 2000). It was also shown that homologues were more commonly to be found distributed across than within pathways (Chapter 4 and Teichmann et al., 2001a). However, Saqi & Sternberg (2001) found that the majority of superfamilies had only one or two members in the SMM repertoire and occurred only in one or two networks — suggesting some families specialise for a particular biological context. There was little order in the process of recruitment (Chapter 4 and Teichmann et al., 2001a) and, when manually derived for \((\beta\alpha)_8\) barrel homologues (Copley & Bork, 2000), phylogenies did not support the notion that the last enzyme in a metabolic chain was necessarily the most ancestral (a state thought to be a logical consequence of retrograde evolution (Copley & Bork, 2000)). In the majority of cases, recruitment of domains conserved chemistry or minor-substrate/co-factor binding; conservation of substrate binding with modification of catalytic activity was rarely observed (Chapters 4 and 5 and Copley & Bork, 2000; Teichmann et al., 2001a; Rison et al., 2002).

Chapter 5 supports the notion of patchwork evolution insofar as enzymes were not commonly found to be recruited from the metabolic neighbourhood. Only 2.6% of enzymes within 11 metabolic steps of one another were found to be homologous. Nevertheless, homology within the metabolic neighbourhood does occur, and when it does, it is more likely to occur at short pathway distances (Chapter 5 and Rison et al., 2002; Alves et al., 2002), including some well known “retrograde-like” instances, suggesting multiple evolutionary mechanisms occurring in concert.

Not only did SMM evolve within catalytic constraints (i.e. the necessity to evolve a chemically efficient network for the production of small molecules), it was also subject to extensive regulatory constraints (to ensure that the SMM is efficiently controlled to deal with changes in both intracellular and extracellular conditions), leading to strategies such as the use of operons, isozymes and the reuse of enzymes (Chapter 5 and van der Meer, 1997; Rison et al., 2002). *E. coli*’s extant SMM
pathways are the result of these pressures.

The picture is complex; further clarification may come from effective phylogenetic analysis of all SMM enzymes (as performed "manually" for metabolic (βα)₈ barrels by Copley & Bork (2000)) and experimental and theoretical investigation of metabolic pathways in not one but many organisms (Dandekar et al., 1999; Forst & Schulten, 2001). The interactions between the genome context, the metabolic context, the functional context and the evolutionary context are certainly worth "mining" for information (see for example Kolesov et al., 2001). Such methods are effective because, as described herein, there are exploitable relationships between all these contexts. It may also be worth moving away from the constraints of the "retrograde versus patchwork" paradigm and to explore alternative, more holistic, theories of pathway evolution (see Chapter 1 and below).

The work presented in this thesis is summarised in Figure 7.1; to a greater or lesser extent, the knowledge gained in one part feeds all others. With the increasing availability of large-scale data and the computational power to store and mine such resources, bioinformatics is already exploiting, and will increasingly exploit, such multi-contextual research.

7.3 Future work

The work presented in this thesis is descriptive, a snapshot of the current state of knowledge of E. coli SMM, and by extension, an investigation of its putative origin. It would be interesting to consider the data in a predictive fashion, that is, to make inferences about the function, metabolic role, genomic localisation, etc. of unknown genes on the basis of their correlations with known genes. An example of such a strategy is the SNAP algorithm described earlier (Kolesov et al., 2001), but many other context based methods have been developed (see for example Pellegrini et al. (1999); Enright et al. (1999); Marcotte et al. (1999) and others reviewed in Eisenberg et al. (2000)).

The ubiquity of "battery domains" (see Section 6.3.2 Rison & Thornton, 2002) is indicative of their special status. Thus, the evolution of SMM "energisation" may be separate from the catalytic evolution. It would be interesting further to investigate recruitment patterns when battery domains are eliminated or, perhaps, when highly promiscuous metabolites are excluded (for example, as done by Alves et al., 2002). It would also be interesting to study only the "battery domains" in the same way that Copley & Bork (2000) concentrated on (βα)₈ barrels.
Figure 7.1: The big picture. The relationships between the research and methods chapters (black boxes) of this thesis are shown. To a greater or lesser extent, all analyses are complementary. Four key concepts (red boxes) “feed” the work. Key papers and resources (italicised) are listed in the blue boxes. Papers written or co-written by the author are in bold.
This thesis has presented a protein-centric view of SMM (Gerrard et al., 2001). Thus, one context remains poorly explored, the "metabolome" — strictly, the complement of all the low molecular weight molecules present in cells. Similarly to investigating the association between enzyme homology and pathway distance, it would be interesting to investigate the change in chemical similarity with pathway distance, perhaps by exploiting the SMILES line notation for entering and representing molecules (Weininger, 1988). Furthermore, the functional context could be refined. Rather than using the EC scheme (Teichmann et al., 2001a; Rison et al., 2002; Alves et al., 2002) or highly simplified schemes (Tamames et al., 1997), it would be interesting to use a finer-grained descriptor. The "Gene Ontology" seems a highly plausible candidate.

Machine learning (ML) techniques are increasingly being exploited in biochemical research. During the course of this thesis, the C4.5 decision tree generation system (Quinlan, 1993) was briefly experimented with to try and identify rules for patterns of EC numbers in the metabolic network. Little came of the work, and the project was abandoned because of lack of time, but mining of the data collected for this thesis using ML remains an interesting challenge.

In particular, the data collected during the course of this thesis concerning enzymes pairs constitute a training set eminently suitable for encoding learnt constraints using machine learning methodologies (Mitchell, 1997). For example, a trained system might predict the metabolic role of an enzyme pair given one or more of the known features of the pair (e.g. genome distance, structural families, pathway distance, EC numbers, etc.)

Work is also continuing, in collaboration with colleagues at the UCL Department of Chemical Engineering, on the linear programming based analysis of the SMM network, in particular looking at properties such error tolerance and investigating highly-connected nodes (Simeonidis et al., 2003).

Finally, as discussed in the Introduction and above, the "retrograde" and "patchwork" theories of pathway evolution are often opposed and yet they are neither mutually exclusive nor competing. Already these notions have been explored (Rison & Thornton, 2002), but they deserve to be thoroughly reviewed (Rison et al., 2003). There may be, in the end, a "grand unified theory of enzyme and pathway evolution". Such a theory will, in all likelihood, be thwarted by Nature. Norman Horowitz ends his 1965 paper regarding the evolution of metabolic pathways with a quote attributed to the geneticist and 1958 Nobel laureate George W. Beadle. The quote is as fitting now as it was then: "It's hard to make a good theory — a theory has to be reasonable, a fact doesn't."
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADP</td>
<td>Adenosine Di-Phosphate</td>
</tr>
<tr>
<td>AICAR</td>
<td>Aminimidazole Carboxamide Ribonucleotide</td>
</tr>
<tr>
<td>AMP</td>
<td>Adenosine Mono-Phosphate</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine Tri-Phosphate</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>bp</td>
<td>Base Pair</td>
</tr>
<tr>
<td>cAMP</td>
<td>cyclic Adenosine Mono-Phosphate</td>
</tr>
<tr>
<td>CAP</td>
<td>Catabolic Activator Protein</td>
</tr>
<tr>
<td>CASP</td>
<td>Critical Assessment of Structure Prediction</td>
</tr>
<tr>
<td>CATH</td>
<td>Class, Architecture, Topology, Homologous superfamily</td>
</tr>
<tr>
<td>CATH-PFDB</td>
<td>CATH Protein Family Database</td>
</tr>
<tr>
<td>CdRP</td>
<td>1-(o-carboxyphenylamino)-1'-deoxyribulose-5'-phosphate</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding Sequence</td>
</tr>
<tr>
<td>COG</td>
<td>Cluster of Orthologous Groups</td>
</tr>
<tr>
<td>DBMS</td>
<td>DataBase Management System</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxy-Ribonucleic Acid</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>FREC</td>
<td>Functionally Related Enzyme Cluster</td>
</tr>
<tr>
<td>FRS</td>
<td>Frame knowledge-Representation System</td>
</tr>
<tr>
<td>FSSP</td>
<td>Families of Structurally Similar Proteins</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HSSP</td>
<td>Homology-derived Secondary Structures of Proteins</td>
</tr>
<tr>
<td>IGPS</td>
<td>Indole-Glycerol-Phosphate synthase</td>
</tr>
<tr>
<td>ImGP</td>
<td>Imidazole Glycerol Phosphate</td>
</tr>
<tr>
<td>ImGPS</td>
<td>Imidazole Glycerol Phosphate Synthase</td>
</tr>
<tr>
<td>IMPALA</td>
<td>Integrating Matrix Profiles and Local Alignments</td>
</tr>
<tr>
<td>ISL</td>
<td>Intermediate Sequence Library</td>
</tr>
<tr>
<td>IUBMB</td>
<td>International Union of Biochemistry and Molecular Biology</td>
</tr>
<tr>
<td>JIPID</td>
<td>Japan International Protein Information Database</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>LCA</td>
<td>Last Common Ancestor</td>
</tr>
</tbody>
</table>

*continued on next page*
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>LISP</td>
<td>LISt Processing</td>
</tr>
<tr>
<td>MGI/MGD</td>
<td>Mouse Genome Informatics/Mouse Genome Database</td>
</tr>
<tr>
<td>MIPS</td>
<td>Munich Information Centre for Protein Sequences</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>NAD(P)</td>
<td>Nicotinamide Adenine Dinucleotide (Phosphate)</td>
</tr>
<tr>
<td>NAD(P)H</td>
<td>Nicotinamide Adenine Dinucleotide (Phosphate), reduced form</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute of Standards and Technology</td>
</tr>
<tr>
<td>npl</td>
<td>New Predecessor List</td>
</tr>
<tr>
<td>NTP</td>
<td>Nucleoside Tri-Phosphate</td>
</tr>
<tr>
<td>OO</td>
<td>Object-Oriented/Orientation</td>
</tr>
<tr>
<td>OO-DBMS</td>
<td>Object-Oriented DataBase Management System</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PEP</td>
<td>Phosphoenolpyruvate</td>
</tr>
<tr>
<td>PIR-PSD</td>
<td>Protein Information Resource-Protein Sequence Database</td>
</tr>
<tr>
<td>PLP</td>
<td>Pyridoxal 5'-Phosphate</td>
</tr>
<tr>
<td>PRA</td>
<td>N-(5'-phosphoribosyl)-anthranilate</td>
</tr>
<tr>
<td>PRAI</td>
<td>N-(5'-phosphoribosyl)-anthranilate isomerase</td>
</tr>
<tr>
<td>ProFAR</td>
<td>Phosphoribosylformimino-5-amino-1-phosphoribosyl-4-imidazole carboxamide</td>
</tr>
<tr>
<td>ProFARI</td>
<td>ProFAR isomerase</td>
</tr>
<tr>
<td>PRPP</td>
<td>5-phosphoribosyl-1-pyrophosphate</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>Position Specific Iterative-BLAST</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position-Specific Score Matrix</td>
</tr>
<tr>
<td>RCSB</td>
<td>Research Collaboratory for Structural Bioinformatics</td>
</tr>
<tr>
<td>rms</td>
<td>root-mean-square</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment and Modeling</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification of Proteins</td>
</tr>
<tr>
<td>SGD</td>
<td>Saccharomyces Genome Database</td>
</tr>
<tr>
<td>SMM</td>
<td>Small Molecule Metabolism</td>
</tr>
<tr>
<td>SNAP</td>
<td>Similarity-Neighbourhood Approach</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SSAP</td>
<td>Sequential Structure Alignment Program</td>
</tr>
</tbody>
</table>

*continued on next page*
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA cycle</td>
<td>Tricarboxylic-acid cycle</td>
</tr>
<tr>
<td>TIM</td>
<td>Triosephosphate Isomerase</td>
</tr>
<tr>
<td>TPP</td>
<td>Thiamine Pyro-Phosphate</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>Translated EMBL</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>TRPS</td>
<td>Tryptophan synthase</td>
</tr>
<tr>
<td>TU</td>
<td>Transcriptional Unit</td>
</tr>
<tr>
<td>UCL</td>
<td>University College, London</td>
</tr>
<tr>
<td>UM-BBD</td>
<td>University of Minnesota Biocatalysis/Biodegradation Database</td>
</tr>
<tr>
<td>WIT</td>
<td>What Is There?</td>
</tr>
<tr>
<td>YPD</td>
<td>Yeast Protein Database</td>
</tr>
</tbody>
</table>
Bibliography


Bibliography


Bibliography


Paper reprints


Abstract In this paper we survey a number of functional classification schemes applicable to genomes. We present the concepts of depth, breadth and resolution as descriptors of the schemes’ scope and architecture and compare selected classifications according to these criteria. We also generate a ‘Combined Scheme’ against which we map six classifications which we believe are representative of the range currently available. The mapping allows the generation of ‘FuncWheels’, which are graphical representations of hierarchical classification schemes. They are used to illustrate similarities and differences in functional space coverage. This survey highlights many issues related to the design and implementation of gene product functional classifications, which are discussed in the light of emerging ‘second-generation’ schemes.

Key words Functional genomics • Protein function • Gene product classification • Ontology

Introduction

The analysis of genes and gene products is usually performed in order to discover, confirm or clarify their function. The function of a gene product is its raison d’être; understanding this function is key to understanding how a limited number of interacting gene products can generate life, from simple unicellular organisms to the incredibly complex multicellular Homo sapiens. The association of functional data with gene products (functional annotation) first appeared in databases of gene products such as SWISS-PROT or PIR, in which protein entries are accompanied by carefully human-generated annotations of their empirically determined or predicted role (Bairoch and Apweiler 1999; Barker et al. 1999). However, although these annotations include keywords chosen from a controlled vocabulary, they are currently not formally organised in a functional annotation scheme, although there have been many efforts to classify such databases on the basis of their annotation (Tamames et al. 1998; Eisenhaber and Bork 1999; Licciulli et al. 1999).

The first extensive gene product functional classification scheme was devised in 1993 to catalogue the 1171 Escherichia coli genes known at the time (Riley 1993). This was some 4 years before the complete genome for E. coli, currently estimated to have approximately 4,300 genes, was sequenced (Blattner et al. 1997). An updated version of the classification scheme was published in 1996 (Riley and Labedan 1996) and regular updates can be found in GenProtEC (Riley 1998a) and EcoCyc (Karp et al. 1999). More recently, genome sequencing projects have been the driving force behind the development of alternative functional annotation schemes.

Once a genome is sequenced, the first step is to identify genes and attempt to annotate the functions of their products. However, in order to understand the overall mechanisms operating, the genes need to be organised according to the biological processes they perform. Such an organisation needs a standardised functional annotation scheme. Functional classification schemes are usually simple hierarchies which begin by defining function in very general terms and become increasingly specific as one progresses down the hierarchy. When dealing with genomes, such schemes allow the gene complement of an organism to be sub-divided into sets of functionally relat-
Table 1 List of gene product classification schemes: references and URLs

<table>
<thead>
<tr>
<th>Functional classification</th>
<th>URL</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenProtEC</td>
<td><a href="http://genprotec.mbl.edu/start/">http://genprotec.mbl.edu/start/</a></td>
<td>Riley 1998a</td>
</tr>
<tr>
<td>EcoCyc</td>
<td><a href="http://ecocyc.pangeasystems.com/">http://ecocyc.pangeasystems.com/</a></td>
<td>Karp et al. 1999</td>
</tr>
<tr>
<td>Sanger Centre (M. tuberculosis)</td>
<td><a href="http://www.sanger.ac.uk/Projects/M_tuberculosis/">http://www.sanger.ac.uk/Projects/M_tuberculosis/</a></td>
<td>Cole et al. 1998</td>
</tr>
<tr>
<td>MIPS: Arabidopsis thaliana Database (MATDB)</td>
<td><a href="http://www.mips.biochem.mpg.de/proj/thal/">http://www.mips.biochem.mpg.de/proj/thal/</a></td>
<td>Mewes et al. 1999</td>
</tr>
<tr>
<td>Proteome.com: YDP and WormPD</td>
<td><a href="http://www.proteome.com/databases/">http://www.proteome.com/databases/</a></td>
<td>Hodges et al. 1999</td>
</tr>
<tr>
<td>MGI: Mouse Genome Database (MGD)</td>
<td><a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a></td>
<td>Blake et al. 1999</td>
</tr>
<tr>
<td>TIGR: Microbial databases</td>
<td><a href="http://www.tigr.org/tdb/mdb/mdb.html">http://www.tigr.org/tdb/mdb/mdb.html</a></td>
<td>n/a</td>
</tr>
<tr>
<td>TIGR: Expressed Gene Anatomy Database (EGAD)</td>
<td><a href="http://www.tigr.org/tdb/egad/egad.html">http://www.tigr.org/tdb/egad/egad.html</a></td>
<td>n/a</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td><a href="http://www.geneontology.org/">http://www.geneontology.org/</a></td>
<td>Ashburner et al. 1999, other documents available on-line</td>
</tr>
</tbody>
</table>

ed gene products and also help to provide an overview of the biology of an organism. There are currently many different schemes used to annotate genomes. Even the interpretation of the Mycoplasma genitalium genome, which with 470 genes is the smallest completed (Fraser et al. 1995), greatly profits from organisation into a scheme.

We surveyed a number of WWW sites with functional classification schemes, and literature references and URLs for these are given in Table 1. We described a number of these schemes in terms of their resolution, depth and breadth; these terms help determine the scope and architecture of the schemes. We then focused on six functional classification schemes which we considered representative of the range currently available: EcoCyc (essentially identical to GenProtEC), TIGR, SubtiList, MIPS/PEDANT, KEGG and WIT. EcoCyc and GenProtEC are updated versions of Riley’s original scheme (Riley 1998a; Karp et al. 1999), while TIGR (Fleischmann et al. 1995) and SubtiList (Moszer et al. 1996) are adaptations of it. The MIPS/PEDANT scheme was developed by the researchers at the Munich Information Centre for Protein Sequences (MIPS) (Frishman and Mewes 1997; Mewes et al. 1997). Finally, KEGG (Ogata et al. 1999) and WIT (Selkov et al. 1998) mainly address regulation and metabolic pathways. Mapping of these schemes onto a ‘Combination Scheme’ allowed us to compare them. This analysis included the generation of FuncWheels, a novel way of graphically depicting gene product functions. Certain schemes, although independently implemented, were not included on the basis of their similarity with schemes already present in the selection. For example, the Mycobacterium tuberculosis genome classification scheme employed at the Sanger Centre is essentially the same as the Riley scheme (Cole et al. 1998) and the COGs scheme is a ‘generalisation’ of the Riley scheme into broader functional categories (Tatusov et al. 1997). Therefore in both of these cases we would expect to see a similar coverage of functional space. Furthermore, we did not include the ‘Gene Ontology’ scheme (Gene Ontology Consortium 1999) for mapping, even though it represents a separate type of functional classification scheme. There were two reasons for this: first, its scope is much larger, and its structure more complex, than the chosen schemes; secondly, the ‘Gene Ontology’ represents a radical rethink of gene product function classification. Therefore, its direct comparison with the chosen schemes would have been difficult and ineffective.

This work highlighted many of the issues involved in functional scheme design and implementation and we discuss these with particular focus on recent developments in this area.

Methods

Classification scheme uploading

All analysed functional classification schemes were available on-line during the course of August 1999 when
the data for this paper were collected. Where relevant and possible, they were uploaded locally and converted to a format suitable for storage in the publicly available PostgreSQL relational database management system (PostgreSQL 1999).

We conceptualised the schemes as trees - sets of connected nodes organised hierarchically. The nodes are functions or functional categories (e.g. 'DNA synthesis' or 'Transport'). Progression from the top (level-1) nodes down to the terminal nodes represents increasingly specific functions. The functions can be identified by means of a hierarchical key (for example, function 5.3.1) in which the first number (5) refers to level-1, the second (3) to level-2 etc.

All the uploaded classification schemes were easily stored in such a format with the exception of the 'Gene Ontology' (Gene Ontology Consortium 1999). The latter is implemented as a directed acyclic graph (DAG), which has a more complex data structure than a tree. A DAG allows a node to have more than one parent and for the edges to distinguish between different types of relationships between nodes. It is not possible to covert a DAG data structure onto a tree structure without some concessions; in particular, we lost the capacity to distinguish between relationship types, and nodes with more than one parent had to be duplicated and inserted separately within the tree structure. Nevertheless, we used this conversion to estimate the depth, breadth and resolution of the 'Gene Ontology'.

Design of the 'Combination Scheme' and scheme mapping

In order to compare the six chosen functional classification schemes, we designed a 'Combination Scheme' (CS) of gene product functions. The CS is not intended as a replacement scheme but was designed solely to facilitate a comparison of the current schemes to appreciate their similarities and differences.

The generation of the CS was iterative. It involved the collation of all functional nodes described in the selected schemes and their organisation into a tentative scheme. Because we wanted the CS to be as simple as possible, the first attempted scheme had two levels only. However, it was soon evident that such a scheme was not viable. We therefore designed a three level scheme that was modified during two rounds of mapping. This generated a CS with a broad coverage of all functions described in the selected schemes without excessive bias towards any one of them. The details of the design of the CS are given below.

All the nodes in all the schemes investigated were collected and obviously duplicated functions or functional categories were eliminated. The complete list was re-organised into a three-level tree with six nodes at the top-level and 73 level-3 nodes. The scheme was manually generated with care but remained arbitrary in many respects. For each scheme investigated, all level-1, 2, 3 and 4 nodes (a total of 1,315 nodes) were compared with nodes in the CS and mapped to the lowest (most specific) CS node possible. To simplify the mapping process, we only allowed a one to one relationship between a node in the mapped schemes and the CS. In certain cases, such a rule made mapping impossible. For example, the node 'Cell growth, Cell division and DNA synthesis' in the MIPS/PEDANT scheme could be mapped onto three different CS nodes. In some instances, where such multi-functional categories overwhelmingly pointed towards one CS node, we mapped onto that node but we usually skipped these functions rather than assign them incorrectly.

In order to keep the CS as universal as possible, we tried to avoid including functions as a separate node which tended to be species specific. For example, 'Sporulation', a property specific to certain organisms including Bacillus subtilis, was present as a function in the SubtiList scheme and could have justifiably been included as an additional 'Organism process' in the CS. However, because the function is specific to a very limited number of organisms, it was subsumed into the more generalised 'Adaptation' category on the basis that 'Sporulation' is usually initiated in response to nutrient starvation.

To identify and eliminate scheme-specific nodes from the CS, we analysed the results of the first mapping and identified all CS nodes associated with only one or two distinct schemes. Each of these nodes was reviewed and either subsumed into another node, combined with other scheme-specific nodes, deleted, reclassified or, in rare instances where the function was considered critical, left unchanged. We then repeated the mapping process to determine coverage of the CS by each of the six selected schemes. A similar iterative process has previously been used to classify SWISS-PROT function annotations (Tamames et al. 1996) in which SWISS-PROT entry keywords were mapped onto a one level scheme, based on the segregation of the Riley scheme into three nodes: 'Energy', 'Information' and 'Communication'. All our mapping procedures were performed using database backed Perl scripts (Wall et al. 1996) and further details on the mapping process, the mapping results and the mapped schemes can be found at http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/.

Generation of FuncWheels

FuncWheels are a graphical representation of all the nodes in a three level classification scheme. The wheel is separated into differently coloured segments each representing a top-level node and proportional in size to the number of level-3 nodes in them. The wheel is also divided into an inner disc and an outer ring. The inner disc of the wheel is divided into segments representing level-2 nodes, again of a size proportional to the number of level-3 nodes in them, whilst the outer ring is divided into equally sized segments each representing a level-3 node (see Fig. 2 for an example of a FuncWheel).
To illustrate the coverage of the CS by mapped schemes, we used FuncWheels in which non-matched CS level-3 nodes were blanked out. In addition, level-3 nodes were considered unoccupied and blanked out if more than two-thirds of their child (level-3) nodes were blanked, unless they spanned only two level-3 nodes when they were considered unoccupied only if neither of the two level-3 nodes was occupied (see Fig. 3 for examples of such ‘coverage’ FuncWheels).

All FuncWheels were generated using a modified version of the software used to generate ‘CATH wheels’ (Martin et al. 1998). Data for the generation of these wheels were extracted from the scheme database using SQL queries (Bowman et al. 1996) and Perl scripts (Wall et al. 1996).

Results

Scheme survey

The functional classification scheme data gathered during this survey are summarised in Table 2, which also includes some data on related classification schemes (e.g. classification of gene products by subcellular localisation or by Enzyme Commission code). Schemes discussed in this paper are in bold in the table, but the other schemes are included for completeness. Table 2 also lists means of accessing the assigned function classification(s) of gene products (e.g. by gene name, by EMBL code etc.).

The surveyed classification schemes were for the most part related to genome sequencing initiatives or analysis of genomes. It is worth pointing out that the ‘Gene Ontology’ scheme – designed by a consortium of researchers affiliated with the ‘Saccharomyces cerevisiae Genome Database’ (SGD), the Drosophila melanogaster database ‘FlyBase’ and the ‘Mouse Genome Informatics’ (MGI/MDG) group – is actually composed of three parts (Gene Ontology Consortium 1999). These are schemes concerning cellular localisation, biological processes and biological function. This distinction of biological process and function is extremely pertinent to the design and implementation of functional classification schemes and will be discussed later.

Single vs multi-organism schemes

Some of the genome-related WWW sites were single organism databases, and the others dealt with multiple organisms. Whilst the MIPS databases included two single organism databases (MYGD and MATD for S. cerevisiae and Arabidopsis thaliana respectively) and a multiple organism database (PEDANT), they all shared one functional classification scheme ‘FunCat’ originally based on yeast gene products but adapted to be applicable to a number of other organisms (Mewes et al. 1997, 1999). Whilst the ‘FunCat’ is used in PEDANT to classify many gene complements, including the partially completed human one, it nevertheless remains yeast orientated, although efforts are being made to extend the scope of the classification (D. Frishman, personal communication). The ‘Gene Ontology’ is being developed with the aim of being applicable to many organisms (Gene Ontology Consortium 1999; Riley 1998b). We note that it is considerably more complex than previous schemes by an order of magnitude.
Table 2 An overview of gene product classification schemes identified during a survey of 16 genome related WWW sites. All the gene product classification schemes encountered are listed; those discussed in the paper are in bold, and where applicable three examples of nodes in the top level of the classification are given. The table also indicates the breadth, depth and resolution of the schemes and lists alternative access routes to gene product data. Schemes empirically determined to support multiple functional annotations of single gene products are flagged in the '1:M annotation' column.

<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resolution</th>
<th>1:M annot.</th>
<th>Alternative access routes to gene product information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single organism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenProtEC</td>
<td>Physiological role</td>
<td>Cell process, Metabolism of small mols, Structural elements</td>
<td>3</td>
<td>10</td>
<td>118</td>
<td>Y</td>
<td>Gene/protein name, Blattner number, SwissProt ID</td>
</tr>
<tr>
<td></td>
<td>Gene type</td>
<td>Enzyme, RNA, Membrane</td>
<td>1</td>
<td>21</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EC scheme</td>
<td>Oxidoreductases, Transferases, Ligases</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EcoCyc</td>
<td>Taxonomy of genes</td>
<td>Cell process, Lipid biosynthesis, Broad regulatory functions</td>
<td>3</td>
<td>6</td>
<td>150</td>
<td>Y</td>
<td>Gene/protein name</td>
</tr>
<tr>
<td></td>
<td>Taxonomy of pathways</td>
<td>Signal-transduction pathways, Intermediary metabolism, Biosynthesis</td>
<td>3</td>
<td>6</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxonomy of reactions</td>
<td>EC-Reactions, Binding reactions, Transport reactions</td>
<td>4</td>
<td>6</td>
<td>305</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxonomy of compound</td>
<td>Hormones, Lipids, All carbo-hydrates</td>
<td>5</td>
<td>16</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institut Pasteur:</td>
<td>Functional categories</td>
<td>Intermediary metabolism, Information pathways, Cell envelope and cell processes</td>
<td>3</td>
<td>6</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubtiList</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institut Pasteur:</td>
<td>Functional categories</td>
<td>Intermediary metabolism and respiration, Information pathways, Cell wall and cell processes</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TuberclList</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPS: MYGD</td>
<td>Functional catalogue</td>
<td>Transport facilitation, Energy, Cellular organisation</td>
<td>4</td>
<td>14</td>
<td>75</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>(FunCat)</td>
<td>Protein classes</td>
<td>ATPases, Transcription factors, Molecular chaperones</td>
<td>4</td>
<td>22</td>
<td>187</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>(ClassCat)</td>
<td>Subcellular localisations</td>
<td>Plasma membrane, Nucleus, Mitochondria</td>
<td>2</td>
<td>15</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(SubcellCat)</td>
<td>EC scheme</td>
<td>As above</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein complexes</td>
<td>Replication complexes, Cell-cycle checkpoint</td>
<td>4</td>
<td>70</td>
<td>317</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CompCat)</td>
<td>Phenotypes</td>
<td>Stress response defect, Sensitivity to antibiotics, Auxotrophy defects</td>
<td>4</td>
<td>11</td>
<td>179</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pathways</td>
<td>Amino acid metabolism, Energy, Signal transduction</td>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPS: MATD (Arabidopsis)</td>
<td>Functional catalogue</td>
<td>Transport facilitation, Energy, Cellular organisation</td>
<td>4</td>
<td>17</td>
<td>255</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>(Arabidopsis thaliana)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genetic element type</td>
<td>tRNA, 5'UTR, Gene/protein</td>
<td>1</td>
<td>14</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 (Continued)

<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resolution</th>
<th>1:M annot.</th>
<th>Alternative access routes to gene product information</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPD (S. cerevisiae) / WormPD (C. elegans)</td>
<td>Functional categories</td>
<td>Binding protein, Ligase, Protein kinase</td>
<td>1</td>
<td>1</td>
<td>55/55 Y</td>
<td>Y</td>
<td>Gene name, Keyword, Sequence similarity</td>
</tr>
<tr>
<td></td>
<td>Cellular role</td>
<td>Amino-acid metabolism, Pol I transcription, Signal transduction</td>
<td>1</td>
<td>1</td>
<td>42/45 Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subcellular localisations</td>
<td>Plasma membrane, Golgi, Nuclear nucleolus</td>
<td>1</td>
<td>1</td>
<td>24/32 Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular environment</td>
<td>Actin-cytoskeleton associated, Protein synthesis factor</td>
<td>1</td>
<td>1</td>
<td>10/10 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genetic properties</td>
<td>XX animals are male, 1 intron, Null lethal</td>
<td>1</td>
<td>1</td>
<td>23/119 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post-translational modifications</td>
<td>Phosphorylation, Ubiquitination, N-linked glycosylation</td>
<td>1</td>
<td>1</td>
<td>30/25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger Centre (M. tuberculosis)</td>
<td>Gene list</td>
<td>Cell process, Lipid biosynthesis, Broad regulatory function</td>
<td>4</td>
<td>6</td>
<td>119</td>
<td></td>
<td>Gene name, Sequence similarity</td>
</tr>
<tr>
<td>MGI: MGD</td>
<td>Phenotypic classification</td>
<td>Biochemical, Anatomical, Physiological</td>
<td>2</td>
<td>8</td>
<td>34</td>
<td></td>
<td>Gene/Protein name, Gene/Protein ID, Sequence</td>
</tr>
<tr>
<td>Multiple organisms dbs</td>
<td>TIGR</td>
<td>Amino acid biosynthesis, Regulatory functions, Cell envelope</td>
<td>2</td>
<td>16</td>
<td>114</td>
<td></td>
<td>Gene name, Function text search, Locus search, Sequence ID</td>
</tr>
<tr>
<td></td>
<td>EGAD</td>
<td>Metabolism, Cell signalling/ cell communication, Cell structure/motility</td>
<td>3</td>
<td>6</td>
<td>49 Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PEDANT databases</td>
<td>Transport facilitation, Energy, Cellular organisation</td>
<td>4</td>
<td>16</td>
<td>240 Y</td>
<td></td>
<td>Sequence ID, Text search, PIR keywords/ superfamilies, PROSITE patterns, PPAM domains</td>
</tr>
<tr>
<td></td>
<td>EC scheme</td>
<td>As above</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Structural classes</td>
<td>All alpha, All beta, Alpha beta</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCOP scheme</td>
<td>All alpha proteins, Alpha plus beta protein, Membrane and cell surface proteins</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathway related</td>
<td>KEGG: GENES</td>
<td>Energy metabolism, Membrane transport, Signal transduction</td>
<td>2</td>
<td>14</td>
<td>105 Y</td>
<td>Y</td>
<td>Gene names, EC numbers, Via DGET/LinkDB, Sequence similarity</td>
</tr>
<tr>
<td></td>
<td>Pathway classification (functional and metabolic)</td>
<td>Energy metabolism, Metabolism of complex lipids, Metabolism of macromols</td>
<td>2</td>
<td>10</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pathway classification (regulatory)</td>
<td>Signal transduction, Ligand-receptor interaction, Molecular assembly</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KEGG: LIGAND/ COMPOUND</td>
<td>Carbohydrate, Lipid, Nucleic acid</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KEGG: LIGAND/ ENZYME</td>
<td>Oxidoreductases, Transfereases, Ligases</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WIT</td>
<td>Intermediate metabolism and bioenergetics, Information pathway, Structure and function of the cells</td>
<td>9</td>
<td>6</td>
<td>3,002</td>
<td></td>
<td>Text search in ORFs, pathways, enzymes, overviews, Ortholog clusters, Operon clusters, Sequence</td>
</tr>
</tbody>
</table>
Table 2 (Continued)

<table>
<thead>
<tr>
<th>Database</th>
<th>Classifications</th>
<th>Examples of top nodes</th>
<th>Depth</th>
<th>Breadth</th>
<th>Resolution</th>
<th>1:M annot.</th>
<th>Alternative access routes to gene product information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ontologies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>Functional</td>
<td>Protein, Ribozyme, Nucleic acid</td>
<td>9</td>
<td>3</td>
<td>1,740</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>primitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cellular</td>
<td>Extracellular, Intracellular, Unlocalised</td>
<td>9</td>
<td>3</td>
<td>385</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>component</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Process</td>
<td>Cell growth and maintenance, Cell communication</td>
<td>11</td>
<td>6</td>
<td>1,667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misc.</td>
<td>COGs</td>
<td>Information storage and processing, Cellular</td>
<td>2</td>
<td>3</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>annotation</td>
<td>processes, Metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scheme depth, resolution and breadth

To gain an understanding of the scope and structure of the surveyed schemes, we collected data on the number of levels, the number of nodes at the top level, and the total number of nodes (see Table 2). These three elements can be used to represent the depth, breadth and resolution of the classifications respectively and are, for a selection of schemes, plotted in Fig. 1.

Depth can be thought of as the potential of the scheme for division into subsets: the greater the depth, the further the scheme allows subdivision into functional groups. For example, the MIPS/PEDANT scheme has a depth of four, and when applied to the *S. cerevisiae* gene complement yields sets of 742 ORFs involved with transcription (level-1), 539 ORFs involved with mRNA transcription (level-2), 411 ORFs involved with mRNA synthesis (level-3) and 30 ORFs involved with chromatid modification (level-4). The depth of a scheme represents the amount of magnification that can be applied to functions; much like a microscope, the higher the magnification, the more specifically one can resolve a particular subset of functions.

The depths of the mapped schemes together with that of the 'Gene Ontology' function and process classifications are plotted in Fig. 1A. The depth indicated in the bar chart is the maximum depth encountered and not all branches of the functional tree necessarily extend that far. Depths ranged from two (TIGR and EcoCyc) to 11 ('Gene Ontology' process scheme). Only the 'Gene Ontology' schemes and the WIT scheme have depths greater than four levels. The WIT database, constructed to aid the reconstruction of metabolic pathways, contains a painstakingly detailed classification of metabolism and information pathway functions (404 terms related to these functions are found at a depth greater than six). The 'Gene Ontology' function and process schemes had a maximum depth of nine and 11 respectively which reflect the intricacy of the scheme.

The next parameter is the resolution of a scheme (Fig. 1B). We used the intuitive hypothesis that schemes with a large number of function nodes were likely to have more specific functional descriptions. To use an analogy from the computer world, if all gene functions are represented as a screen — where the fundamental unit is function rather than pixels — the greater the number of function nodes, the higher the resolution. Resolutions ranged from 52 for SubtiList to 3,002 for WIT. Again the size of the WIT scheme is apparent, as is that of the 'Gene Ontology' schemes, which have a combined resolution of over 3,500 nodes, illustrating the minutiae that has gone into designing these schemes. Depth and resolution are closely linked: the greater the depth and resolution of a scheme, the finer its granularity (Gene Ontology Consortium 1999).

The breadth of the schemes, plotted in Fig. 1C, and represented by the number of nodes at the top level, helps to illustrate the coverage of the scheme. The broadest schemes, TIGR and MIPS/PEDANT, had 16 nodes at the top level and the narrowest is the section of the 'Gene Ontology' dedicated to gene product function with a breadth of three. TIGR and MIPS/PEDANT do offer good coverage of function but judging a scheme by its breadth can be misleading. Whilst the 'Gene Ontology' function ontology has a depth of three, this is because, at the top level, the ontology distinguishes between proteins, ribozymes and nucleic acids. The protein node itself has 16 level-2 nodes (e.g. 'signal transduction' and 'structural protein') many of which tend to be top-level nodes in the other schemes. Therefore, a scheme with a limited breadth does not necessarily have a narrow coverage of function.

The Combination Scheme

The CS, designed to allow comparison between schemes, was generated by compiling all level-1, level-2, level-3 and level-4 nodes in the six selected schemes and joining, splitting, deleting or renaming them during two
rounds of mapping. The first pass mapping was performed to identify CS nodes biased towards one particular node. Of the first-pass level-3 CS nodes, 17 were found to be associated with only one scheme and 12 nodes to be associated with two. As our aim was to avoid such bias, we modified the CS to reduce their incidence. Of the 29 level-3 nodes identified as potentially scheme-specific, 15 nodes were variously grouped into combined nodes, four nodes were deleted, three nodes were subsumed into other nodes and seven nodes were kept unchanged. In the first pass mapping, we skipped 149 of the 864 nodes.

The resulting version of the CS still had three levels and six level-node schemes but now only had 55 level-3 nodes and, with minor modifications, became the working version shown in Table 3 and illustrated as a FuncWheel in Fig. 2. Second-pass mapping of the selected schemes to this CS confirmed that the incidence of over-specific nodes had been minimised. Only 139 nodes were skipped during this second round mapping. The mapping also generated the data used in the generation of FuncWheels for the six selected schemes. Nevertheless, mapping of selected schemes onto the CS was difficult to complete. All schemes use umbrella terms (especially at the higher levels) and some of these did not resolve well onto our CS; by extension, it was not always trivial to unambiguously reclassify the children nodes of such umbrella terms within the CS. The CS is amongst the 'smallest' of the schemes, with only 77 nodes, and yet it could accommodate all the other schemes combined, even those with markedly more nodes such as the 254 belonging to the MIPS/PEDANT scheme. This is a good indicator of the level of subsuming involved in generating and mapping to the CS, and explains why we do not recommend its use as a substitute scheme.

The mapping was also subjected to a number of arbitrary assignments, e.g. when distinguishing functions relating to energy metabolism from those concerned with small molecule metabolism. As far as possible, we tried to be consistent; for example, the tricarboxylic acid (TCA) cycle, a functional node found in many of the schemes, was always mapped onto 'energy metabolism - carbon' (CS 1.1.2) regardless of whether it was under a different parent node (e.g. 'Carbohydrate metabolism') in the mapped scheme.

It is interesting to note that the final CS is similar to the eight top-node scheme employed by Tamames et al. (1997) in their analysis of functionally related genes in Haemophilus influenzae and E. coli although it was designed entirely independently. The Tamames scheme was adapted from the TIGR scheme (Fleischmann et al. 1995) and found to be a good compromise between functional specificity and ease of use for the analysis of genomes.

Functional scheme comparison

A full list of mapping assignments, along with further details regarding the mapping process, can be found in our WWW site (http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/). The mapping allowed us to compare
Table 3 The ‘Combination Scheme’ (CS). The hierarchical CS was used as a common reference to compare various classification schemes. The CS has six level-1 nodes, 16 level-2 nodes and 55 level-3 nodes. Numbers represent the key of the functions (e.g. CS 4.1.1 is ‘cell membrane’).

<table>
<thead>
<tr>
<th>Level-1 Nodes</th>
<th>Level-2 Nodes</th>
<th>Level-3 Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Energy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1.1 autotrophic (energy) metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1.2 energy metabolism (carbon)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1.3 energy transfer/ATP-proton motive force</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2 Macromolecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.1 polysaccharides, lipopolysaccharides</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.2 proteoglycans, glycoproteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.3 phospholipids, glycolipids, lipoproteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3 Small molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.1 amino acid metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.2 nitrogen/sulphur metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.3 nucleotide/nucleoside metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.4 phosphorus metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.5 carbohydrate metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.6 lipid, fatty acid and sterol metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.7 biosynthesis of vitamins, co-factors and prosthetic groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.8 secondary metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Cell processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.1 cell division</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.2 signal transduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.3 protein targeting/protein destination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.4 cell regulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2 Organism processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2.1 adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2.2 protection responses/detoxification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2.3 responses to stimuli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Transport</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Large molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.1 protein, peptide transport</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.2 transport of nucleic acids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2 Small molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2.1 ion channels/pores/ion transporters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2.2 sugar and carbohydrate transporters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2.3 amino-acid/amino transporters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2.4 nucleoside, nucleotide, purine and pyrimidine transporters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2.5 ABC transporters/transport ATPases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Structure and organisation of structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1 Cell envelope/membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1.1 cell membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1.2 cell wall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 Cell exterior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2.1 surface structures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2.2 surface polysaccharides/antigens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3 Ribosome related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3.1 ribosomal RNAs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3.2 ribosomal proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4 Other structural elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4.1 chromosome related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4.2 organelle related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Information Pathways</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1 DNA related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1.1 DNA synthesis and replication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1.2 DNA restriction/modification and repair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1.3 DNA recombination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1.4 DNA degradation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2 RNA related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2.1 RNA synthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2.2 transcription related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2.3 RNA modification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2.4 RNA degradation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3 Protein related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.1 protein synthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.2 aminoacyl-tRNA synthetases/transferases and aminoaoyl-tRNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.3 translation related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.4 protein modification/phosphorylation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.5 protein folding/chaperoning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3.6 protein degradation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Miscellaneous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1 Elements of external origin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1.1 phage/virus related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1.2 transposon and related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1.3 plasmid/collincin related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2 unclassified/unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2.1 unclassified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2.2 unknown function</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Schemes by generating a set of coverage FuncWheels as shown in Fig. 3. In these, CS nodes not represented in each of the six selected schemes are blanked out (see the Methods section for full details on the generation of FuncWheel). Blanked nodes can be determined by comparison of the coverage FuncWheels with the CS FuncWheel in Fig. 2.

The most extensive coverage of the CS is provided by the MIPS/PEDANT scheme with only five level-3 nodes unoccupied (Fig. 3D). The MIPS/PEDANT scheme is also the only scheme to have all its level-2 nodes occupied. Conversely, the KEGG scheme, with 11 out of 16 level-2 nodes blanked, has the lowest overall coverage: level-1 segments for 'Processes', 'Transport', and 'Information pathways' are almost entirely blanked out although the scheme has good coverage of metabolism (Fig. 3E).

The WIT scheme has good overall coverage except for the 'Processes' level-1 segment (and the relatively trivial 'Miscellaneous' level-1 segment). WIT is also unsurpassed in its coverage of metabolism and is the only scheme with no blanks in that segment (Fig. 3F).

It comes as no surprise that EcoCyc (Fig. 3A) and TIGR (Fig. 3B) exhibit very similar coverage as they are both based on the Riley scheme. The SubtiList scheme is partially based on the Riley scheme but a number of functions have been combined and adapted for B. subtilis with consequent partial loss of CS coverage (Fig. 3C).

In the future, we will use FuncWheels to graphically depict the functional coverage of fully sequenced genomes and we hope to gain insight into the functional distinctions that characterise genomes. A similar comparison of genomes on the basis of their gene product function distribution performed on 44 organisms using a very simplified three-node scheme, highlighted differences between viruses, bacteria, eukaryotic unicellular organisms, plants and animals (Tamames et al. 1996).
Discussion

Mapping and scheme comparison limitations

Clearly, the comparison of schemes depends on their mapping to the CS. This mapping is not straightforward and is constrained by the requirement for one-to-one correspondence between a node in the mapped schemes and a node in the CS. Therefore, the absence of mapping to a CS node can mean one of three things:

1. The CS node is not represented in the mapped scheme. For example, the KEGG scheme, at the time of data gathering, did not explicitly describe functions pertaining to 'Structure', and therefore this segment in the KEGG FuncWheels (Fig. 3E) is completely blanked out.

2. The mapping process has assigned nodes that could have mapped to the 'missing' CS node elsewhere. For example, Fig. 3B shows that the CS nodes 'Cell membrane' (CS 4.1.1) has not been 'mapped to' by the TIGR scheme, yet this scheme has a 'Cell Envelope' node. Two of the nodes under the TIGR 'Cell envelope' node could have been mapped to CS 4.1.1: 'biosynthesis of surface polysaccharides and lipopolysaccharides' and 'lipoproteins'. The former was more accurately mapped to 'Surface polysaccharides/antigens' (CS 4.2.2), the latter was mapped to 'metabolism of phospholipids, glycolipids and lipoproteins' (CS 1.2.3) and therefore, the CS 4.1.1 node appears unoccupied.

3. The mapping process could not resolve ambiguity of broad-coverage nodes. In the SubtilList scheme FuncWheel, the majority of transport related functions (third segment) are blanked out. However, the SubtilList scheme does include the function node 'Transport/binding proteins and lipoproteins'; most of the CS level-3 transport related nodes could be subsumed by this broad function, but because we cannot map this node specifically to any of them, they appear unoccupied.

In view of these limitations, we reiterate that the comparison of the mapped schemes to the CS is a means of getting an approximate overview of the schemes' coverage. A different person repeating the mapping would doubtless have emerged with somewhat different coverage FuncWheels but not, we believe, to the extent of changing the overall conclusions that could be drawn from them.

In addition, although the breadth, depth and resolution descriptors used herein offer a good handle to compare...
better than one with small breadth, depth and low resolution. Broad schemes tend to be used to offer users rapid access to large functional categories, but this means that super-sets of these categories must be constructed manually. For example, to generate the equivalent of the 'Information Pathways' node in the CS, the 'Transcription', 'Translation' and 'DNA metabolism' nodes of the TIGR scheme must be combined. Deeper schemes allow users to identify gene products associated with quite specific functions without having to resort to alternative functional information databases (e.g. SWISS-PROT), but complicate access to gene product data. High-resolution schemes may indicate focus on a particular area of functional classification (e.g. the WIT scheme and metabolism) or simply reflect the extent of the scheme. A high-resolution scheme may be crucial for the expert user but may prove dauntingly complex to others. Different depths, breadths and resolutions reflect different functional classification strategies and goals on the part of their implementers and cater to different needs on the part of their users. This is well illustrated by the COGs scheme (Tatusov et al. 1997) where the combination of some nodes in the 'Riley scheme' generates nodes with broader functional coverage (i.e. coverage remains the same but depth, breadth and resolution are reduced). This is needed to classify the COGs, which group together related proteins, with similar but sometimes non-identical functions. Conversely, the WIT scheme requires very detailed description of function (i.e. a deep, broad scheme with high resolution) to allow the development of metabolic models (Selkov et al. 1998).

Meaning of scheme level

In some classification schemes, levels have a semantic value. For example, in the Enzyme Commission (EC) scheme, a four-level hierarchical scheme of enzyme-catalysed reactions, the first level represents the major class of enzyme activity (e.g. 'transferases' or 'hydrolases'), and the second, the group or bond acted upon (e.g. 'transferring phosphorus containing groups' or 'acting on peptide bonds') (IJMB 1992). Such semantic 'level-meaning' is absent in the surveyed schemes. Levels are often used to divide functions into subsets, but the rationale for this subdividing is dependent on the parent node (e.g. if the parent node is 'amino-acid metabolism', the children nodes usually relate to the metabolism of a specific amino-acid) rather than an intrinsic property of the level. Resolution and depth in the schemes is therefore not consistent for all branches of the functional trees. It seems very unlikely that an overall functional classification scheme could be designed with semantically meaningful levels. Perhaps though, such meaning could be implemented within specific subsets of the scheme, for example by classifying all transport related functions using a system such as the 'Transport Commission' system (Saier 1998).

Function, apples, and oranges

One of the main issues bearing on functional classification schemes derives from a more philosophical question: "What is function?" Function is an umbrella term, e.g. a gene product can be described in terms of its biochemistry, molecular activity, cellular function and physiological role (Rastan and Beeley 1997). These functions are distinct and different. Consider the human serine protease trypsin: biochemically it catalyses the hydrolysis of peptide bonds following lysine or arginine residues in peptides, its molecular activity is as a proteolytic enzyme, its cellular function is protein degradation, and its physiological role is to aid digestion. Such distinctions are rare in functional classification schemes. In her review of systems for cataloguing the functions of gene products, Riley (1998b) points out that many schemes juxtapose the 'apples and oranges' of function and combine different aspects of gene product function, such as biochemical and physiological function, into a one-dimensional list. This problem is inherent in the surveyed schemes, which all mix 'apples and oranges'. Similarly the CS includes, for example, the nodes 'cell regulation' (CS 2.1.4), a physiological function, and 'ion channels' (CS 3.2.1) a molecular function. The current schemes cannot be merely re arranged to tackle this; separating the apples from the oranges requires a fundamental reclassification. This remains one of the most pressing and complex issues to be resolved for effective functional classification of gene products.

The 'Gene Ontology' illustrates a possible solution to this problem by distinguishing function in terms of three organising principles: gene product function, process and cellular localisation (Gene Ontology Consortium 1999). The function of a gene product is defined as 'a capability that a physical gene product (or gene product group) carries as a potential'. To avoid confusion with the more general use of the term function, this organising principle is also known as a 'functional primitive'. Examples of functional primitives include broad terms (e.g. 'enzyme' and 'transporter') and narrower ones (e.g. 'adenylate cyclase'). Process is defined as 'a biological objective accomplished via one or more ordered assemblies of functions'; e.g. 'cell growth and maintenance', or more specifically 'pyrimidine metabolism'. The division between organising principles is, however, not always definitive. The term 'signal transduction', for example, exists within both the function and process categories.

Multi-dimensionality and multi-functionality

The obvious solution to dealing with the umbrella term 'function' would be to distinguish carefully all these different aspects of function and to describe a gene prod-
uct's function in terms of each of them. This solution is encapsulated in the concept of multi-dimensionality of classification schemes as proposed by Riley (1998b). The three organising principles of the 'Gene Ontology' represent three functional dimensions (biochemical for functional primitive, cellular and physiological for process and spatial for the cellular localisation). Such a classification is invaluable in understanding the role of a gene product. This is illustrated by the comprehensive-ly annotated Yeast Protein Database (YPD) (Hodges et al. 1999). Each gene product in the YPD is annotated in up to six different dimensions: genetic properties, functional category, post-translational modification, cellular role and subcellular localisation (see Table 2 for examples of nodes in these categories). Although each of these dimensions is only a list (i.e. a scheme with only one level, and resolution equal to breadth), the combined information described by these six parameters permits the gene product to be positioned very accurately within the functional space.

Another aspect of multi-dimensionality concerns the hierarchical classification of functions within schemes; certain functions can be involved in a number of more generalised functional classes. In the 'Gene Ontology', the functional node 'ATP-binding and phosphorylation-dependent chloride channel' is an example of an 'intracellular ligand-gated ion channel', a 'chloride channel' and a 'transmembrane conductance regulator'. This is handled in the 'Gene Ontology' by conceptualisation of the scheme as a DAG; a simple tree-like hierarchy could not contend with such complexity.

Finally, many proteins are multi-functional: capable of performing a variety of biological roles, sometimes simultaneously (particularly with multidomain proteins). The biological role of a protein may also depend on its environment or localisation (Todd et al. 1999). In Table 2, we indicate the schemes that we have found empirically to include multiple functional assignments for gene products.

Current schemes

In this paper we have focused on six schemes mapped to the CS (EcoCyc, TIGR, SubtiList, MIPS, WIT and KEGG) and the 'Gene Ontology'. Two broad families of schemes emerged from our survey: (1) genome related schemes and (2) schemes related to the interaction networks of gene products.

The genome related schemes are EcoCyc, TIGR, SubtiList, MIPS and the 'Gene Ontology'. Two of them (EcoCyc and TIGR) are current implementations or derivations of Riley's original classification (Riley 1993). As a consequence, they can essentially be thought to represent the same scheme (implemented with trees of different breadth, depth and resolution). The SubtiList scheme was derived from an adapted combination of parts of the WIT related 'Metabolic Pathways Database (MPW) and of the Riley scheme (Moszer et al. 1996; Selkov et al. 1998). In addition, the scheme includes a number of functions specific to B. subtilis. In terms of their coverage of the CS, no doubt because of their relation to the Riley scheme, the three schemes are quite similar even though SubtiList appears to have a noticeably smaller coverage of the CS than the other two schemes. This is partly due to mapping limitations and partly because the SubtiList scheme was designed with the specific needs of the B. subtilis research community in mind, and therefore focuses on functional aspects of major relevance to them. The original Riley scheme was designed for the unicellular prokaryotic eubacteria E. coli, and this bias will exist in all derivative schemes. With schemes such as TIGR that are applied to diverse gene complements, such a bias could be problematic.

The MIPS scheme shares a lot of the Riley scheme but extends it to encompass a number of further functions. Some of these functions (e.g. signal transduction) exist in all organisms but are not explicitly listed in the Riley based schemes, whilst others are present to allow better coverage of eukaryotic functions by the scheme (e.g. organelle related functions). The MIPS scheme can, in essence, be thought of as a superset of Riley schemes which begins to address the issue of generating functional schemes applicable to multiple and diverse organisms.

As we have previously mentioned, genome sequencing initiatives are the main driving force in the development of functional classification schemes. Nevertheless, the vast majority of genome-sequencing initiatives have been focused on unicellular micro-organisms. Of the 24 complete genomes listed in NCBI's 'Entrez Genomes', only one, Caenorhabditis elegans, is multicellular. Both Riley's scheme and the MIPS scheme were designed for classification of the genomes of unicellular organisms. Therefore, there is a great paucity of functional nodes concerning the interaction between cells in many schemes.

WIT and KEGG are databases of gene product interactions. They deal with functions performed by the concerted actions of gene products in pathways and complexes. Both the WIT and the KEGG functional classification schemes generally classify the function of a gene product by association with a pathway or complex. This helps explain why both these classification schemes have good coverage of metabolism. At the time of data gathering, the KEGG scheme had only minimal coverage of non-metabolism related functions, but a recent visit to the KEGG homepage (http://www.genome.ad.jp/kegg) confirmed that the KEGG scheme is being extended to include a number of non-metabolism-related functions. The WIT scheme had good coverage of the CS dealing with transport, structure and information-related pathways in addition to metabolism. This association of gene products to pathways and complexes is very relevant to their function: all but the simplest of biological roles in cells are performed by interactions of gene products.

The 'Gene Ontology' is representative of the 'next generation' of functional classification schemes. Rather
than updating existing schemes, the 'Gene Ontology' has been designed from scratch and addresses many of the problems and issues we have discussed in this paper. The 'Gene Ontology' is multi-dimensional, separating the concepts of 'functional primitive', 'process' and 'localisation'. Its more complex architecture allows it to accommodate functional descriptions that are examples of more than one parent node. The scheme is being developed for classification of the gene complements of both unicellular and multi-cellular organisms. We did not attempt to map the 'Gene Ontology' to the CS, but we are confident that it would have occupied all of the CS nodes.

The most extensive scheme currently in use and applied to a significant number of genomes is the MIPS scheme. However, perhaps one of the most notable conclusions is that all genome related schemes (other than the 'Gene Ontology') cover broadly the same set of functions and there is little to make one scheme overwhelmingly superior to another. The schemes are tantali­ singly similar but unfortunately different enough to make direct comparisons between them difficult. With respect to the pathway and metabolism schemes, WIT has the most extensive functional classification but KEGG has built more generalised pathways that may be more accessible to many researchers. Certainly, consideration of gene-product interactions in pathways and complexes will play an important part in any future functional classification scheme.

The recent application of the 'Gene Ontology' for annotation of gene products identified in the Adh region of the *D. melanogaster* genome (Ashburner et al. 1999) illustrates the potential of such a scheme.

Future schemes

Functional classification schemes will become an increasingly critical element of genome databases. We believe that future schemes should have a controlled vocabulary and be integrated within an ontology which will not only classify functional nodes but control their grammar and semantics (Baker et al. 1999; Schulze-Kremer 1998). Ideally, they should be applicable to all species but still be capable of accommodating very specific functions and allow cross-species functional correspondence where possible. They will also have to be able to contend with environment and location dependent changes in the function of gene products. This will only be possible if multiple functional assignments for gene products are permitted. Furthermore, the most effective functional classification schemes will be multi-dimensional which will allow for accurate positioning of gene products in the function space. To deal with these multiple parameters, schemes will undoubtedly need to explore more complex structures than simple trees.

The increasing availability of multi-cellular genomes demands the development of more complete schemes that will have to classify not only the functions related to intercellular communication but also those related to the more complex organisation of multi-cellular organisms (e.g. tissues and organs).

As the quantity of information on gene products increases at an unparalleled pace, it is imperative that the quality of functional annotation improves. The 'Gene Ontology' represents a promising development in this area. On the other hand, its very complexity and scope may be an obstacle to its widespread implementation. There is perhaps a need for a less extensive scheme, spanning the gap between simple, tree-like classification schemes and the 'Gene Ontology', or perhaps we should consider having both highly detailed and simplified functional schemes catering to different sets of users (Gilbert 1998).

Whichever schemes are developed, there is certainly a need to try and use a single, standardised format or to provide tools to map between the different schemes. The logical progression is towards functional classifications that are applicable to all organisms and cover all functional dimensions. The power of such schemes will only be realised when applied consistently over multiple genomes to allow comparison of organisms. Functional genomics will then be ready for its golden age.

Acknowledgements We thank Dr. Neil Stoker for helpful discussions. S.R. is supported by GlaxoWellcome studentship. This is a publication from the Bloomsbury BBSRC structural biology centre.

References


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

Sarah A. Teichmann*, Stuart C. G. Rison†, Janet M. Thornton†,2
Monica Riley†, Julian Gough† and Cyrus Chothia†

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

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

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

4MRC Laboratory of Molecular Biology, Hills Road, Cambridge
CB2 2QY, UK

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. 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.

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 Horowitz 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.1 Jensen 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...
active-site residues produce new catalytic properties for enzymes and, hence, the formation of new pathways.\(^{1-11}\)

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 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. 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. 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 defined on the basis of structural information and which belong to one of 202 families. The 27 domains clustered by 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 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.

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 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

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

B. Number and size of protein families

<table>
<thead>
<tr>
<th>Family size (n)</th>
<th>Number of families of size n</th>
<th>Family size (n)</th>
<th>Number of families of size n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>33</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
Evolution of Metabolic Pathways

Table 3. Domain combinations

<table>
<thead>
<tr>
<th>A. The number of different families from which partners are drawn</th>
<th>Number of families whose members can be linked to n different families</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number n of different families from which partners are drawn</td>
<td>Number of families</td>
</tr>
<tr>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Adjacent regions of unknown domain type</td>
<td>15</td>
</tr>
<tr>
<td>Total no. families</td>
<td>213</td>
</tr>
</tbody>
</table>

B. Protein families that have many partner families

<table>
<thead>
<tr>
<th>Protein family</th>
<th>Family size</th>
<th>Number of partner families</th>
<th>Families N or C-terminal to the protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NAD(P)-binding Rossmann domain</td>
<td>51</td>
<td>12</td>
<td>N7 C5</td>
</tr>
<tr>
<td>2. Glutathione synthetase ATP-binding domain</td>
<td>11</td>
<td>6</td>
<td>N4 C2</td>
</tr>
<tr>
<td>3. Thiamine diphosphate-binding fold (THDP-binding)</td>
<td>21</td>
<td>4</td>
<td>N1 C2 N + C:1</td>
</tr>
<tr>
<td>4. Regulatory domain in the amino acid metabolism</td>
<td>10</td>
<td>4</td>
<td>N2 C2</td>
</tr>
<tr>
<td>5. Class I glutamine amidotransferases (GAT)</td>
<td>8</td>
<td>4</td>
<td>N2 C2</td>
</tr>
<tr>
<td>6. β-Galactosidase/glucuronidase domain</td>
<td>3</td>
<td>3</td>
<td>N1 C1 N + C:1</td>
</tr>
<tr>
<td>7. FAD/NAP(P)-binding domain</td>
<td>19</td>
<td>3</td>
<td>N3</td>
</tr>
<tr>
<td>8. Copper amine oxidase, domains 1 and 2</td>
<td>2</td>
<td>3</td>
<td>N1 C2</td>
</tr>
<tr>
<td>9. Fe-4S ferredoxins</td>
<td>5</td>
<td>3</td>
<td>N2 C1</td>
</tr>
<tr>
<td>10. Glycosyltransferases</td>
<td>7</td>
<td>3</td>
<td>N1 C1 N + C:1</td>
</tr>
<tr>
<td>11. N-terminal nucleophile amidohydrolases</td>
<td>4</td>
<td>3</td>
<td>N3</td>
</tr>
<tr>
<td>12. Cobalamin (vitamin B12)-binding domain</td>
<td>2</td>
<td>3</td>
<td>N1 C2</td>
</tr>
</tbody>
</table>

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 multiple use of a particular member. Thus, a protein that functions in n pathways can be seen as 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

<table>
<thead>
<tr>
<th>Number of pathways (n)</th>
<th>Number of proteins active in n pathways</th>
<th>Number of these proteins with assigned domains</th>
<th>Number of domains in the matched proteins</th>
<th>Virtual homologs: number of domains in the other (n − 1) different pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>96</td>
<td>81</td>
<td>131</td>
<td>131</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>30</td>
<td>43</td>
<td>86</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12</td>
<td>23</td>
<td>69</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>131</td>
<td>201</td>
<td>304</td>
</tr>
</tbody>
</table>
Evolution of Metabolic Pathways

Family Size against Number of Pathways

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 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 Commission (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. 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. 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 SMMP 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. Triplet 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 (α/β)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 cofac-
tor or minor substrate-binding site and eight con­
serve a catalytic mechanism. The other five con­
serve 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
(αβ)6 barrels trpC and trpA in tryptophan biosynthesis, and hisA and hisF in histidine biosynthesis, and hisA and hisF in histidine biosynthesis. trpC is a bifunctional enzyme consisting of two (αβ)6 barrels, one of which is N-(3′-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 neither "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 phosphoenolpyru-
vate carboxylase, ppc, in an EC class 4 reaction.
These "parallel" enzymes belong to the same family of (αβ)6 barrels. There are three phosphori-
boyltransferase 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 phosphori-
boylpyrophosphate synthase (prsA). In deoxyopyr-
imidine nucleotide/nucleoside metabolism, dCTP
deaminase (dcd) is followed by the related dUTP
pyrophosphatase (dut). Finally, there are two
members of the inosine monophosphate dehydro-
genase (αβ)6 barrel family in nucleotide metabo-
lism: IMP dehydrogenase (guaD) 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 of many domains is shared. An example of
this type of scenario is given in Figure 3: furA 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 sub-
strates but produce different products (Figure 2(b)).
The two enzymes can either carry out related reac-
tions 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>
<thead>
<tr>
<th>Superfamily and pathway</th>
<th>Enzymes</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphoenolpyruvate/pyruvate (αβ)6 barrels in fermentation</td>
<td>pykF/pykA, ppc</td>
<td>The pykA/pykF and ppc both have phosphoenolpyruvate as their substrate and belong to different EC classes.</td>
</tr>
<tr>
<td>Ribulose-phosphate binding (αβ)6 barrels in tryptophan biosynthesis*</td>
<td>trpA, trpC</td>
<td>Consecutive enzymes in different EC classes.</td>
</tr>
<tr>
<td>P-binding α/β barrels in histidine, purine and pyrimidine biosynthesis</td>
<td>hisA, hisF</td>
<td>Consecutive enzymes in different EC classes.</td>
</tr>
<tr>
<td>Phosphoribosyltransferases (PRTases) in histidine, purine and pyrimidine biosynthesis</td>
<td>prsA, purF and prsA, pyrE</td>
<td>Consecutive pairs of enzymes in EC class 2 (transferases).</td>
</tr>
<tr>
<td>dUTPase domains in deoxyxypirimidine nucleotide/nucleoside metabolism</td>
<td>dcd, dut</td>
<td>Consecutive enzymes in EC class 3 (hydrolases).</td>
</tr>
<tr>
<td>Inosine monophosphate dehydrogenase (αβ)6 barrels in nucleotide metabolism</td>
<td>guaB, guaC</td>
<td>Enzymes one step apart in EC class 1 (oxidoreductases).</td>
</tr>
</tbody>
</table>

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.

* 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.
Evolution of Metabolic Pathways

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 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

<table>
<thead>
<tr>
<th>Homologous pair</th>
<th>Subsequent enzyme(s)</th>
<th>Homologous domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different substrates, similar reactions, same product</td>
<td>malP, glgP</td>
<td>malZ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aldA/aldB or fucO</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pfkA/pfkB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ppsA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Homologous pair</th>
<th>Subsequent enzyme(s)</th>
<th>Homologous domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same substrates, different cofactors, similar reactions, same product</td>
<td>maeA, maeB</td>
<td>ppsA*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class II aldolase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAD(P)-binding Rossmann fold domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amino acid dehydrogenase domain and NAD(P)-binding Rossmann domain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Homologous pair</th>
<th>Subsequent enzyme(s)</th>
<th>Homologous domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different substrates, similar reactions, different products: subsequent enzyme uses several substrates simultaneously or has multiple substrate specificity</td>
<td>thiD, thiM</td>
<td>thiE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ndk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>serC</td>
</tr>
</tbody>
</table>

B. Homologous pairs that act on the same substrate

<table>
<thead>
<tr>
<th>Preceding enzyme(s)</th>
<th>Homologous pair</th>
<th>Homologous domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same substrate, similar reactions</td>
<td>araC</td>
<td>trpDE, pabAB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trpD/pabB: class I glutamine amidotransferase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>glcB/aceB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carAB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>serA, pdxB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>prsA*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>purF, pyrE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>purU, pykF, ppc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uxaC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>deoD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>thaD</td>
</tr>
</tbody>
</table>

* 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 superfamilies in SCOP) with two members, cts and pssA, which are transferases of a similar type. pssA and cts 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, lipX and lipX, 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-acetylglicosamine-1-phosphate uridylytransferase activity. A further sequence family with conservation of chemistry encompasses lipX, dktA and murG, which are all transferases with similar
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.

**Polysoprenoid 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

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

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.

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).
maIS and maIZ both have a glycosyltransferase domain, and galP and malP are both phosphorylases that have a β-glucosyltransferase and glycogen phosphorylase domain, as shown in Figure 4.

Glucoseoxygenesis. In glucoseoxygenesis, 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; 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 rhhamnose catabolism (Figure 3) are each one gene apart on the E. coli chromosome, and are homologous and...
Evolution of Metabolic Pathways

(a) Leucine Biosynthesis

1. **isopropylmalate isomerase**
   - EC 4.2.1.33

2. **acnA: aconitase**
   - EC 4.2.1.3

3. **3-isopropylmalate dehydrogenase**
   - EC 1.1.1.85

(b) Carnitine metabolism, CoA-linked

1. **carnitine-CoA ligase**
   - EC 6.3.2.-

2. **O-succinylbenzoyl-CoA ligase**
   - EC 6.2.1.26

3. **carnitine racemase**
   - EC 5.-.-.-

4. **naphthoate synthase**
   - EC 4.1.3.36

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 biosynthesis 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, who also point out the pathway duplication of the proteins in the prep operon and proteins in the glyoxylate shunt. The proteins in the prep 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 ubiquit-none 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 fusO 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 identified genes in E. coli that are potential candidates for horizontal transfer within the last 10⁸ 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 (galF, 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.

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. in the tryptophan biosynthesis pathway. Jensen 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. and for a pathway that degrades a xenobiotic pesticide by Copley. On the basis of the distribution of 38 (aB)₆ homologous barrel structures in central metabolism, Copley & Bork 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., Huynen et al. and Makarova et al. 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 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.

The details of the iterative HMM procedure SAM-T99, is described by Gough et al. 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. and the parameters used to construct this library of models were calibrated by Gough et al. 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. 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

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

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

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

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

© 2002 Elsevier Science Ltd. All rights reserved

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

---

**Introduction**

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

A number of theories have been advanced to explain the evolution of enzyme-catalysed metabolic networks from the "primordial soup". As early as 1945, Horowitz proposed the retrograde model of pathway evolution, followed by Ycas and Jensen, who suggested a patchwork model. In the retrograde evolution model, pathways evolve "backwards" from a key metabolite. The model presumes the existence of a chemical environment where both key metabolites and potential intermediates are available; each time the environmental supplies of a key metabolite are used up, the organism recruits an enzyme capable of transforming an intermediate into this key metabolite. Every time the environmental reserves of the intermediate drop prohibitively, an enzyme is similarly recruited to catalyse the transformation of another metabolite into the intermediate, etc. In 1965, Horowitz restated his theory to take into account the discovery of operons. At the time, the clustering of genes involved in known pathways into operons (e.g. leucine and tryptophan) along with a consideration of the probable origin of operons led him to suggest that operons would cluster genes with overlapping specificities, suggesting structural homology and common ancestry; enzymes within a pathway would tend to be recruited by duplications "within" a pathway. In its strictest form, however, "the stepwise backwards route does not demand that the enzymes are evolutionarily related".

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

Therefore, within this large network of possible interactions (including spontaneous non-enzymatic reactions), many paths, some synthesising key metabolites, may have existed, albeit at a very low level. Duplication of genes in such key-metabolite synthesising paths, followed by their specialisation, would account for extant pathways. Furthermore, fortuitous evolution of a novel enzyme-catalysed chemistry could bring into play environmental substrates previously unavailable to the metabolic network. This novel intermediate may then become a new precursor to a key metabolite, even if it is several enzymatic steps away. Retrograde evolution is generally thought to suppose a "substrate-driven" evolution as, for neighbouring enzymes in a pathway, the product of one enzyme will be the substrate for the next. By contrast, patchwork evolution is thought to be "chemistry-driven", by recruitment and specialisation of broad substrate specificity enzymes capable of performing the required catalysis.

The structural and evolutionary anatomy of SMM pathways

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

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

Pathway distance

Here, we make use of a measure known as the pathway distance: the number of distinct metabolic steps separating two enzymes (see Figure 1). By metabolic step, we mean an enzyme-catalysed modification of one or more substrates into chemically distinct compounds. This concept is similar to that of reaction frames found in the EcoCyc database from which we extract our information on the E. coli metabolic pathways. EcoCyc reaction frames are computational objects encapsulating an enzyme-catalysed substrate modification. The frame contains the reactants and products of the modification, and is associated with one or more pathways. The reaction frame is associated with one or more enzymes using a linking object. Conceptually, our metabolic steps are the product of this linking, i.e. the enzyme(s) catalysing the transition from reactant(s) to product(s). Indeed, in most cases, the number of "our" metabolic steps and the number of reaction frames in an
Figure 1. Pathway distance illustrated in glycolysis. The pathway shown is glycolysis as represented in the EcoCyc database. Each enzyme-catalysed reaction (blue arrows) represents a metabolic step and therefore a unit of pathway distance. Enzymes catalysing the steps are shown in pink; major metabolites are listed in red; co-factors and minor metabolites are in black.

For the enzymes pgi and pykF (located at the beginning and end of the pathway), the longest route (via tpiA) has a pathway distance of eight steps (traversing the metabolic steps catalysed by: (i) pfkA and pfkB, (ii) fbaA and fbaB, (iii) tpiA, (iv) epd and gapA, (v) pgk, (vi) pgm1, pgmA and pgmB, (vii) eno, and (viii) pykF). The minimal pathway distance between pgi and pykF is seven steps (avoiding tpiA). Pathway distances relative to pgi are in green; distances from glyceraldehyde-3-phosphate onwards show the minimal and maximal distances. pgi, phosphoglucose isomerase; pfkA and pfkB, 6-phosphofructokinase-1 and -2; fbaB and fbaA, fructose bisphosphate aldolase class I and II; tpiA, triose phosphate isomerase; epd, glyceraldehyde-3-phosphate dehydrogenase 2; gapA, glyceraldehyde-3-phosphate dehydrogenase-A complex; pgk, phosphoglycerate kinase; pgm1 and pgmB, phosphoglycerate mutase 1 and 2; pgm1, phosphoglycerate mutase (co-factor-independent); eno, enolase; pykF and pykA, pyruvate kinase 1 and II.

EcoCyc pathway are identical and differ only when we merge two EcoCyc reaction frames into one (see Methods). Pathway distance has been used independently in a recent work, where it is called "metabolic distance."^4 Using our measure of pathway distance, adjacent enzymes therefore have a pathway distance of 1. By extension, enzymes catalysing the same metabolic step in a pathway (for example, in Figure 1, pfkA and pfkB) can be thought of as having a pathway distance of 0.

Gene and context

The analysis of aspects of the genome other than the predicted amino acid sequence of the proteins encoded by the genes has been described as the "context of a gene."^7 Here, we consider three contexts: the genome (i.e. the relative location of SMM enzyme genes on the E. coli chromosome); metabolism (i.e. the relative location of enzymes within the SMM network); and the evolutionary context.

Much work has already been done regarding the spatial organisation of genes in bacteria. Tamames et al.,^8 considering Haemophilus influenzae and E. coli, observed that functionally related genes (as classified within simplified scheme derived from GenProtEC's gene classification scheme^9,10) were neighbours more often than functionally unrelated genes. A strong correlation between genomic clustering and function was detected when considering a large number of genomes,^11 in particular, for genes in close proximity not just in one, but in many genomes. Recently, the concepts presented by Overbeek et al.,^12 were generalised and implemented within a function prediction algorithm that connects genes likely to share functional similarity (in particular, involvement in common metabolic pathways) by analysing orthology and genomic localisation of genes.13 Such correlations are, however, strongly dependent on phylogenetic distance.14,15,16

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

<table>
<thead>
<tr>
<th>A. Pathways</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. original EcoCyc pathways considered</td>
<td>102</td>
</tr>
<tr>
<td>No. final pathways</td>
<td>82</td>
</tr>
<tr>
<td>No. 82 final pathways composed of two or more of the original 102 EcoCyc pathways</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Reaction frames</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. frames in 82 analysed pathways</td>
<td>619</td>
</tr>
<tr>
<td>No. distinct frames</td>
<td>581</td>
</tr>
<tr>
<td>No. (%) of these 581 frames used more than once</td>
<td>33 (5.68)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. Gene assignments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. (%) 581 reaction frames with no known genes</td>
<td>59 (10.15)</td>
</tr>
<tr>
<td>No. genes in 619 analysed reaction frames</td>
<td>776</td>
</tr>
<tr>
<td>No. distinct genes</td>
<td>594</td>
</tr>
<tr>
<td>No. (%) of these 594 distinct genes used more than once in the pathways</td>
<td>79 (13.30)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D. GenBank identifiers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. (%) 594 distinct genes assigned a GenBank PID</td>
<td>586 (98.75)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E. Structural (CATH) and sequence domain families</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. CATH domain families</td>
<td>220</td>
</tr>
<tr>
<td>No. sequence families pre-linkage to structural families</td>
<td>137</td>
</tr>
<tr>
<td>No. sequence families post-linkage to structural families</td>
<td>117</td>
</tr>
<tr>
<td>No. families</td>
<td>337</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F. Domain family assignments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. (%) 586 genes with known PID assigned to one or more CATH domain families</td>
<td>382 (65.19)</td>
</tr>
<tr>
<td>No. (%) 586 genes with known PID assigned to one or more sequence domain families only</td>
<td>98 (16.72)</td>
</tr>
<tr>
<td>No. (%) 586 genes with known PID assigned to one or more CATH and/or sequence domain families</td>
<td>480 (81.91)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>G. Genomic locations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. (%) 594 distinct genes with an identifiable chromosomal location</td>
<td>584 (98.32)</td>
</tr>
</tbody>
</table>

Analysing pathways

SMM pathways have been analysed before.\textsuperscript{1,11,12,16-26} Here, we extend our previous work\textsuperscript{1,13} by analysing patterns of domain distribution and recruitment within the E. coli SMM. We gain deeper insight by exploring a trinity of contexts (evolutionary relationships of genes, genomic location of genes and metabolic environment of enzymes) rather than considering each pathway as a “bag of genes”. Furthermore, as much as possible, we analyse the SMM as a single network rather than a collection of arbitrarily defined pathways. Such a “stepwise” analysis allows us to detect hitherto unobserved patterns of recruitment as well as clarify the metabolic range of SMM gene clustering.

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

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

Results

Small molecule metabolism pathways

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

Many of the pathways described separately in EcoCyc possess a high level of overlap, i.e. stretches of the same reaction frames are found in both. In our previous work,\textsuperscript{13} genes found in reaction frames reused in different EcoCyc pathways were identified as virtual homologues. These virtual homologues reflect the “inter-connectedness” of the pathways, which can be considered more realistically as a network. Here, since we wanted to consider the whole network rather than traditionally defined separate pathways, we dealt with such duplication by merging pathways with three or more reaction frames in common. We began with 102 pathways, composed of 738 reaction frames; following iterative merges, our final dataset contained 82 pathways composed of 619 reaction frames. Of the original 102 EcoCyc pathways, 68 were left unchanged by the merging procedure, one was deleted (as it was found to be represented entirely in other pathways) and the
remaining 33 were merged into 14 pathways, accounting for the 82 (68 + 14) final pathways. For example, our largest merged pathway was formed when merging the EcoCyc pathways GLYCOLYSIS/TCAC/GLYOX-BYPASS, GLYCOLYSIS/E-D, ANAERES1- PWY, FERMENTATION-PWY, GLUCONEO-PWY and GLYCOL-GLYOX-BYPASS (respectively, glycolysis/tricarboxylic acid cycle/glyoxylate bypass, glycolysis and Entner-Doudoroff, anaerobic respiration, fermentation, gluconeogenesis and glycol metabolism and degradation). The total number of frames in the six individual pathways was 81, the final number in the merged pathway was 42, illustrating the large overlap between individual pathways.

We briefly compared our final pathways to the 89 SMM pathways that we identified in the metabolic pathway section of the KEGG database.7 The 14 pathways created from the merger of two or more of the original EcoCyc pathways tended to be similar in size, and sometimes substantially larger, than their KEGG equivalent; the majority of the other pathways were smaller than their KEGG equivalent. However, the KEGG pathways are composite pathways, combining reactions occurring in a number of different organisms into one representation.7 When considering only the portions of the KEGG pathways predicted by KEGG curators to occur in E. coli, all our pathways appeared to be of a similar or larger size than their KEGG equivalent.

Our merging procedure ensured that no two pathways in our final dataset shared three or more consecutive reaction frames. It is worth noting that even if we had merged our pathways to completion (i.e. until no two distinct pathways in the final set shared a reaction frame) we would not have ended up with a single network representing all of E. coli SMM. This is because certain pathways are connected only by a common metabolite frame, rather than a reaction frame, and we considered only connectivity between reaction frames. A measure of the remaining level of overlap is that 33 of the 581 distinct reaction frames in the 82 final pathways are found in more than one pathway. Nevertheless, the merges represent a transition from the traditional representation of SMM as distinct pathways towards a network representation.5,23

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

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

### Structural annotation, sequence families and evolutionary relatedness

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

We used the Gene3D database44 to assign 382 (65.1%) of the 586 E. coli SMM genes considered to one or more of 220 structural (CATH) families. Also, to find all possible evolutionary relationships, we used sequence comparison methods to analyse whole genes and gene regions of greater than 75 residues for which no structural assignments could be made. Using PSI-BLAST28 and DIVCLUS,26 as described in Methods, we identified a further 117 sequence domain families. An additional 98 enzymes with no structural assignments were classified into a sequence family, bringing the total number of enzymes assigned to at least one structural or sequence family to 480 (82%) (see Table 1).

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

---

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

‡ http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/
Homology and pathway distance

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

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

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

<table>
<thead>
<tr>
<th>Pathway distance</th>
<th>No. pairs</th>
<th>No. homologous pairs</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>660</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>491</td>
<td>16</td>
<td>5.0 × 10⁻³</td>
</tr>
<tr>
<td>3</td>
<td>425</td>
<td>11</td>
<td>6.2 × 10⁻¹</td>
</tr>
<tr>
<td>4</td>
<td>367</td>
<td>9</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>344</td>
<td>8</td>
<td>0.23</td>
</tr>
<tr>
<td>6</td>
<td>284</td>
<td>3</td>
<td>0.71</td>
</tr>
<tr>
<td>7</td>
<td>287</td>
<td>3</td>
<td>0.72</td>
</tr>
<tr>
<td>8</td>
<td>294</td>
<td>3</td>
<td>0.74</td>
</tr>
<tr>
<td>9</td>
<td>254</td>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>176</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>11</td>
<td>129</td>
<td>2</td>
<td>0.37</td>
</tr>
</tbody>
</table>

For each pathway distance analysed, the number of pairs considered is listed (pairs were considered only if at least one domain assignment was available for each enzyme) and the number of homologous pairs is given. Statistically significant p-values (cut-off: 0.075) are in bold; the homology percentage observed at these distances is not the consequence of a chance distribution of enzymes.
chemistry is often linked to the co-factors used,30 so not all instances of recruitment were classified. Furthermore, the sequence domain recruitments were not classified. The most common explanation for domain recruitment is conservation of the catalytic mechanism. This accounts for 39 of the 113 instances of domain duplication (34.5%). Conservation of co-factor binding comes second, accounting for 35 (31%) of the cases. The least common apparent cause of domain duplication is conservation of substrate binding, occurring in six instances (5.3%). We were unable to classify the remaining 33 domain duplications unambiguously (See Table 3).

### Homology and gene intervals

Similarly to pathway distance, we considered the relationship between gene intervals and gene homology. We can assign a genome position for 584 (98%) of the 594 distinct genes present in our pathways. Therefore, for the majority of enzyme pairs in our pathways, we can derive a gene interval (i.e. the number of genes on the *E. coli* genome separating the two genes encoding the enzymes in the pair). This generates a discrete distribution of gene intervals. A total of 4405 *E. coli* gene pairs, for which both genes had at least one identifiable genomic location, therefore, the largest gene interval possible is 2202 (since we consider only the minimal gene interval on the circular chromosome).

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

### Gene intervals and pathway distance

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

The scatter plot in Figure 4 shows no obvious pattern but binning revealed some trends. At each pathway distance, we grouped the enzyme pairs into gene-interval bins as described above. The relative contributions of the first three bins at various pathway distances is illustrated in Figure 5. There is an evident correlation between pathway distance and the proportion of genes with low gene intervals (zero to five genes) (see Figure 5(a)). To determine to what extent this correlation was due to clustering of metabolic genes into operon transcriptional units, we obtained a list of *E. coli* transcriptional units from the RegulonDB database36 and used it to flag *E. coli* genes known or predicted to be part of operon structures. Figure 5 shows (b) the subset of all pairs in which both genes are in an operon and (c) where both genes are predicted not to be part of an operon. In the former case, the trend observed in Figure 5(a) is more marked, whilst in the latter case it disappears.

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

### Homology within reaction frames

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

---

**Table 3. Domain conservation explanations**

<table>
<thead>
<tr>
<th>Domain conservation explanation</th>
<th>No. (%) instances</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry conserved</td>
<td>39 (34.5)</td>
<td>MetB and MetC (PLP-dependent aspartate aminotransferase like domain)³⁹</td>
</tr>
<tr>
<td>Co-factor/minor substrate</td>
<td>35 (31.0)</td>
<td>PurD and PurT (ATP-grasp fold)³⁷</td>
</tr>
<tr>
<td>Substrate binding conserved</td>
<td>6 (5.3)</td>
<td>TrpA and TrpC (TIM barrel)³⁹</td>
</tr>
<tr>
<td>Unclassified</td>
<td>33 (29.2)</td>
<td></td>
</tr>
</tbody>
</table>

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

---

³⁶ http://www.genome.wisc.edu/pub/analysis/m5forfs.txt
Of the 339 possible pairs of proteins co-located within a reaction frame (e.g. a reaction frame containing enzymes A, B and C would have possible pairs A–B, A–C and B–C), 66 (19.5%) were complete isozymes (e.g. the aconitases AcnA and AcnB) and 29 (8.5%) were partial isozymes (e.g. the aspartate kinase LysC and the homologous bifunctional MetL and ThrA aspartate kinase/homoserine dehydrogenases, which have only the aspartokinase catalytic domain in common). For 244 (72%) of the protein pairs within a frame, no homology was detected.

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

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

Figure 4. Gene interval and pathway distance. Gene intervals are plotted against pathway distance for the 3495 gene pairs where both these measures are available.
Figure 5. Pathway distance and gene intervals. At each pathway distance (x-axis), the percentage of enzyme pairs with a gene interval of zero to five genes (blue diamonds), six to 50 genes (pink square) and 51 to 500 genes (yellow triangle) is plotted for (a) all pairs, (b) pairs with both enzymes predicted to be in operons and (c) pairs with both enzymes predicted to be out of operons (operon prediction from Saldago et al. [10]).
by virtue of having the same domain make-up but protein B might be 50 or more residues longer than protein A, suggesting an unidentified additional domain in protein B.

The 95 (complete and partial) isozyme pairs cluster into 59 sets of homologous proteins (with five instances of reaction frames containing more than one set of distinct isozymes). Where possible, we assigned one or more rationales for the isozymes, identifying nine such reasons. The nine reasons are listed in Table 4 and the 59 sets of isozymes we identified are described in detail in Table 5. To illustrate the scenarios presented above, we look at a couple of examples selected from Table 5. Nitrase reductases NarG, NarZ and NapA are all homologous. NarG and NarZ are complete isozymes, having the same domain make-up, whilst NapA is a partial homologue to both NarG and NarZ, since NapA contains two domains not detected in NarG or NarZ. The homology for this set of proteins is therefore described as C[NarG, NarZ]/P in Table 5. NuoM, NuoN and NuoL are all subunits of NADH dehydrogenase 1. They have an identical domain make-up but NuoM is 509 residues long, NuoN is 425 residues long and NuoL is 613 residues long. The homology for this set of proteins is therefore described as C[104, 188] in Table 5; all members of the set are homologues and differences in size, relative to the largest protein (here NuoL) are given.
Most commonly (13 cases), the isozymes had different preferred substrates or minor substrates. For example, AnsA and AnsB both catalyse transamination of aspartate to asparagine but AnsA uses NH₃ as the amine “donor” whilst AnsB uses glutamine; FabA and FabZ have different length preferred fatty-acid chain substrates. The isozymes were often active during “different conditions” (11 instances); for example, the fumarases FumA and FumB are active during aerobic and anaerobic, respectively. In nine cases, the isozymes have different roles, commonly one isozyme was catabolic and the other biosynthetic (e.g. Air and DadX). In seven cases, the isozymes were part of the same enzymatic complex or constituents of separate (but functionally related) complexes. These cases were difficult to explain unambiguously, although between homologous complexes, homologous polypeptides often performed similar roles (e.g. see the formate dehydrogenases). Different regulation accounted for seven sets of isozymes. For example, the aldolases AroF, AroG and AroH are all subject to different feedback control. We observed different kinetics (six sets), alternative cellular localisations (three sets), different co-activity (two sets) and different heterogeneous groups (one set).

### Inline reuse

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

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

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

The case of LpdA warrants special attention; this is a dihydrolipoyl dehydrogenase and is a subunit in both the pyruvate and α-ketoglutarate dehydrogenase complexes. Whilst the overall chemistries performed by these complexes is different, in both cases, the dihydrolipoyl dehydrogenase subunit re-oxidises dihydrolipoamide, a co-factor used in the reactions catalysed by the other subunits of the complexes. In a sense, LpdA therefore acts independently from the rest of the complex and performs the same chemistry on the same substrate at each point of recruitment."
<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. accA, accD</td>
<td>C</td>
<td>Complex</td>
<td>Acetyl-CoA carboxyltransferase</td>
<td>AccA and AccD form the a-2-β-2 complex of acetyl CoA carboxyltransferase</td>
</tr>
<tr>
<td>1. entE, entF</td>
<td>P</td>
<td>Complex</td>
<td>Enterobactin synthase complex</td>
<td>EntE and EntF are part of the enterobactin synthase complex and have similar, but distinct, substrates</td>
</tr>
<tr>
<td>1. hcaC, hcaE</td>
<td>P</td>
<td>Complex</td>
<td>Dioxynase</td>
<td>HcaE forms the large α-subunit of 3-phenylpropionate dioxygenase. HcaC is a ferredoxin</td>
</tr>
<tr>
<td>1. nuoG, nuoL</td>
<td>C</td>
<td>nuoL, nuoF/P; C[104,188]</td>
<td>Complex</td>
<td>NADH dehydrogenase</td>
</tr>
<tr>
<td>1. fdnG, fdoG, fdhP; 2. fdoH, fdnH, 3. fdnl, fdol</td>
<td>C</td>
<td>fdnG, fdoG/P; C; C</td>
<td>Complex conditions</td>
<td>Formate dehydrogenases</td>
</tr>
<tr>
<td>1. hyaB, hyfG, hycE, hybC; 2. hyfA, hyfH, hyCF, hyCB; 3. hyaA, hyfL, hyCG, hyb0; 4. hyfB, hyfD, hyFP</td>
<td>C</td>
<td>hyfG, hycE/I/P; C</td>
<td>hyfA, hyfB; C</td>
<td>hyfH, hycF/I/P; C</td>
</tr>
<tr>
<td>1. narG, narZ, napA; 2. narH, narY; 3. narL, narV</td>
<td>C</td>
<td>narG, narZI/P; C; C</td>
<td>Complex localisation</td>
<td>Nitrate reductases</td>
</tr>
<tr>
<td>1. aceB, glcB</td>
<td>C[190]</td>
<td>Conditions</td>
<td>Malate synthases</td>
<td>GdcB is most active in cells grown on glyoxylate. AceB is active in the glyoxylate bypass</td>
</tr>
<tr>
<td>1. acnA, acnB</td>
<td>C</td>
<td>Conditions</td>
<td>Aconitases</td>
<td>AcnB is mainly catabolic, AcnA is a stabler maintenance enzyme</td>
</tr>
<tr>
<td>1. fumA, fumB</td>
<td>C</td>
<td>Conditions</td>
<td>Fumarases</td>
<td>FumA is aerobic, FumB is anaerobic</td>
</tr>
<tr>
<td>1. glpA, glpD</td>
<td>P</td>
<td>Conditions</td>
<td>Glycerol-3-phosphate dehydrogenases</td>
<td>GlpA forms part of the GlpAB catalytic dimer of glycerol-3-phosphate dehydrogenase. GlpO is aerobic, GlpAB is anaerobic</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. speC, speF</td>
<td>C</td>
<td>Conditions different roles regulation</td>
<td>Ornithine decarboxylases</td>
<td>SpeF is degradative and inducible, especially at low environmental pH. SpeC is biosynthetic and constitutively expressed.</td>
</tr>
<tr>
<td>1. sodA, sodB</td>
<td>C</td>
<td>Conditions heterogenous group</td>
<td>Superoxide dismutases</td>
<td>SodA complexes with manganese (Mn) and is aerobic. SodB complexes with iron (Fe) and is both aerobic and anaerobic.</td>
</tr>
<tr>
<td>1. aroK, aroL</td>
<td>C(66)</td>
<td>Conditions kinetics</td>
<td>Shikimate kinases</td>
<td>AroK has a higher Km than aroL. The enzymes are differently repressed by tyrosine and tryptophan.</td>
</tr>
<tr>
<td>1. treA, treF</td>
<td>C</td>
<td>Conditions localisation</td>
<td>Trehalases</td>
<td>TreA is periplasmic, TreF is cytoplasmic. TreA is active under conditions of high osmolality.</td>
</tr>
<tr>
<td>1. cysK, cysM</td>
<td>C</td>
<td>Conditions substrate</td>
<td>Acetylserine lyases</td>
<td>CysK is acetylserine lyase A. CysM is acetylserine lyase B and can use thiosulphate instead of sulphide (H2S). CysM is required for efficient cysteine biosynthesis during anaerobic growth.</td>
</tr>
<tr>
<td>1. pheA, tyrA</td>
<td>P</td>
<td>Different co-activity</td>
<td>Chorismate mutases</td>
<td>PheA acts as both a chorismate mutase and phenate dehydratase whilst TyrA acts as a chorismate mutase and a phenate dehydrogenase. Both are succeeded by TyrB which turns the product of the former into l-phenylalanine and the latter into l-tyrosine.</td>
</tr>
<tr>
<td>1. relA, spoT</td>
<td>C</td>
<td>Different co-activity</td>
<td>ppGpp synthases</td>
<td>RelA is a ppGpp synthase and a GTP pyrophosphokinase. SpoT is a ppGpp synthase and a ppGpp pyrophosphohydrolase.</td>
</tr>
<tr>
<td>1. tdcB, ilvA</td>
<td>C(185)</td>
<td>Different roles</td>
<td>Threonine dehydratases</td>
<td>IbVa is biosynthetic, TdcB is catabolic.</td>
</tr>
<tr>
<td>1. entC, menF</td>
<td>C</td>
<td>Different roles kinetics</td>
<td>Isochorismate synthases</td>
<td>EntC is the enterobactin synthesis-specific isochorismate synthase and catalyses a reversible reaction. MenF is the menaquinone synthesis-specific isochorismate synthase and catalyses an irreversible reaction.</td>
</tr>
<tr>
<td>1. alr, dadX</td>
<td>C</td>
<td>Different roles regulation</td>
<td>Alanine racemases</td>
<td>DadX (catabolic) is induced; Alr (biosynthetic) is constitutive.</td>
</tr>
<tr>
<td>1. sdaA, sdaB, sdhY</td>
<td>C(sdaA, sdaB)/P</td>
<td>Different roles regulation substrate</td>
<td>l-Serine/l-threonine deaminases</td>
<td>SdaA and SdaB are l-serine and l-threonine deaminases; SdhY is only an l-threonine deaminase.</td>
</tr>
<tr>
<td>1. argD, astC</td>
<td>C</td>
<td>Different roles substrate</td>
<td>Transaminases</td>
<td>AstC (catabolic) has a higher affinity for succinylornithine than for acetylornithine. ArgD is anabolic.</td>
</tr>
<tr>
<td>1. epd, gapA</td>
<td>C</td>
<td>Different roles substrate</td>
<td>Dehydrogenases</td>
<td>GapA is the effective glyceraldehyde-phosphate dehydrogenase (GAPDH) with some possible erythrose-4-phosphate dehydrogenase (EPDH) activity. Epd is mainly involved in PLP biosynthesis as an EFDH but has low level GAPDH activity.</td>
</tr>
</tbody>
</table>

(continued)
Table 5 Continued

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Isozyme rationale</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. git A, prpC</td>
<td>C</td>
<td>Different roles substrate</td>
<td>Citrate synthases</td>
<td>PyrC is a methylcitrate synthase with only minor citrate synthase activity. GltA is the effective citrate synthase.</td>
</tr>
<tr>
<td>1. pflB, tdcE</td>
<td>C</td>
<td>Different roles substrate</td>
<td>Pyruvate/2-ketobutyrate formate lyases</td>
<td>PflB’s principal substrate is pyruvate, TdcE’s principle substrate is 2-ketobutyrate but both can use the other’s main substrate.</td>
</tr>
<tr>
<td>1. cadA, ldcC</td>
<td>C</td>
<td>Kinetics regulation</td>
<td>Lysine decarboxylases</td>
<td>CadA is the most active decarboxylase. It is also more thermostable and has a low optimum pH ldcC is expressed weakly, less active and thermostable, but has a broad pH range with a higher optimum pH.</td>
</tr>
<tr>
<td>1. pykA, pykF</td>
<td>C</td>
<td>Kinetics regulation</td>
<td>Pyruvate kinases</td>
<td>PykF is remarkably stable. PykA shows only limited cooperativity among phosphoenolpyruvate binding sites.</td>
</tr>
<tr>
<td>1. gpt, hpt</td>
<td>C</td>
<td>Kinetics substrate</td>
<td>Phosphorybosyltransferases</td>
<td>Hypoxanthine is the main substrate for hpt, guanine the main substrate for gpt, but both enzymes can use the other’s favoured substrate.</td>
</tr>
<tr>
<td>1. pdxK, pdxY</td>
<td>C</td>
<td>Kinetics substrate</td>
<td>Pyridoxine/pyridoxal kinases</td>
<td>There are two distinct activities: pyridoxal kinase (PL) and pyridoxine kinase (PN). PdxK, pyridoxal kinase, has high PN and moderate PL activity. PdxY, pyridoxal kinase 2, has low PN and high PL activity.</td>
</tr>
<tr>
<td>1. glpQ, ugpQ</td>
<td>C(111)</td>
<td>Localisation substrate</td>
<td>Glycerolphosphoryl phosphodiesterases</td>
<td>GlpQ is periplasmic. UgpQ is cytoplasmic. They act on different ranges of phosphodiesterase.</td>
</tr>
<tr>
<td>1. aroF, aroG, aroH</td>
<td>C</td>
<td>Regulation</td>
<td>2-Dehydro-3-deoxy-phosphohexonate aldolases</td>
<td>These three aldolases have different feedback control and account for different percentages of aldolase activity: AroG (80%), AroF (20%) and AroH (1%).</td>
</tr>
<tr>
<td>1. cis, ybhO</td>
<td>C(73)</td>
<td>Regulation substrate</td>
<td>Cardiolipin synthases</td>
<td>YbhO can use different substrates; however, it does not seem to have in vivo activity.</td>
</tr>
<tr>
<td>1. ansA, ansB</td>
<td>C</td>
<td>Substrate</td>
<td>Transaminases</td>
<td>Both AnsA and AnsB catalyse transamination of aspartate to asparagine. AnsA uses NH₃ as the amine donor whilst AnsB uses glutamine. FabZ has broad substrate specificity acting on short to long fatty acid chains; FabA acts mainly on intermediate-length fatty acid chains. FAB is active in fatty acid elongation whilst FabB, used in membrane phospholipid synthesis, is not.</td>
</tr>
<tr>
<td>1. fabA, fabZ</td>
<td>C</td>
<td>Substrate</td>
<td>β-Hydroxyacyl-ACP dehydrogenases</td>
<td>FabB is active in fatty acid elongation whereas FabZ is used in membrane phospholipid synthesis.</td>
</tr>
<tr>
<td>1. fabB, fabF</td>
<td>C</td>
<td>Substrate</td>
<td>Acyltransferases</td>
<td>FabB is active in fatty acid elongation whereas FabZ is used in membrane phospholipid synthesis.</td>
</tr>
<tr>
<td>1. ackA, tdcD</td>
<td>C</td>
<td>Unknown</td>
<td>Acetate/propionate kinases</td>
<td>AckA function predicted by homology to AckA.</td>
</tr>
<tr>
<td>1. agaY, gatY</td>
<td>C</td>
<td>Unknown</td>
<td>Tagatose 1-6 bis-phosphate aldolases</td>
<td></td>
</tr>
<tr>
<td>1. aldA, aldB</td>
<td>C(63)</td>
<td>Unknown</td>
<td>Aldehyde dehydrogenases</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 5 Continued

<table>
<thead>
<tr>
<th>Homologous genes</th>
<th>Homology</th>
<th>Enzymatic activity</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. argF, argl</td>
<td>C</td>
<td>Unknown</td>
<td>Ornithine carbamoyltransferase</td>
</tr>
<tr>
<td>1. ddlA, ddlB</td>
<td>C(58)</td>
<td>Unknown</td>
<td>d-Alanine-2-alanine ligases</td>
</tr>
<tr>
<td>1. gntK, glnK</td>
<td>C</td>
<td>Unknown</td>
<td>Glutarnine transferases</td>
</tr>
<tr>
<td>1. gpmA, gpmB</td>
<td>C</td>
<td>Unknown</td>
<td>Phosphoglycerate mutases</td>
</tr>
<tr>
<td>1. ilvB, ilvN, ilvH</td>
<td>C(C67)</td>
<td>Unknown</td>
<td>Acetoxyhydroxybutanoate synthases (AHAS)</td>
</tr>
<tr>
<td>1. metL, thrA, lysC</td>
<td>C</td>
<td>metL, thrA</td>
<td>/P</td>
</tr>
<tr>
<td>1. rfbA, rffH</td>
<td>C</td>
<td>Unknown</td>
<td>dTDP-glucose pyrophosphorylases</td>
</tr>
<tr>
<td>1. rfbB, rffG</td>
<td>C</td>
<td>Unknown</td>
<td>dTDP-glucose 4,6-dehydratases</td>
</tr>
<tr>
<td>1. talA, talB</td>
<td>C</td>
<td>Unknown</td>
<td>Transaldolases</td>
</tr>
<tr>
<td>1. tktA, tktB</td>
<td>C</td>
<td>Unknown</td>
<td>Transketolases</td>
</tr>
</tbody>
</table>

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

Discussion

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

The pathway distance range considered herein (1-11 steps) corresponds, in essence, to the "within pathways" of our previous work. Here, we show that homology within pathway distances 1-11 is essentially localised to the shortest of these distances, and that overall recruitment of homologous proteins is rare within this range. Even at pathway distance 2, the distance at which recruitment of homologous proteins is most likely, less than 5% of the possible enzyme pairs share one or more domains (see Figure 2). Of the 3711 enzyme pairs considered (i.e. all the pairs at distances 1-11), only 95 (2.56%) show homology. However, we know recruitment is a common feature in SMM pathways, we can therefore conclude that much of the homology observed previously is the consequence of recruitment from distances greater than 11 steps, from other pathways or indeed from non-SMM genes. Nevertheless, we do observe 95 homologous pairs within pathways. These have a bias for short distances, with pathway distances 1, 2 and 3 accounting for two-thirds of the cases of homology (see Table 2). When homology does occur, our data show that it is most likely at the shortest pathway distances. Two patterns emerge: at a global level, recruitment events from the metabolic neighbourhood are rare. Recruitment does take place, but it

Table 6. Inline reuse in E. coli SMM

<table>
<thead>
<tr>
<th>Steps</th>
<th>No. inline recruitments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5+</td>
<td>None</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Pathway</th>
<th>Intervening genes</th>
<th>No. IF</th>
<th>Recruitment type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>DgoA</td>
<td>Galactonate catabolism</td>
<td>DgoK</td>
<td>1</td>
<td>MF</td>
<td>DgoA is a multifunctional enzyme, it first catalyses the dehydration of D-galactonate, DgoK then phosphorylates the product and the product of the phosphorylation is then lysed by DgoA acting this time as an aldolase</td>
</tr>
<tr>
<td>metL/</td>
<td>Homoserine biosynthesis</td>
<td>Asd</td>
<td>1</td>
<td>MF</td>
<td>Mett, is a bifunctional enzyme performing two non-consecutive reactions, first the phosphorylation of aspartate, then, after the dehydrogenase Asd, MetL oxidises l-aspartate-semialdehyde to homoserine. ThrA is an isozyme of MetL, similarly bifunctional and catalysing the same steps as described for metl</td>
</tr>
<tr>
<td>thrA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tktA/</td>
<td>Pentose phosphate pathway</td>
<td>[talA, talB]</td>
<td>1</td>
<td>MS</td>
<td>TktA catalyses the major transketolase activity in E. coli. In this pathway, it acts both on ribose-5-phosphate and xylose-5-phosphate, producing the substrates for the next reaction catalysed by transaldolases talA and talB, which, in turn, produce one of the two substrates for the second transketolase reaction listed in the pentose phosphate pathway. TktB catalyses the minor transketolase activity in E. coli; it is an isozyme of TktA and performs reactions identical with those listed for tktA</td>
</tr>
<tr>
<td>tktB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RelA</td>
<td>ppGpp metabolism</td>
<td>GppA</td>
<td>1</td>
<td>MS</td>
<td>GTP pyrophosphokinase catalyses the synthesis of guanosine 5'-triphosphate 3'-diphosphate (pppGpp) as well as guanosine 3',5'-bispyrophosphate (ppGpp) by transferring the pyrophosphoryl group from ATP to GTP or GDP, respectively. Phosphatase GppA catalyses the transition from pppGpp to ppGpp</td>
</tr>
<tr>
<td>deoD</td>
<td>Nucleotide metabolism</td>
<td>Add</td>
<td>1</td>
<td>MS</td>
<td>DeoD is a ubiquitous purine nucleoside phosphorylase multiply recruited within nucleotide metabolism. DeoD catalyses the generalised reaction purine nucleoside + orthophosphate = purine + α-D-ribose 1-phosphate. In this instance of reuse, DeoD phosphorylates Add (deoxyadenosine deaminase/adenosine deaminase) and DeoD mutually bracket one another (i.e. the chain deoD, add, deoD, add occurs in the nucleotide metabolism pathway). Functions of deoD and add are described above. HisB encodes a single polypeptide possessing the two enzyme activities: histidinol-P phosphatase and imidazole-glycerol phosphate dehydratase. The intervening enzyme, HisC, acts as a histidine phosphate aminotransferase</td>
</tr>
<tr>
<td>add</td>
<td>Nucleotide metabolism</td>
<td>DeoD</td>
<td>1</td>
<td>MS</td>
<td></td>
</tr>
<tr>
<td>hisB</td>
<td>Histidine biosynthesis</td>
<td>HisC</td>
<td>1</td>
<td>MF</td>
<td></td>
</tr>
<tr>
<td>Ndk</td>
<td>Pyrimidine ribonucleotide/side metabolism</td>
<td>PyrG</td>
<td>1</td>
<td>MS</td>
<td></td>
</tr>
<tr>
<td>Udk</td>
<td>Pyrimidine ribonucleotide/side metabolism</td>
<td>Cdd</td>
<td>1</td>
<td>MS</td>
<td></td>
</tr>
<tr>
<td>purB</td>
<td>Nucleotide metabolism</td>
<td>purH; purA</td>
<td>2</td>
<td>2</td>
<td>MS</td>
</tr>
<tr>
<td>ubG</td>
<td>Ubiquinone synthesis</td>
<td>ubih; ubiE; ubiF</td>
<td>3</td>
<td>3</td>
<td>MS</td>
</tr>
</tbody>
</table>

(continued)
Table 7 Continued

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

Nearby pathway enzymes are clustered in the genome

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

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

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

Interestingly, a similar "plateau" at pathway distance 4 was observed by Kolesov et al. and a
The median size of 3 “same-pathway” gene clusters was observed by Overbeek et al.,23 with the latter considered an underestimate by the researchers.

The observation that, in prokaryotes, functionally related genes cluster and that these genes often participate in the same biosynthetic pathway is neither unexpected nor novel, and this clustering is generally accepted to be the consequence of the operon gene organisation of prokaryotes. However, this relationship has not previously been explored quantitatively for the whole SMM of an organism and verified on a set of known and predicted operons. By correlating gene interval and pathway distance we “measure” the range of the clustering. Analysis of the E. coli genome suggests an average operon size of three to four genes.30 We conclude that, in general, for E. coli SMM enzymes, operons cluster blocks of three to four genes all within a short (four steps or less) pathway distance of one another. These operons are possibly co-regulated at a higher level in “uber-operons”44. This observation constitutes an important rationale for the often-exploited use of genomic co-localisation in gene function prediction.

**Genome distance, pathway distance and homology**

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

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

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

**Operons, Inline-reuse, Isozymes and regulation**

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

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

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

Few of our isozymes are co-located within an operon structure and therefore within a short distance of one another (only five out of our 59 sets had all isozymes in a set within five genes of one another). We have previously found lateral gene transfer not to play a key role in this observation.15 Adjacent genes would suggest a recent duplication event or strong evolutionary pressure to keep the genes nearby. It would appear that for our set of isozymes, the duplication events are not recent and evolutionary pressure has acted to separate the genes to allow segregation of transcriptional
control and/or future specialisation of the isozymes.

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

Conclusion

The data presented here add some support to the growing body of evidence suggesting patchwork evolution as the prevailing pathway evolution strategy.13,34,24 Recruitment from the metabolic neighbourhood (1–11 steps) is rare, as is conservation of substrate binding with a change in associated chemistry. Nevertheless, homology within the metabolic neighbourhood does occur, and when it does, it is more likely to occur at short pathway distances, including some well-known "retrograde-like" instances, suggesting multiple evolutionary mechanisms occurring in concert. We are observing catalytic constraints (i.e. the necessity to evolve a chemically efficient network for the production of small molecules), and we are observing extensive regulatory constraints (to ensure that the SMM is controlled efficiently to deal with changes in both intracellular and extracellular conditions). E. coli's extant SMM pathways are the result of these pressures.

The picture is complex; further clarification may come from effective phylogenetic analysis of all SMM enzymes (as performed "manually" by Copley & Bork for TIM barrels24) and experimental and theoretical investigation of metabolic pathways in not one but many organisms.46 Nevertheless, the interaction between the genome context, the metabolic context and the evolutionary context is certainly worth "mining" for information (e.g. see Kolesov et al.19). Such methods are effective because, as described here, there are exploitable relationships between all these contexts.

Methods

Generating the pathway dataset

The SMM pathways analysed in this work were obtained from the EcoCyc database.7 Pathway data were downloaded and converted to a format suitable for easy parsing by a number of Perl scripts.6 In keeping with the EcoCyc architecture, we downloaded data describing the pathway frames, and data describing reaction and enzyme-reaction frames,39 and stored them locally using a relational database management system (postgresql). Some of the data were edited manually following update reports (Alida Pellegrini-Toole and Monica Riley, personal communication).

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

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

Gene identification and ancillary data

The dataset generated above described the reaction frames and their relationships. Reaction frames describe a metabolic transition in terms of the substrates, products, co-factors and enzyme(s) catalysing that step.38,57 For all calculations involving the enzymes themselves, we needed to assign genes to reaction frames. Most of these assignments were obtained directly from EcoCyc, with some additional manual correction. In EcoCyc, genes are commonly described by their symbol (e.g. gapA or pgk) or their Blattner number (e.g. b3919 or b2926), but the Gene3D structural assignment procedure (see below) required GenBank protein identifiers (PIDs).44 We converted gene symbols and Blattner numbers to GenBank identifiers using a conversion list obtained from GenProtEC,77 which was edited manually following update reports (Margrethe Serres, personal communication).

Genomic location and gene intervals

Genes were assigned a chromosomal location by consulting the Gene Tablet for E. coli1 using the GenBank identifiers described above. We derived a gene order with genes ordered, irrespective of their strand, on the basis of their boundaries (i.e. starting at position 1 on the circular chromosome and numbering genes by scanning clockwise for boundaries, regardless of whether the boundary was a start codon, as would be the case for genes on the (+) strand or as stop codon, as would be the case for genes on the (−) strand); the ranking obtained was nearly identical with the Blattner

1 http://www.genome.wisc.edu/pub/analysis/m52orfs.txt
numbering. The gene interval is a measure of the number of genes separating two genes as derived from the aforementioned ordering (e.g. a gene interval of zero for genes sides by side, of one for two genes separated by a third gene, of two for genes separated by two other genes, etc.)

Assignments to structural and sequence families

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

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

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

Calculating pathway distances

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

Estimating p-values

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

Acknowledgments

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

References

Small Molecule Metabolism Enzymes in E. coli.


*Edited by G. von Heijne*

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